Analysing Data using Linear Models

With applications in R

Stéphanie M. van den Berg

2024-07-02

Contents

Preface

Target audience

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.

Why linear models?

Starting from linear models gives students a great start into the world of data analytics. Starting from the linear regression model, there is a whole world to explore of versatile models that can handle almost any data problem in the social sciences. It also gives an entry into the world of machine learning, as many linear models are also used in that context. Being familiar with linear regression and logistic regression gives student a head start when moving to that field later on and something to build on.

This book is unique in that it is about linear models but has a non-technical focus. Many texts on linear models start from linear algebra with complicated formulas for vectors and matrices. In contrast, this text sticks to relatively simple regression equations and as far as formulas are concerned, they do not go beyond addition, multiplication, division, taking the root or taking the square of something. Anybody with a diploma from secondary education should be able to follow the equations, maybe with a little assurance from a teacher.

R

Although this book is about analysing data using R, it is not a book on R itself. Online there are many resources to get a first introduction to R. In this book we provide the student with example R code that can be copied and tweaked to make it usable for their own datasets. In order to understand the code and to be able to use it, the student should get acquainted with the tidyverse way of coding in R, and particularly the pipe operator `%>%`. The most basic functions that we use in this book are `mutate()`, `filter()`, `select()`, `pivot_longer()`, `pivot_wider()`, `group_by()`, `summarise()` and `ggplot()`. The student should also be familiar with the difference between numeric variables and factors, know how to read in data files, and how to install and work with R packages. The rest is explained in this book.

How to read the book

Finally, some important tips on how to read this book. Although the focus is non-technical, some things need to be explained. But it is often hard for a student to figure out to what extent something should be understood in order to put the theory into daily practice. To guide the student, at the end of every chapter there is a list of take-away points. They form the essence of the learning goals. There is also a list of key concepts. If the student reads these take-away points and key concepts and feels they understand what is meant by them, they can stop reading and try to put the theory into practice by analysing some data on their own. If they get stuck, they can go back to the text again to see where they missed some key insights. Remember: analysing data is a skill, you learn by doing.

In order to further help the student to distinguish between what is essential and what is more detailed background information, we put the more detailed background information into grey boxes. For the day-to-day application of linear models, it is sufficient to understand what is in the main text. Whenever a student wants more explanation of how things actually work and why things are as they are, this can be found in the grey boxes. There are also a couple of sections that can be skipped entirely without losing track of the narrative of the book. These sections are indicated by having "(advanced)" in their title.

Note on statistical reporting and scientific notation

When analysing data with R, the output shows a lot of numbers, with varying numbers of decimals. When reporting the output, we adhere to the style

proposed by the American Psychological Association, which dictates that statistics should be reported to at most 2 decimals, and p-values to 2 or 3. Here we use 3 decimals for p-values.

Further, when numbers become either very large or very small, R shows output in scientific notation . If R shows a number like `8.77e- 2`, this should be read as 8.77×10^{-2}, which is equivalent to 0.0877. The easiest way to think about it is that you take the number before the e and then if you see a -2 after the e, you move the decimal dot in the number two places to the left. If you see a $+2$ after the e, you move the decimal dot two places to the right. In this way, `8.77e+ 1` should be read as 87.7, and `8.77e+ 2` should be read as 877. `8.77e+ 0` should be read as 8.77.

Disclaimer

Some of the data sets used in this book have been generated for the sole purpose of demonstrating statistical principles. These are not real data and no conclusions should be based on them.

Acknowledgements

Earlier editions of this work, based on Sweave files, were supported by the Faculty of Behavioural, Management and Social Sciences (BMS) at the University of Twente, the Netherlands. The current bookdown edition was generously supported by a BMS WSV grant awarded to the author. Jolien van Straalen-Pas and Marian van Dijk offered many suggestions that improved the text substantially. Others, including students, spotted a lot of errors, big and small, and made helpful suggestions. For the errors remaining, the author takes all responsibility. Part of the work was done while hosted at the Biostatistics department at the University of Southern Denmark.

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*.

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).

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

Table 1.2: Data matrix on teachers.

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

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

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 four vectors into a tibble.

```r
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 look at long format. The way data are stored in a matrix is important because for linear models, it is usually required to have the data in long format.

Suppose we measure depression levels in four clients, 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**. In other words, **depression_1** represents the score that was

Table 1.3: Data matrix with depression scores 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

recorded first during therapy, and **depression_4** represents the score that was recorded at the very end of the therapy.

This way of representing data on a variable that was measured more than once is called *wide format*. We call it *wide* because we have several columns where we put the depression scores in, which leads to a wide data matrix.

Note that this is only one way of looking at four measures of depression. Here, we have four depression variables: there is depression measured at time point 1, there is depression measured at time point 2, and so on, and each of these four variables varies only across clients (i.e., **depression_1** has different values for different clients).

An alternative way of storing multiple depression scores per client, is that depression is really only one variable and that it varies both across clients (some clients are more depressed than others) *and* across time (sometimes you feel more depressed than at other times).

Therefore, instead of using multiple columns, we put all the depression scores into one column with many rows. That way, the data matrix becomes long, 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 depression variables, we have only one variable for depression, and one extra variable **time** that indicates to which time point a particular depression measure refers to. Check that Tables 1.3 and 1.4 give us the exact same information, for instance what do both data matrices tell us about the third measure of the second client?

Now let's look at an example, where the advantage of long format becomes clear. Suppose the clients were measured twice and that the depression measures were taken on different days for different clients. Client 1 was measured on Monday and Tuesday, while client 2 was measured on 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 Friday for client 2, and missing values for measures on Wednesday thru Sunday for patient 1. The matrix has 2 rows and 8 columns, so 16 cells.

Table 1.6 shows the same data in long format. Now it has 4 rows and 3 columns, so 12 cells. A bit smaller than the data in wide format.

Table 1.4: Data matrix with depression scores 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

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

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

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

client	day	depression
1	Monday	5
1	Tuesday	6
2	Saturday	8
2	Sunday	7

Thus, storing data in long format is often more efficient in terms of storage room: you do not need so many cells with missing data. But a more important reason for preferring long format over wide format is purely practical: when analysing data using linear models, software packages like R require your data to be in long format. With long format, we mean that the most important variable in your analysis, the variable that you wish to understand better or to predict, should be stored in only one column. Thus, if you want to understand how depression scores are different for different individuals, and how the scores are different for different days, then you should put all the depression scores into one column. The same goes for visualisation. When using `ggplot()` for visualisation, as we do in this book, the data should be in long format too.

However, we will also come across some analyses without 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

Suppose we have the following dataframe on depression measures for two clients for every day of the week.

```
data_wide
```

```
## # A tibble: 2 x 8
##   client Monday Tuesday Wednesday Thursday Friday Saturday Sunday
##    <int>  <int>   <int>     <int>    <int>  <int>    <int>  <int>
## 1      1      5       6        NA       NA     NA       NA     NA
## 2      2     NA      NA        NA       NA     NA        8      7
```

We see for each client, the variable depression is measured twice. The measurements were done on different days for clients 1 and 2, so many values

are not there. In R, missing values are indicated by `NA` (not available). We would like to see all of the depression scores in one column, which means we have to transform this data set into long format.

```
data_wide %>%
  pivot_longer(cols = Monday:Sunday, # columns that need to be restructured
               names_to = "day", # name of new variable with old column names
               values_to = "depression") # name of new variable with values
```

```
## # A tibble: 14 x 3
##    client day       depression
##     <int> <chr>          <int>
##  1      1 Monday             5
##  2      1 Tuesday            6
##  3      1 Wednesday         NA
##  4      1 Thursday          NA
##  5      1 Friday            NA
##  6      1 Saturday          NA
##  7      1 Sunday            NA
##  8      2 Monday            NA
##  9      2 Tuesday           NA
## 10      2 Wednesday         NA
## 11      2 Thursday          NA
## 12      2 Friday            NA
## 13      2 Saturday           8
## 14      2 Sunday             7
```

If we use `values_drop_na = TRUE` we only get the rows that actually contain information about the depression levels, which leads to a smaller dataframe.

```
data_wide %>%
  pivot_longer(cols = Monday:Sunday, # columns that need to be restructured
               names_to = "day", # name of new variable with old column names
               values_to = "depression", # name of new variable with values
               values_drop_na = TRUE) # leave out rows with missing values
```

```
## # A tibble: 4 x 3
##   client day      depression
##    <int> <chr>         <int>
## 1      1 Monday            5
## 2      1 Tuesday           6
## 3      2 Saturday          8
## 4      2 Sunday            7
```

- The `cols` argument describes which columns need to be reshaped. In this case, it is all columns from **Monday** to **Sunday**
- The `names_to` argument gives the name of the variable that indicates from which column the data come, i.e. **day**
- The `values_to` argument gives the name of the variable that will be created from the depression scores stored in the cells, i.e. **depression**

1.4.2 From long to wide

Suppose we have the following dataframe called `data_long`. It contains depression levels at the beginning of therapy and at the end of therapy.

```
data_long
```

```
## # A tibble: 6 x 3
##   client time   depression
##    <int> <chr>       <int>
## 1      1 before          9
## 2      1 after          13
## 3      2 after          12
## 4      2 before         14
## 5      3 before         19
## 6      3 after          15
```

Suppose for some type of analysis, we need the data in wide format. We can use `pivot_wider()` to do that.

```
data_long %>%
  pivot_wider(names_from = time,
              values_from = depression)
```

```
## # A tibble: 3 x 3
##   client before after
##    <int>  <int> <int>
## 1      1      9    13
## 2      2     14    12
## 3      3     19    15
```

- The `names_from` argument gives the name of the variable that will be used for the new column names, i.e. **time**

- The `values_from` argument gives the name of the variable that stores the values that you wish to see spread out across several columns. Here that is **depression**

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 obtained the values of 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

Table 1.7: 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

name the nationality of a person, you have a nominal variable. Table 1.7 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.

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.

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

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, 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
```

[2]For instance, you may find that the assumptions of your linear model are not met, see Chapter 7.

```
## 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).

```
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 1.8. 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).

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

Furthermore, we see that there are quite a lot of children with ages between 5 and 8, but not so many children with ages below 3 or above 14. The advantage

Table 1.8: 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

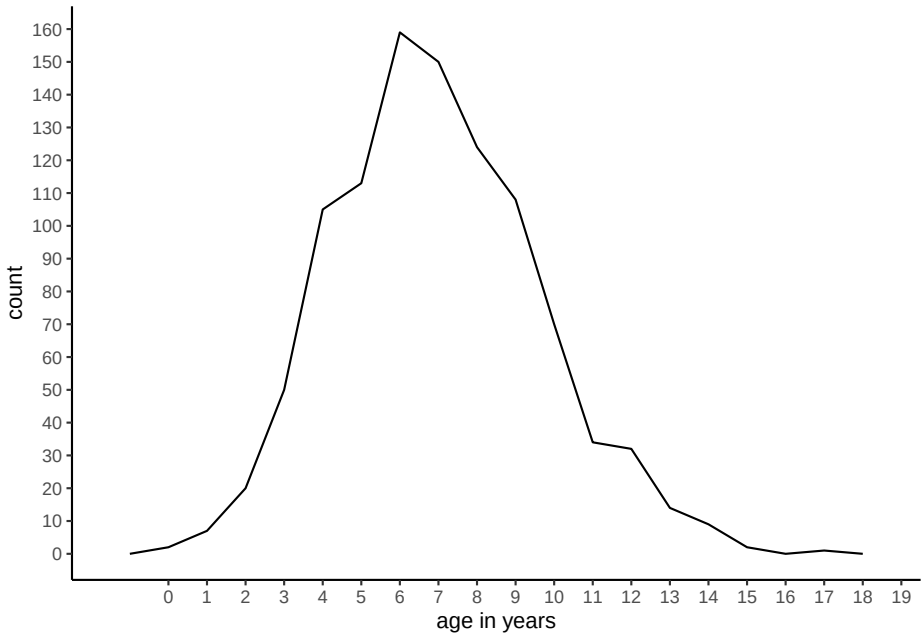


Figure 1.1: A frequency plot.

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

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

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

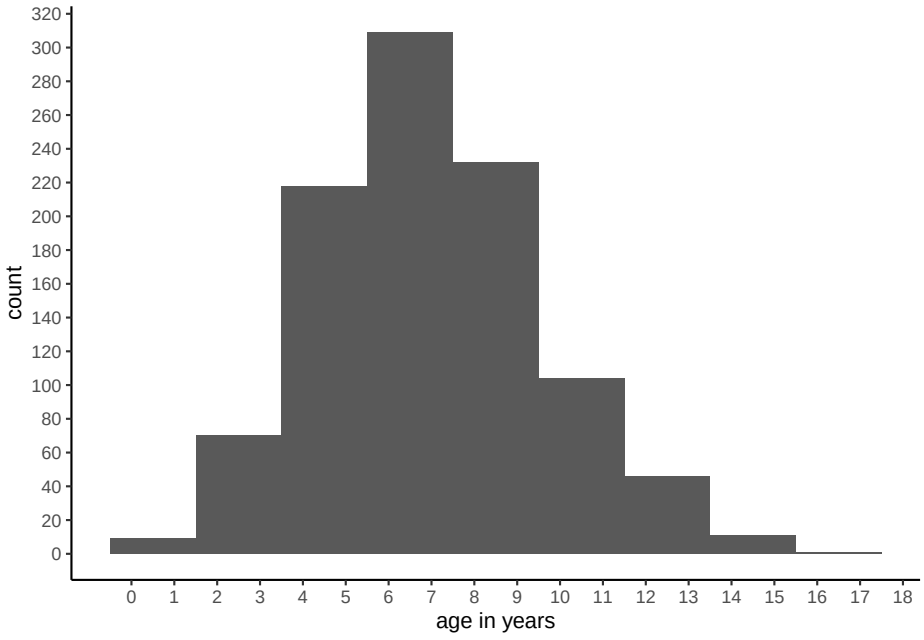


Figure 1.2: A histogram.

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.

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 1.8 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 1.8.

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 1.8. 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 `tabyl()` from the janitor package:

```
library(janitor)
mtcars %>%
  tabyl(cyl)
```

```
##  cyl  n percent
##    4 11 0.34375
##    6  7 0.21875
##    8 14 0.43750
```

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. The last column is the proportion (the term `percent` is misleading here). The table tells us that 34% of the cars have 4 cylinders.

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

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

```
##  cyl  n percent proportion
##    4 11 0.34375    0.34375
##    6  7 0.21875    0.21875
##    8 14 0.43750    0.43750
```

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

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

```
##  cyl  n percent proportion cumfreq cumprop
##    4 11 0.34375    0.34375      11 0.34375
##    6  7 0.21875    0.21875      18 0.56250
##    8 14 0.43750    0.43750      32 1.00000
```

The table tells us that 56% of the cars have 6 cylinders or less.

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

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

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

```
mtcars %>%
  ggplot(aes(x = mpg)) +
  geom_histogram(breaks = seq(5, 40, 5)) +
  scale_y_continuous(breaks = seq(0, 12, 1), minor_breaks = NULL)
```

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.

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.

[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.

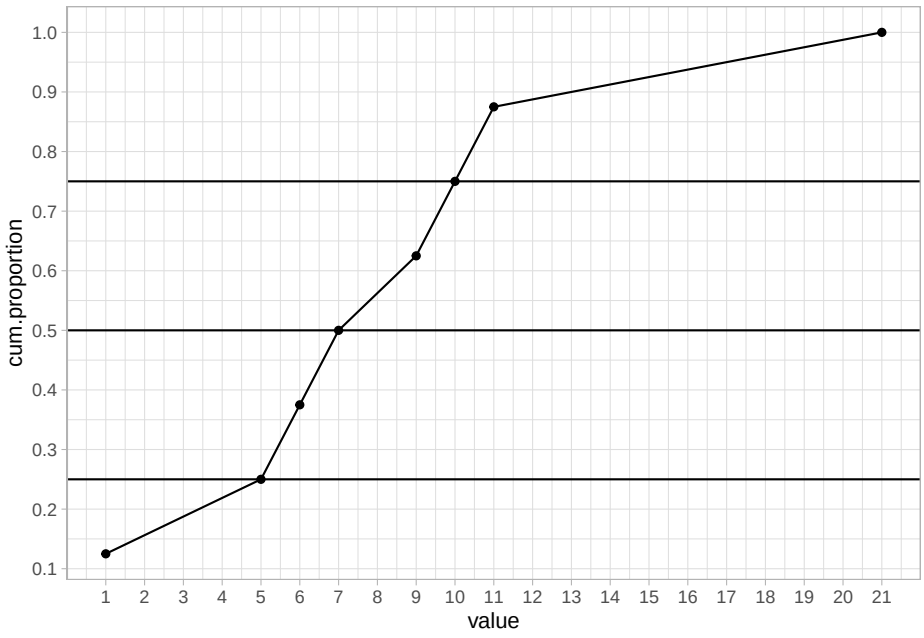


Figure 1.3: Cumulative proportions.

Alternatively, we could also use the frequency table (Table 1.8). First, if we want to have 25% of the children that are the youngest, and we know that we have 1000 children in total, we should have $0.25 \times 1000 = 250$ children in the first group. So if were to put all the children in a row, ordered from youngest to oldest, we want to know the age of the 250th child.

In order to find the age of this 250th child, and we look at Table 1.8, 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

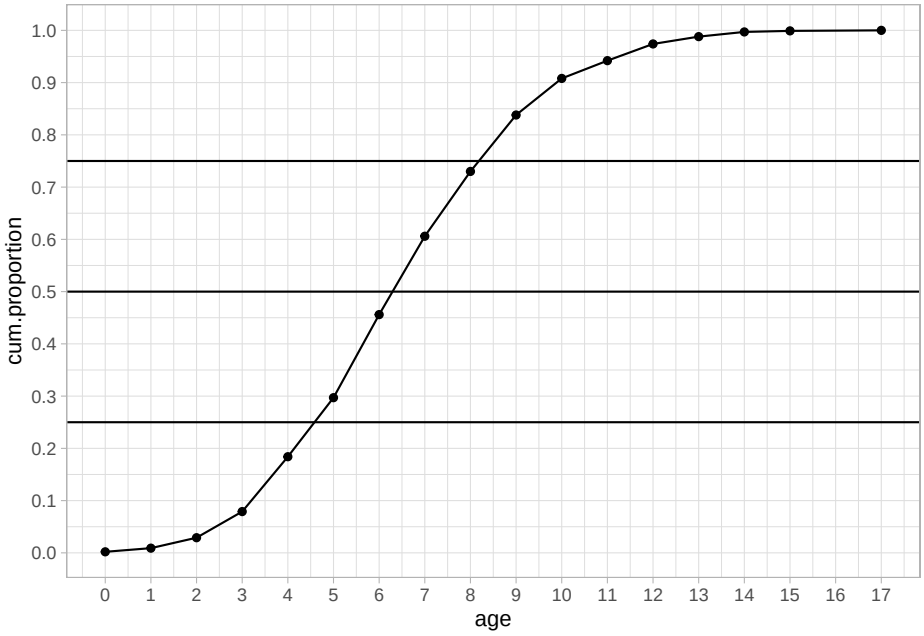


Figure 1.4: Cumulative proportions.

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:

```
mtcars$mpg %>% # select the values for the mpg variable
  quantile(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. In such formulas we often use X or Y to represent a particular variable. Suppose if the values of variable Y are 1, 2 and 3, then we denote the mean by $\bar{Y}$ (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]:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

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

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{Y_1 + Y_2 + Y_3}{n} = \frac{6 + (-3) + 21}{3} = \frac{24}{3} = 8$$

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

A closer look at the summation sign $\sum$

The sigma symbol $\sum$ tells us to sum a couple of things. We usually see something like this:

$$\sum_{i=1}^{3}$$

The number below the Sigma symbol tells us where we have to start making the sum, and the number above the Sigma symbol tells us where we have to stop making the sum. Right behind the Sigma symbol it says what we have to add. For instance,

$$\sum_{i=1}^{3} i$$

tells us that we have to sum a number of values for i, when i runs from 1 to 3. It is short for making the following sum:

$$\sum_{i=1}^{3} i = 1 + 2 + 3$$

since we substitute i for the values starting at 1 and ending at 3. Similarly, we could have the following sum:

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$$

,

because we count the x-values for which the subscripts run from 1 to 3.

Suppose x is a series (vector) of several values, say 5 values.

$$x = (7, 8, 4, 3, 2)$$

Then, when we compute $\sum_{i=1}^{3} x_i$, we only need the first three values of x.

$$\sum_{i=1}^{3} x_i = 7 + 8 + 4 = 19$$

If we have a lot of values for x, say n values, where n can be anything, and we want to add all the numbers up, we can write in a short way as

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

A bit shorter:

$$\sum_{i=1}^{n} x_i$$

Even shorter:

$$\sum_{i} x_i$$

It's best to be clear about what values for what letter the summation has to be done. For instance,

$$\sum_{k=2}^{3} (k + i)$$

means that we have to use two different values for k, and then add things up. Thus we have:

$$\sum_{k=2}^{3} (k + i) = (2 + i) + (3 + i)$$

As an exercise, write out the following sum to check your understanding of summation.

$$\sum_{l=0}^{2} (k - l)$$

The result should be equal to $3k - 3$.

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.

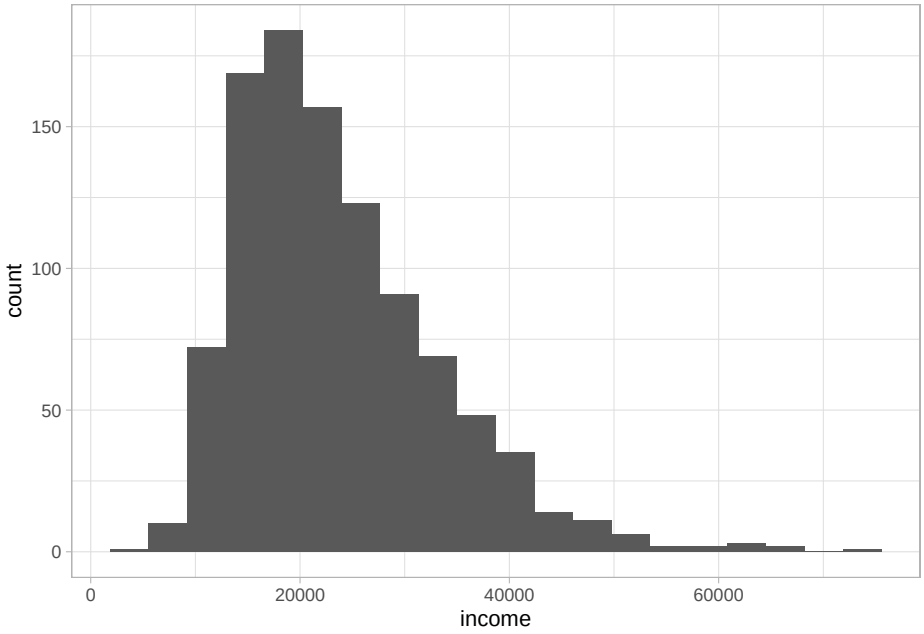


Figure 1.5: Distribution of yearly income.

Table 1.9: Four series of values and their respective medians and means.

X1	X2	X3	median	mean
4	5	8	5	6
4	5	80	5	30
4	5	800	5	270
4	5	8000	5	2670

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 remains the same.

This can be seen even more clearly by looking at the example in Table 1.9. 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 1.9). 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.

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, XL. 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.

Table 1.10: 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

"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 1.10. 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!).

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 cm) 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 it would be better to 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 $\bar{Y}$), you subtract this mean from each of these n values ($Y_i - \bar{Y}$), then you take the squares of these deviations, $(Y_i - \bar{Y})^2$,

and then add them up (take the sum of these squared deviations, $\sum(Y_i - \bar{Y})^2$. In formula form, this process looks like:

$$SS = \sum_i^n (Y_i - \bar{Y})^2$$

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:

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

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 $SS = 4 + 0 + 4 = 8$.

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

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]

$$\text{Var}(Y) = \frac{SS}{n} = \frac{\sum_i (Y_i - \bar{Y})^2}{n}$$

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[9]:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\text{Var}(Y)} = \sqrt{\frac{\sum_i (Y_i - \bar{Y})^2}{n}}$$

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

[8] Online you will often find the formula $\frac{\sum_i (Y_i - \bar{Y})^2}{n-1}$. 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.

[9] The greek letter sigma is used to indicate summation, but also refers to the variance. Capital Sigma, $\sum$, refers to summation, lower-case sigma, σ, refers to standard deviation.

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:

$$z = \frac{y - \bar{Y}}{\sigma}$$

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 $\frac{\sum_i (Y_i - \bar{Y})^2}{n-1}$ and $\sqrt{\frac{\sum_i (Y_i - \bar{Y})^2}{n-1}}$, respectively. We will discuss this further in Chapter 2. If you want to use the formula $\frac{\sum_i (Y_i - \bar{Y})^2}{n}$, 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)) %>%
  dplyr::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
## # i 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.

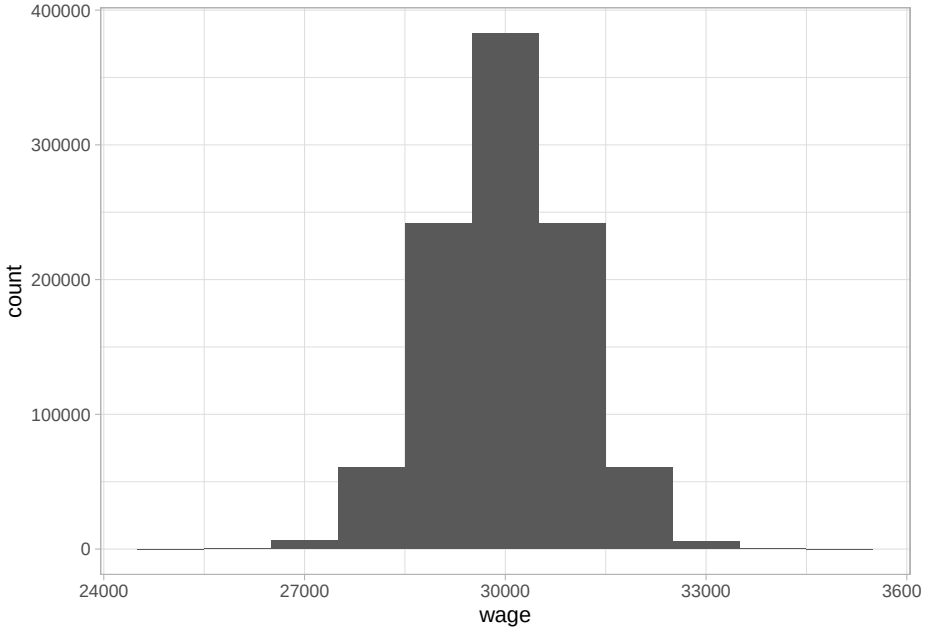


Figure 1.6: A histogram of wages with bin width 1000.

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.

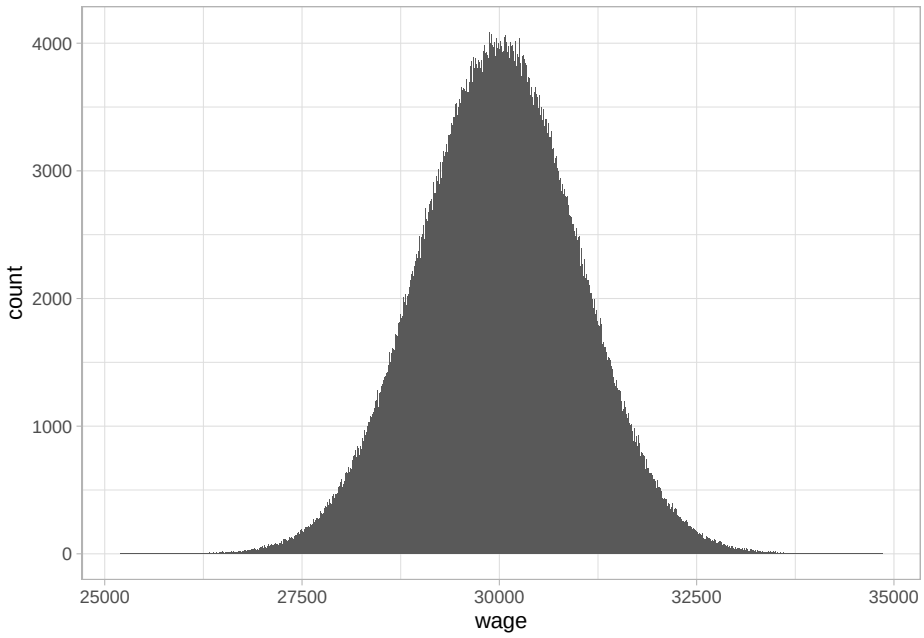


Figure 1.7: A histogram with bin width 10.

1.18 Density plots in R

Density plots can be obtained using `geom_density()`:

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

A histogram can be obtained using `geom_histogram()`:

```
mtcars %>%
  ggplot(aes(x = mpg)) +
  geom_histogram(bins = 8) # when you want to divide the data into 8 bins
```

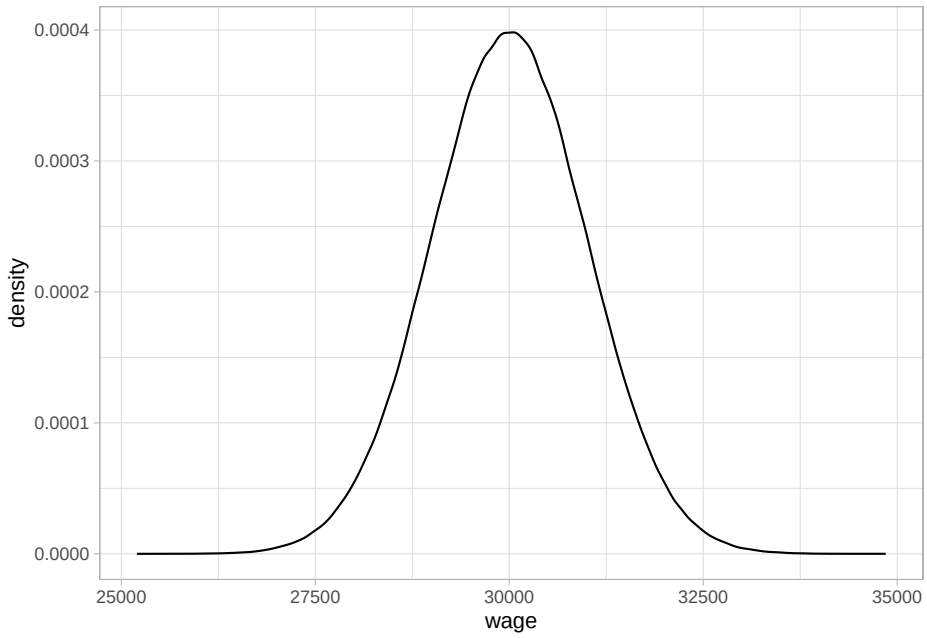
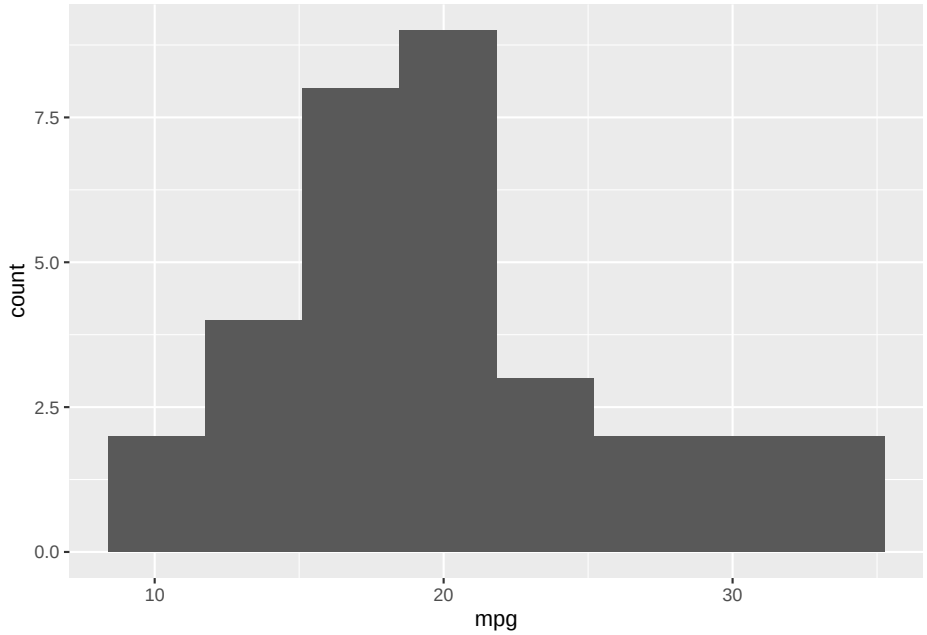


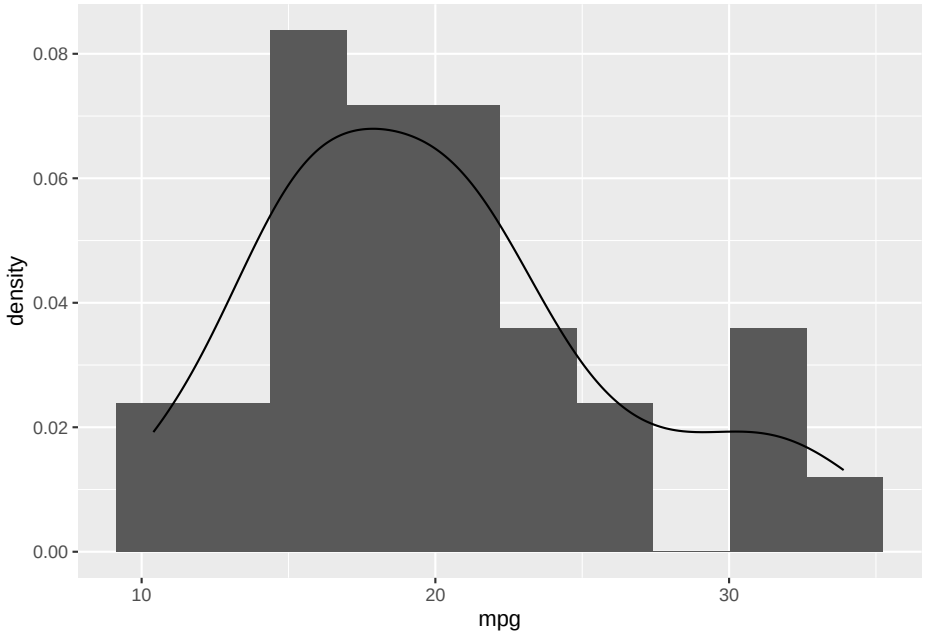
Figure 1.8: A density plot of the wage variable.



Often a combination is made of a histogram and a density plot. It's a bit

tricky to do because the histogram has counts on the y-axis, and the density has density on the y-axis, which is mostly smaller than 1. To put the counts on the same density scale, you can use:

```
mtcars %>%
  ggplot(aes(x = mpg)) +
  geom_histogram(aes(y = after_stat(density)), bins = 10) +
  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

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

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.

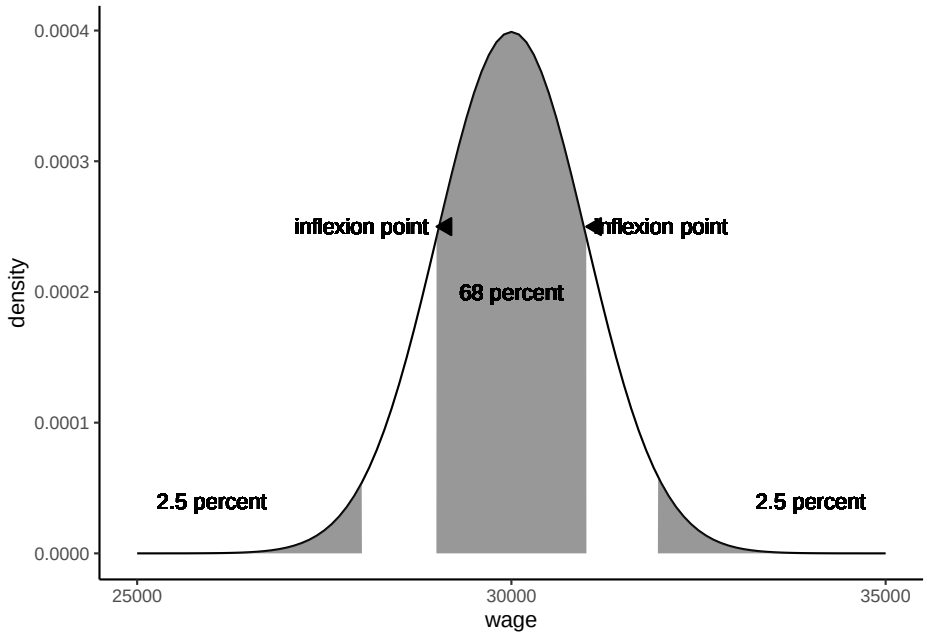


Figure 1.9: The theoretical normal distribution with mean 30,000 and standard deviation 1000.

In interactive Figure 1.10 you can see the shape of the normal distribution change as a function of mean and standard deviation. Try out different values for the mean and see what happens to the distribution. Then try out different values for the standard deviation and see what effect it has on the shape.

Please Wait

Figure 1.10: [Interactive] The shape of the normal distribution depends on only two parameters: the mean and the standard deviation. Change their values and see what it does to the density function.

Mathematicians have discovered many interesting things about the normal distribution. If the distribution of a variable closely resembles the normal

distribution, you can infer many things. One thing we know about the normal distribution is that the mean, mode and median are always the same. Another thing we know from theory is that the inflexion points[10] 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:

$$z = \frac{y - \bar{Y}}{\sigma}$$

Table 1.11 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 1.11. 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.

[10]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 1.11: 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

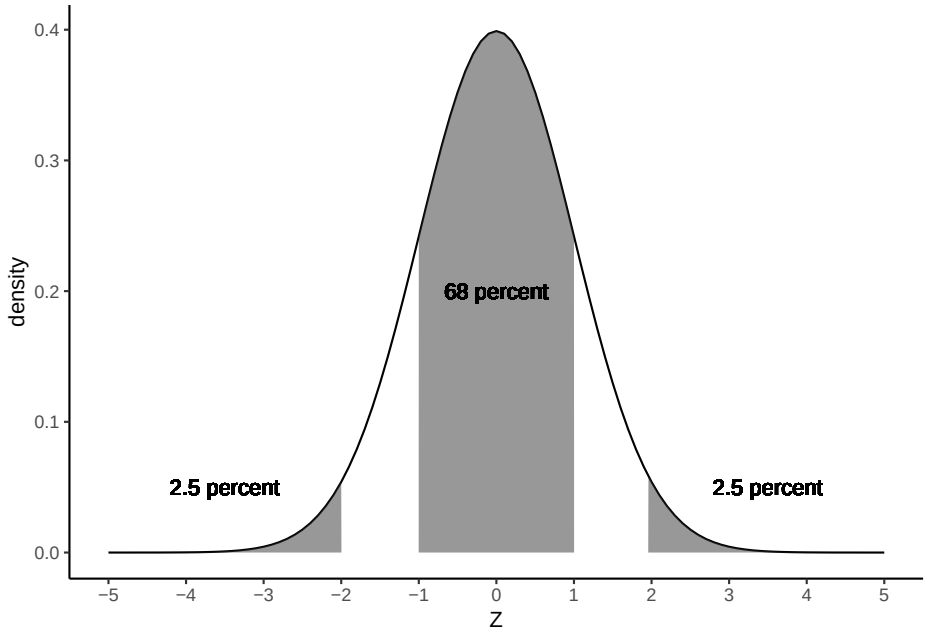


Figure 1.11: The standard normal distribution.

How do we know that 68% of the observations lie between the two inflexion points? Similar to proportions and cumulative proportions, we can plot the cumulative normal distribution. Figure 1.12 shows the cumulative proportions curve for the normal distribution. Note that we no longer see dots because the variable Z is continuous.

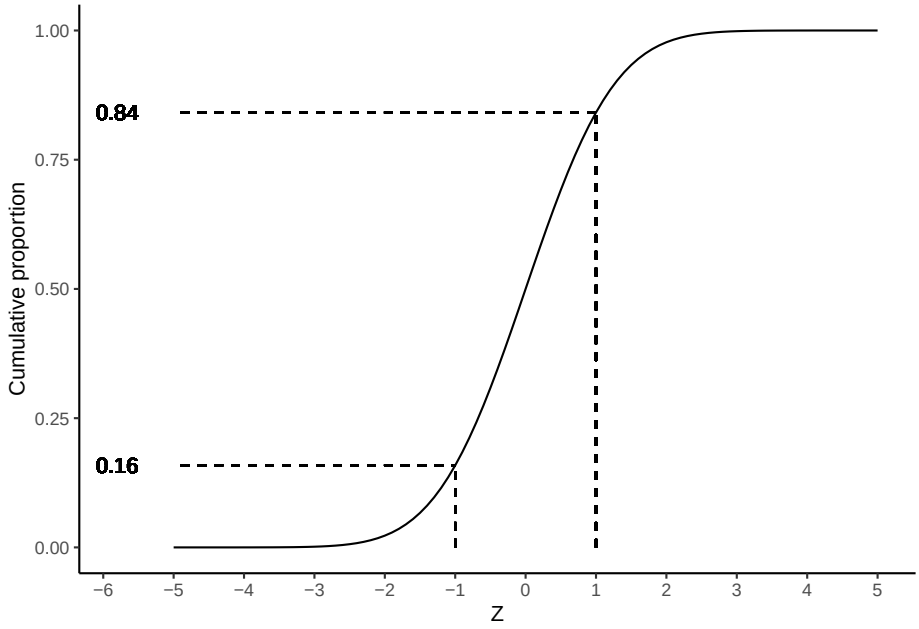


Figure 1.12: The cumulative standard normal distribution.

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[11] or in Appendix A, but also using R. Table 1.12 gives a short list of quantiles. From this table, you see that 1% of the z-values is lower than -2.33, and that 25% of the z-values is lower than -0.67. We also see that half of all the z-values is lower than 0.00 and that 10% of the z-values is larger than 1.28, and that the 1% largest values are higher than 2.33.

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

Table 1.12: 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

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.

Thus, if we return to our wages with mean 30,000 and standard deviation 1,000, we know from Table 1.12 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 1.12 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 1.12 we see that the 0.10 quantile is a z-value of -1.28. This tells us that the 0.10 quantile for the personality scores lies at 1.28 standard deviations below the mean. The mean is 18, so the 0.10 quantile for the personality scores lies at 1.28 standard deviations below 18. The standard deviation is 2, so this amounts to $18 - 1.28 \times 2 = 15.44$. This tells us that 10% of the scores on this test are 15.44 or lower.

Such handy tables are also available for other theoretical distributions.

Theoretical distributions are at the core of many data analysis techniques, including linear models. In this book, apart from the normal distribution, we will also encounter other theoretical distributions: the t-distribution (Chapter 2), the F-distribution (Chapter 6), the chi-square distribution (Chapters 2, 8, 14, 15 and 16) and the Poisson distribution (16).

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 75.33 or less, 50% of the values are 100.00 or less, and 95% of the values are 124.67 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 1.13 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.

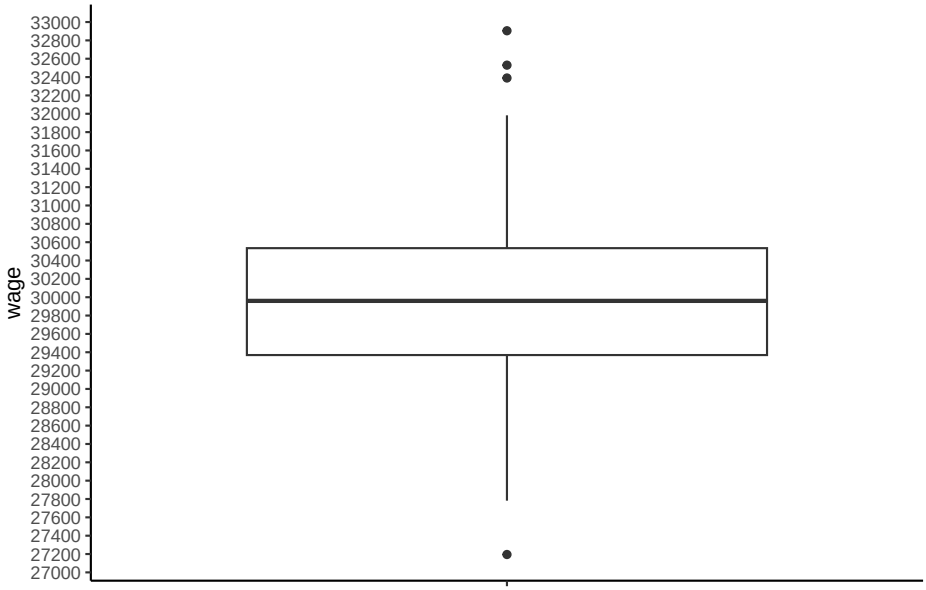


Figure 1.13: A box plot of the wages earned by a sample of 150 administrative clerks.

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 1.13, 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 1.13. 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 $30800 + 2100 = 32900$.

Table 1.13: A frequency table of nationalities.

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

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 1.13), but a *bar chart* shows the distribution in a more dramatic way, see Figure 1.14.

Sometimes, counts of values of a categorical variable are displayed as a *pie chart*, see Figure 1.15. Pie charts are however best avoided. First, because compared to

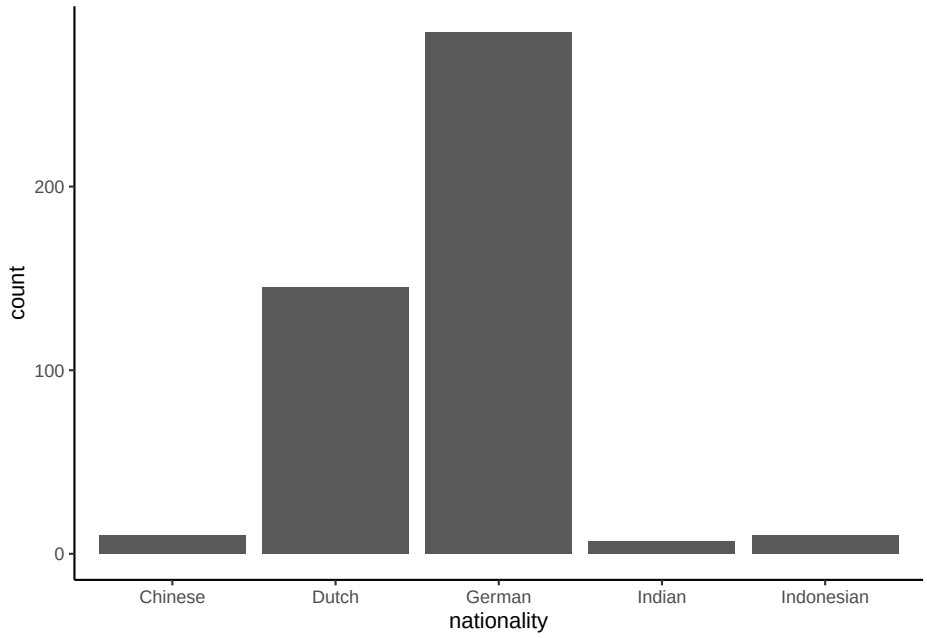


Figure 1.14: A bar chart of the observed nationalities in a lecture hall.

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 1.14 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 1.15. In sum, pie charts are best replaced by bar charts.

Ordinal variables are often visualised using bar charts. Figure 1.16 shows the variation of the answers to a Likert questionnaire item, where Nairobi inhabitants are asked "To what degree do you agree with the statement that the climate in Iceland is agreeable?". With ordinal variables, make sure that the labels are in the natural order.

1.24 Visualising categorical and ordinal variables in R

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

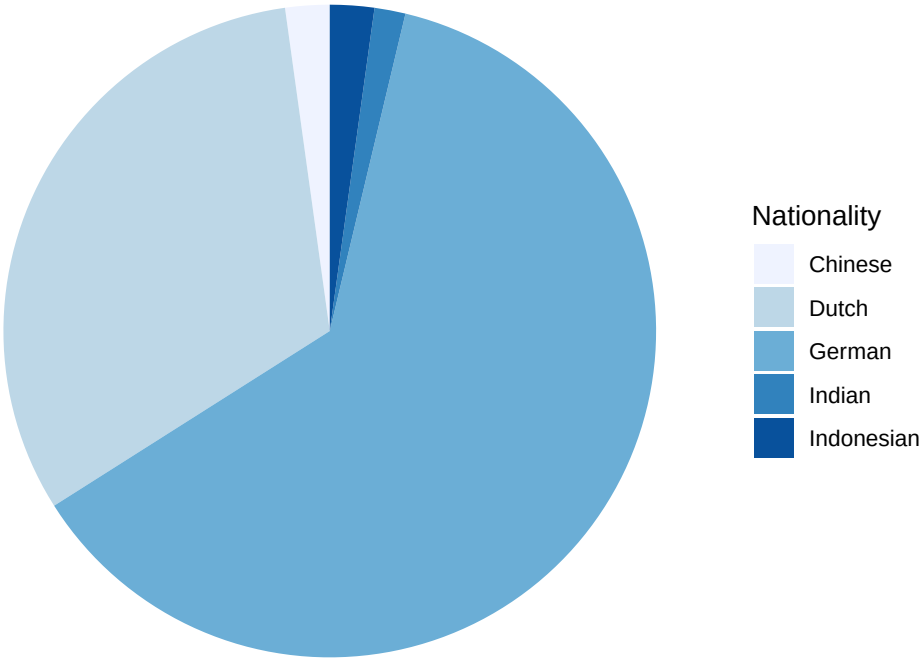


Figure 1.15: A pie chart of nationalities.

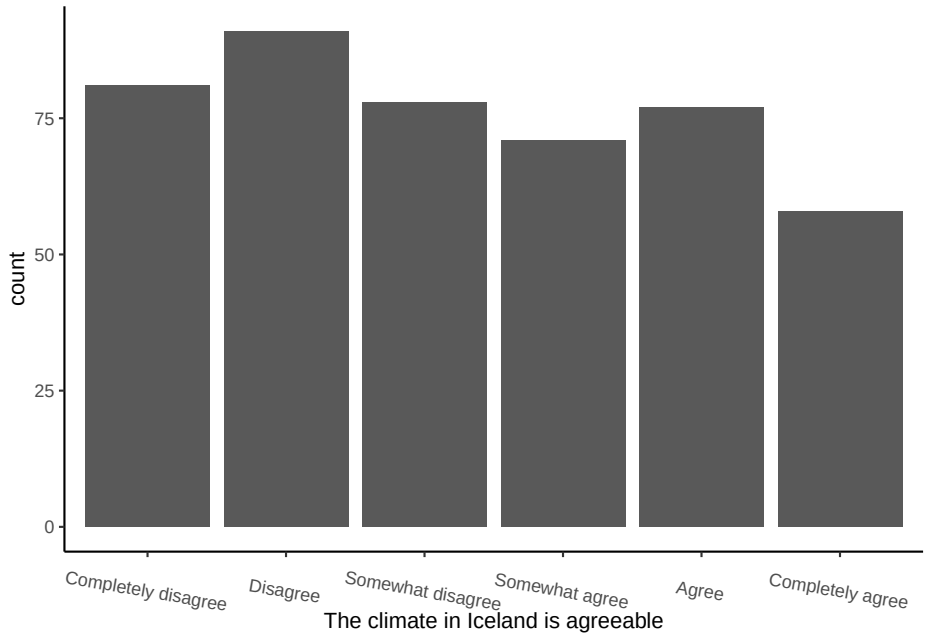


Figure 1.16: Opinions on the climate in Iceland.

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

If you really want a pie chart, then do:

```
mtcars %>%
  tabyl(cyl) %>%
  ggplot(aes(x = "",
             y = percent,
             fill = factor(cyl))) +
  geom_col(width = 1) +
  coord_polar(theta = "y") +
  xlab("") +
  ylab("") +
  theme_void() +
  scale_fill_brewer(palette = "Blues") +
  labs(fill = "Cylinders")
```

Table 1.14: Cross-tabulation of colour and length for twenty pencils.

	Colour	
	blue	**red**
long	4	0
short	8	8

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 1.14.

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 1.15. We see that the table is much larger than Table 1.14. 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

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

	Colour	
Length	**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

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.

The alternative for two variables where one is categorical and the other one is numeric, is to create a *box plot*. Figure 1.17 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. This value is less than 1.5 times $5.03 = 7.55$ 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

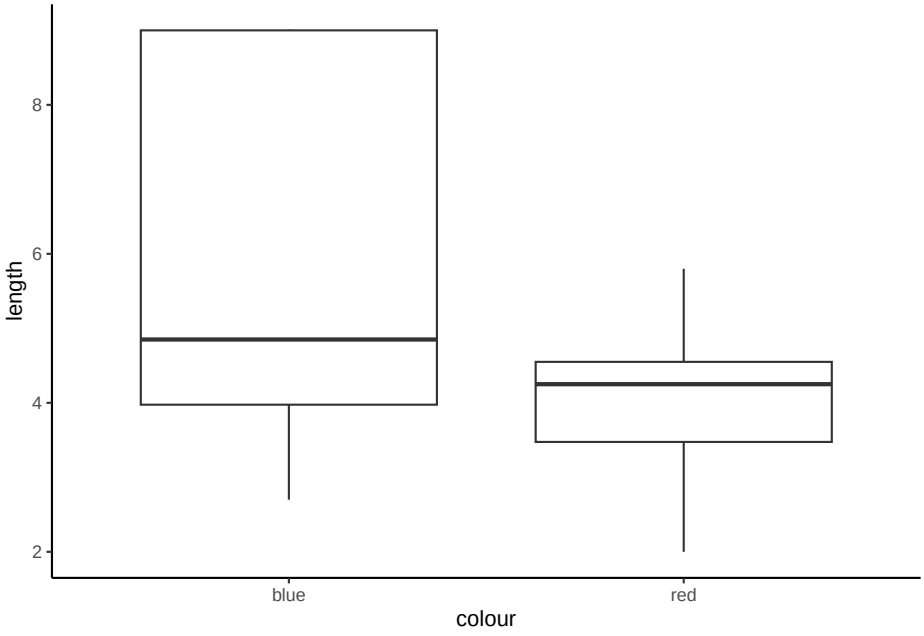


Figure 1.17: A box plot of the pencil data.

1.17 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 1.16 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 1.18 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 1.19 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 1.20.

Table 1.16: Cross-tabulation of length (rows) and weight (columns) for twenty pencils.

Length	Weight					
	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

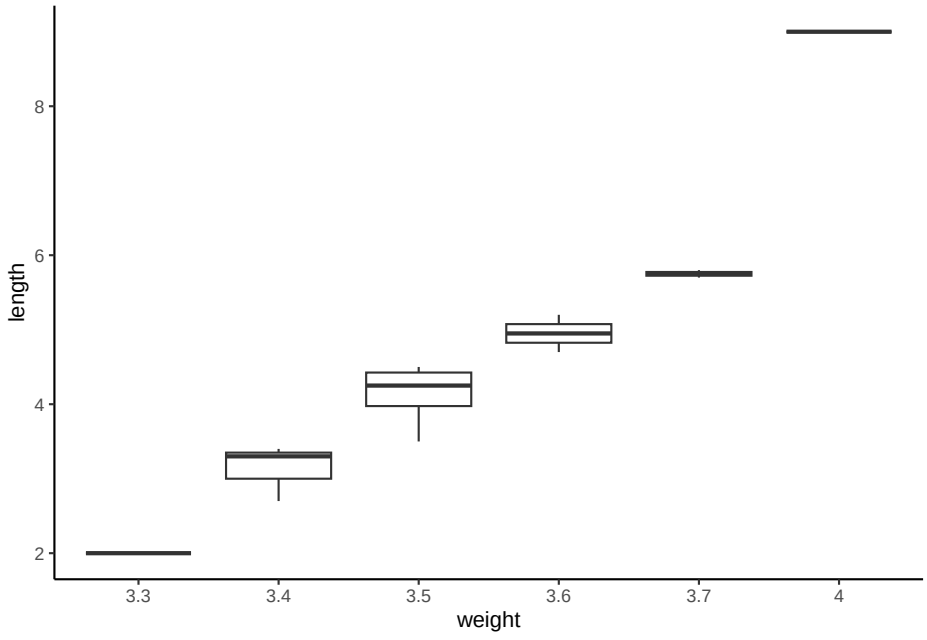


Figure 1.18: A box plot of the pencil data.

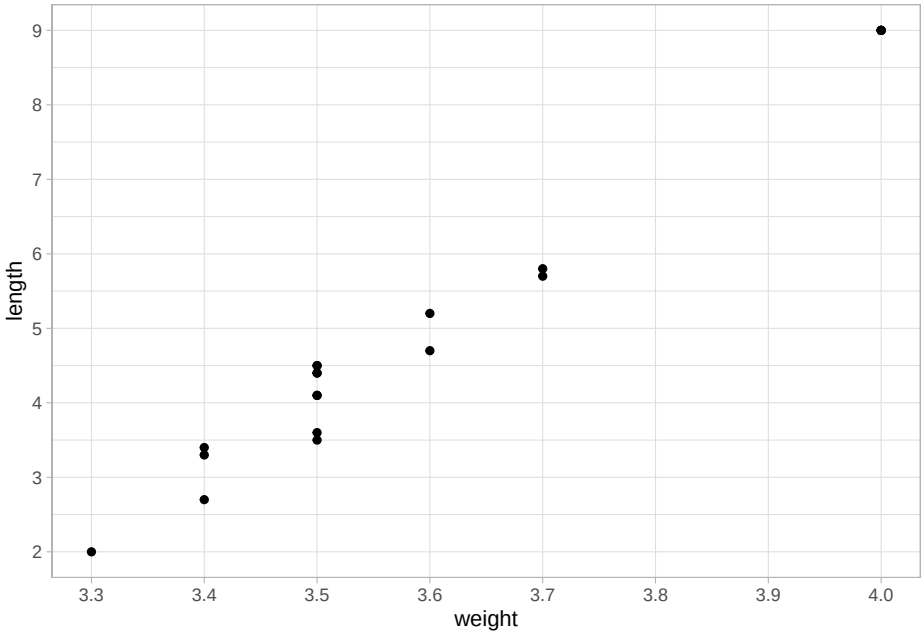


Figure 1.19: A scatter plot of length and weight.

You see that by visualising two variables, important patterns may emerge that you can easily overlook when only looking at the values. Cross-tables, box plots and scatter plots are powerful tools to find regularities but also oddities in your data that you'd otherwise miss. Some such patterns can be summarised by straight lines, as we see in Figure 1.20. 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()`:

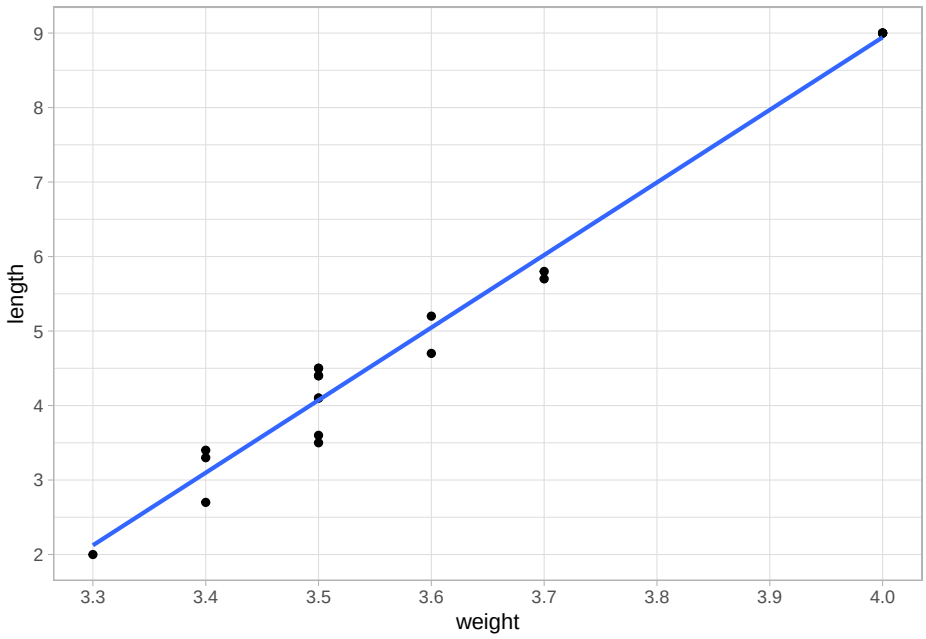


Figure 1.20: A scatter plot of length and weight, with a straight line that summarises the relationship.

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

A cross table for two categorical variables can be made using `tabyl()` from the `janitor` package:

```
# install.packages("janitor") # only run when not installed earlier
library(janitor)
mtcars %>% tabyl(cyl, gear)
```

```
##  cyl  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).

The `janitor` package includes functions that help you make tables that are prettier than this basic one.

1.27 Take-away points

- Data on units and variables can be stored in a data matrix.
- Every column in a data matrix is a variable.
- Data can be stored in long or wide format.
- There are different kinds of variables in terms of their measurement level.
- How you describe variables and how you visualise them depends on their measurement level.
- There are several ways to see how two variables co-vary.

Key concepts

- Data
- Units
- Variables
- Variation
- Data matrix
- Wide and long format

- Measurement level
- Numeric/ordinal/categorical
- Interval variable
- Ratio variable
- Dichotomous variable
- Nominal variable
- (Cumulative) frequency
- (Cumulative) proportion
- Histogram
- Quartiles, quantiles, percentiles
- Central tendency
- Mean, median, mode
- (Interquartile) range
- Sum of squares (SS)
- Variance (σ^2) and standard deviation (σ)
- Standardised score/ z-score
- Density plot
- (Standard) normal distribution
- Empirical rule
- Box plot
- Bar chart
- Pie chart
- Cross-table
- Scatter plot

1.28 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 10 shows how you can make elaborate statements about differences between groups of observations, in case one of the predictor variables is a categorical variable.

Chapter 11 discusses dealing with research questions that arise not *before* the data analysis, but *during* the data analysis.

Chapters 12 and 13 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 14 discusses alternative methods to study variables that are repeatedly measured in the same research unit.

Chapters 15 and 16 discuss *generalised linear models*. These are models where the outcome variable is not numeric and continuous. Chapter 15 discusses *logistic regression*, a method that is appropriate when the outcome variable has only two values, say "yes" and "no", or "pass" and "fail". Chapter 16 discusses methods 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. An often used method is the Pearson chi-square statistic that can be used when you have two categorical variables and you want to analyse crosstables. For more complicated situations we introduce *Poisson regression*.

Chapter 2

Inference about a mean

2.1 The problem of inference

For the topic of inference, we turn to an example from biology. 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

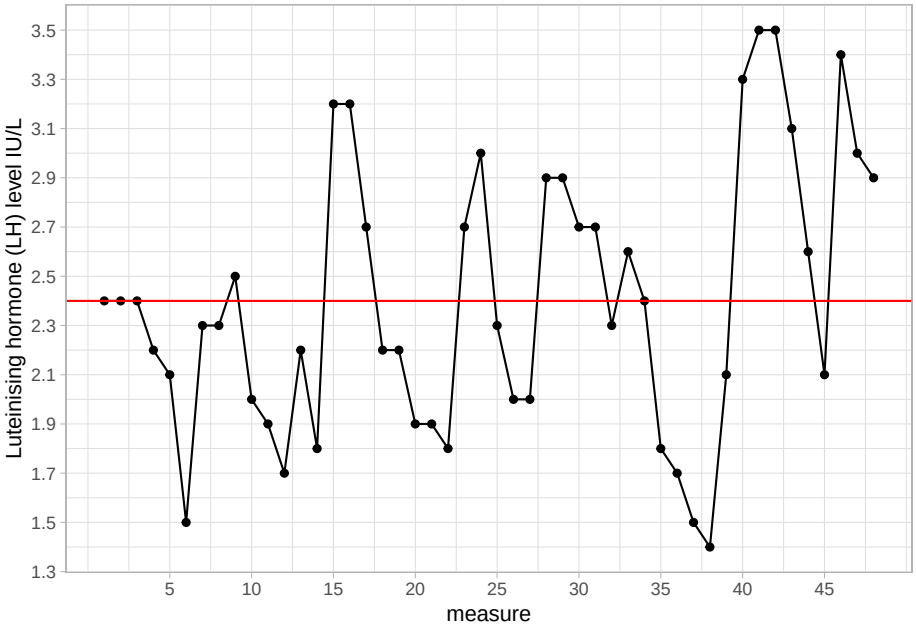


Figure 2.1: Luteinising hormone levels measured in one female, 48 measures taken at 10-minute intervals.

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 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.

Table 2.1: Imaginary data on elephant height when 5 random samples (columns) of 10 elephants (rows) are drawn from the population data. For each sample, the mean and the variance are computed.

elephant	**sample1**	**sample2**	**sample3**	**sample4**	**sample5**
1	3.22	3.08	2.90	3.31	2.93
2	3.21	2.49	2.93	3.41	3.21
3	4.12	3.14	3.71	2.63	3.55
4	3.63	3.47	2.72	3.31	2.77
5	3.01	3.79	3.82	3.83	3.38
6	3.14	2.82	3.82	3.40	3.34
7	3.46	3.24	3.38	3.28	2.90
8	3.07	2.91	3.49	3.08	2.56
9	3.28	3.60	3.39	2.85	2.59
10	3.47	2.55	3.01	3.23	3.25
mean	3.36	3.11	3.32	3.23	3.05
variance	0.10	0.17	0.15	0.10	0.11

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.

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.

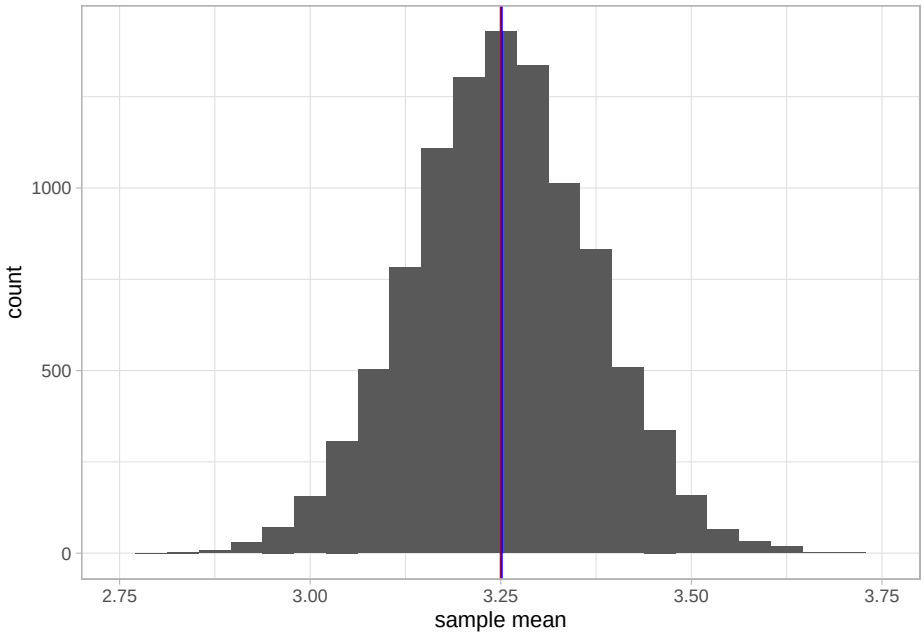


Figure 2.2: A histogram of 10,000 sample means when the sample size equals 10.

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 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.

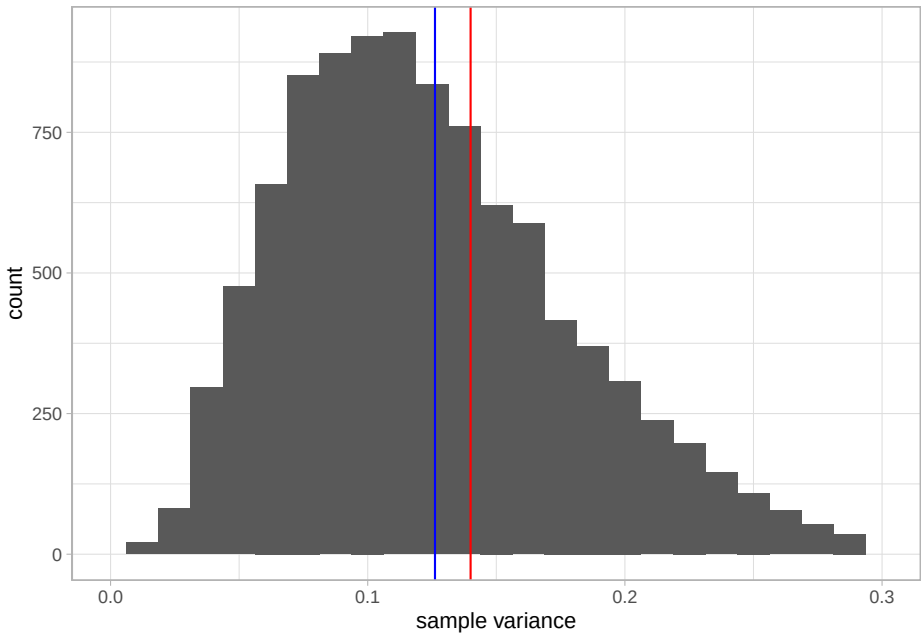


Figure 2.3: A histogram of 10,000 sample variances when the sample size equals 10. The red line indicates the population variance. The blue line indicates the mean of all variances observed in the 10,000 samples.

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, 14, 15 and 16. 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!

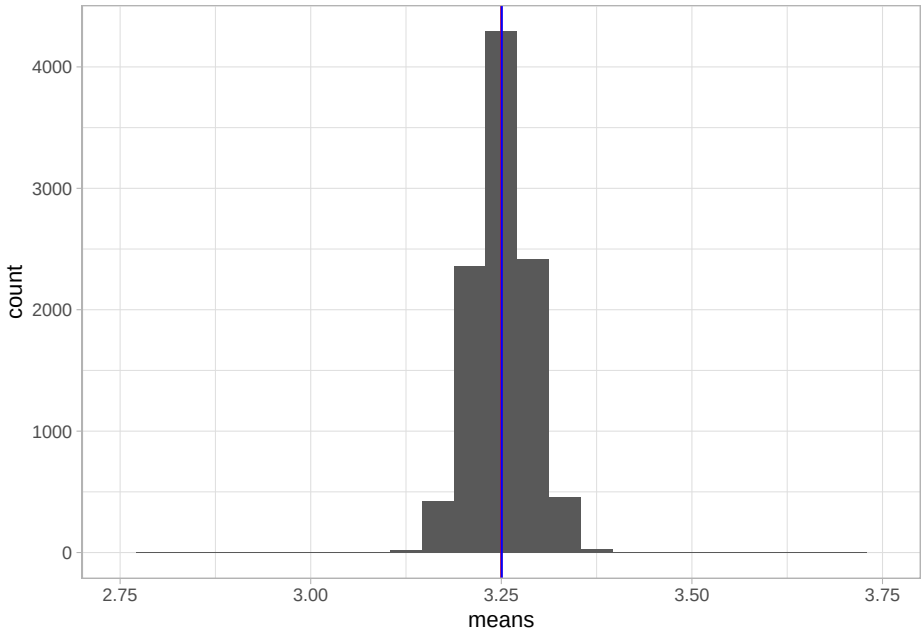


Figure 2.4: A histogram of 10,000 sample means when the sample size equals 100.

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.

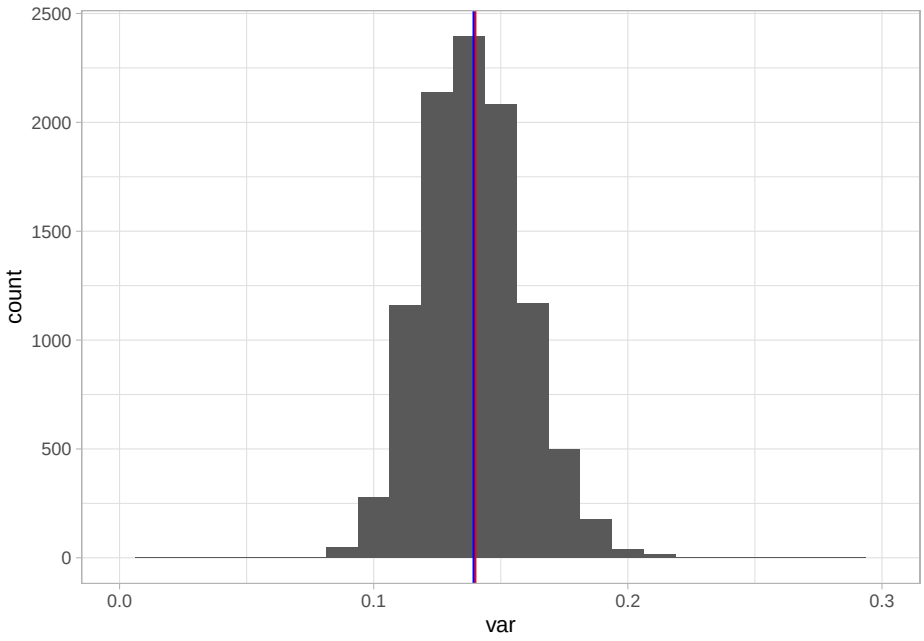


Figure 2.5: A histogram of 10,000 sample variances when the sample size equals 100.

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 of the population mean, but that the variance of the sample data is not an unbiased

[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.

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

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

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 $\frac{n}{n-1}$:

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

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:

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

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.

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

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.

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 sum 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 sum 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 2.80 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.

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

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.

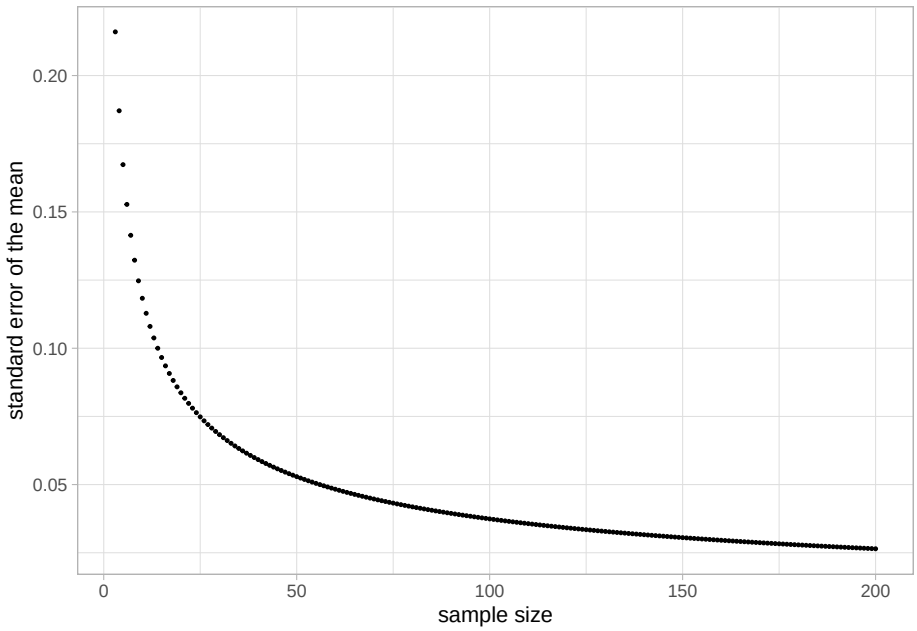


Figure 2.6: Relationship between sample size and the standard error of the mean, when the population variance equals 0.14.

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?

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:

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

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):

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

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 $\sqrt{\frac{0.14}{4000}} = 0.006$ and $\mu = 3.25$, so that 68% of the sample means lie between 3.244 and 3.256.

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 3.244 and larger than 3.256. 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 $3.25 - 1.96 \times \sqrt{\frac{0.14}{4000}} = 3.238$ or larger than $3.25 + 1.96 \times \sqrt{\frac{0.14}{4000}} = 3.262$. 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:

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

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 $3.25 - \text{MoE}$ and $3.25 + \text{MoE}$.

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.

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

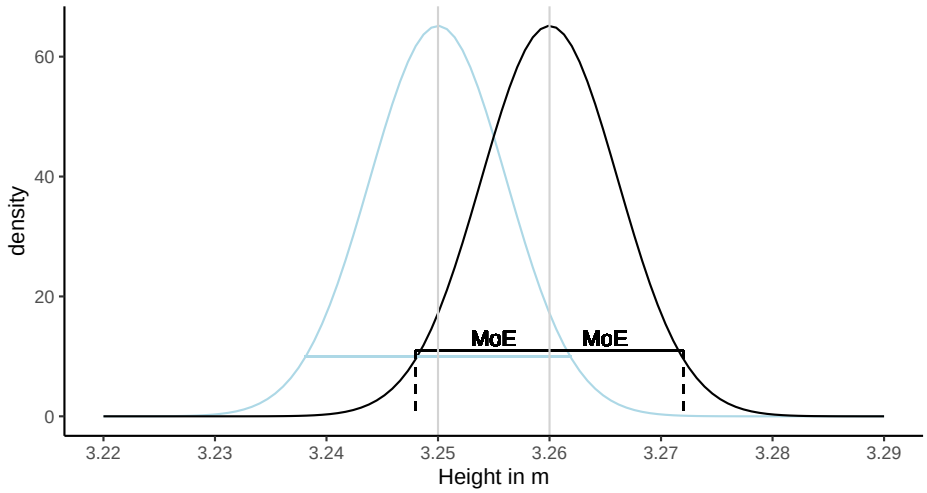


Figure 2.7: Illustration of the construction of a 95% confidence interval. Suppose we find a sample mean of 3.26 and a sample variance of 0.15, with $n = 4000$. The black curve represents the sampling distribution if the population mean would be 3.26 and a variance of 0.15. In reality, we don't know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

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 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.

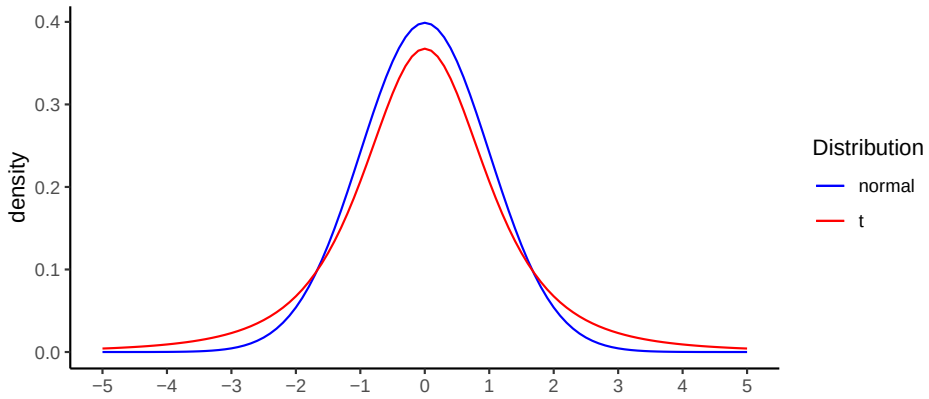


Figure 2.8: Distribution of *t* with sample size 4, compared with the standard normal distribution.

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 $3.26 - 3.18 \times \sqrt{\frac{0.15}{4}} = 2.64$, and +3.18 times the standard error away from the mean of 3.26, is equal to $3.26 + 3.18 \times \sqrt{\frac{0.15}{4}} = 3.88$. 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

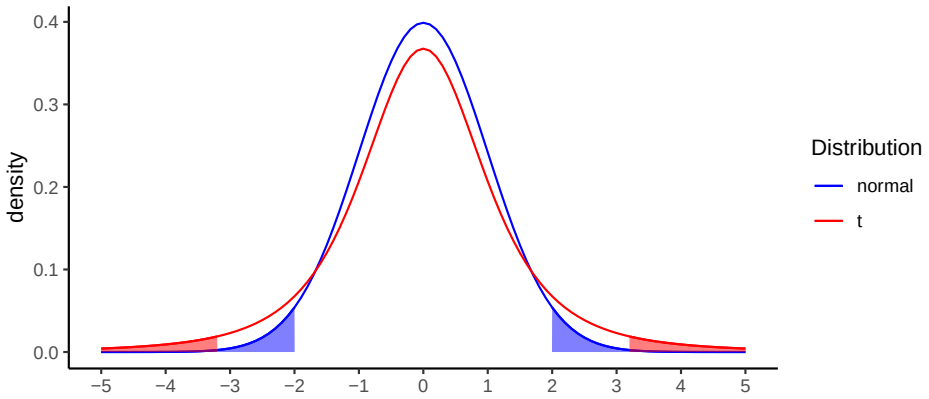


Figure 2.9: Distribution of *t* with sample size 4, compared with the standard normal distribution. Shaded areas represent 2.5% of the respective distribution.

true 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 mean is a certain value, say $\mu = m_0$, we assume that the standard error of the mean is equal to $\sigma_{\bar{y}}$, and we know that if we would look at many many samples and compute standardised sample means (by using sample means $\bar{y}$ and sample variances s^2), 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 the population variance. Then you imagine what sample means would be reasonable to find and what sample variances. But of course, it's in fact the opposite: you only know the mean and variance of *one* sample 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

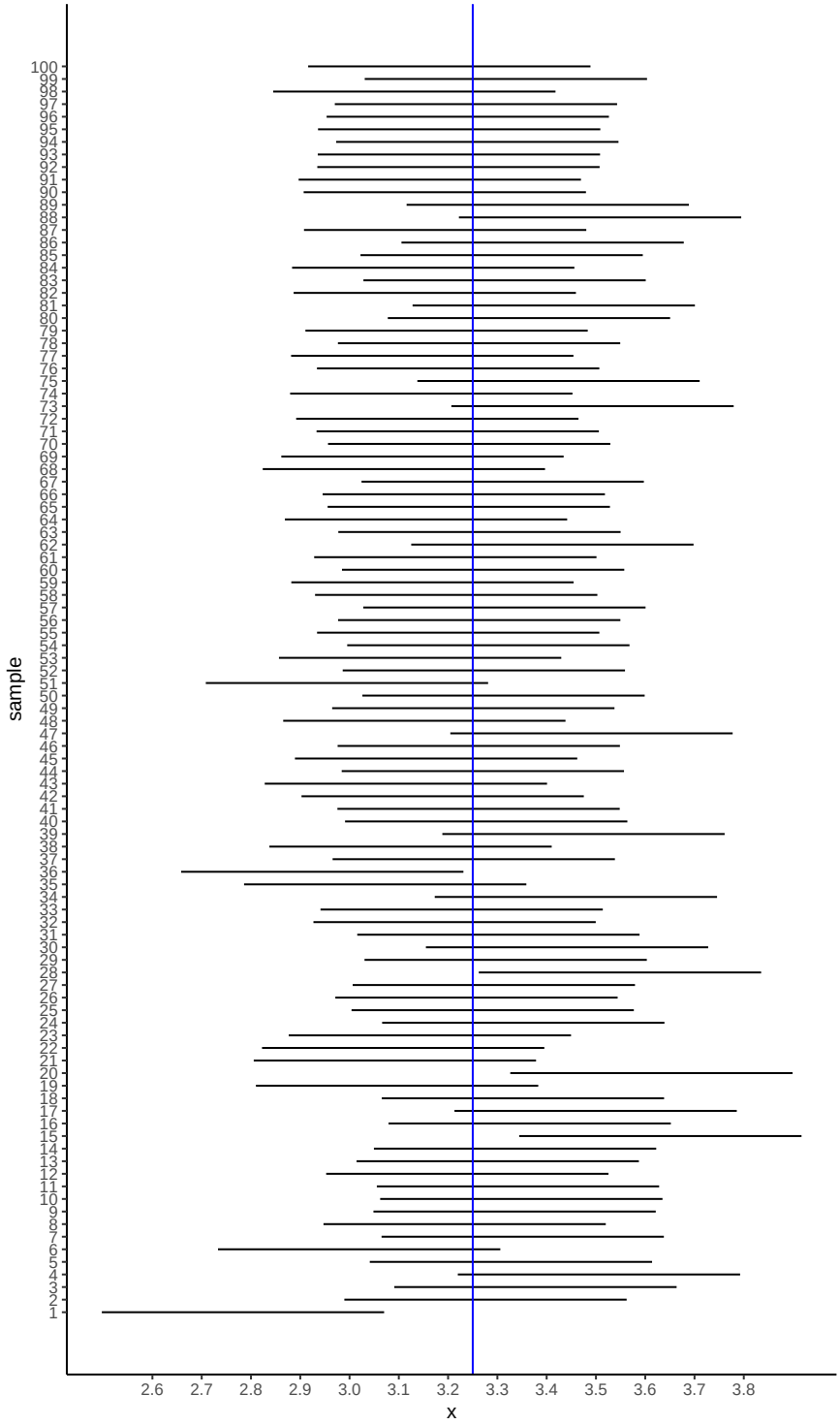


Figure 2.10: Confidence intervals.

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 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 many degrees of freedom 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.

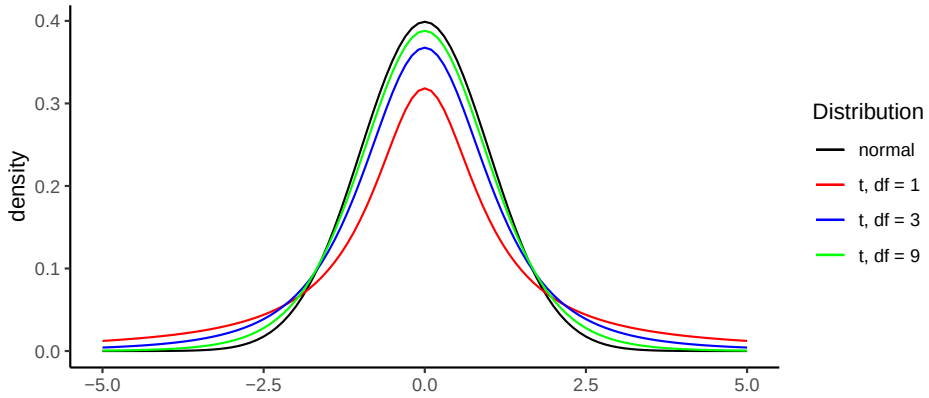


Figure 2.11: Difference in the shapes of the standard normal distribution and *t*-distributions with 1, 3 and 9 degrees of freedom.

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.

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

Table 2.2: Quantiles for the standard normal and several t-distributions with varying degrees of freedom.

		t-distribution				
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

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 2.09 to 3.92.

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.

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 32 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:

```
mtcars$mpg %>% mean()
```

```
## [1] 20.09062
```

```
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. In a report we can state:

> "In a our sample of 32 cars, we found a mean mileage of 20.1 miles per gallon. The 99% confidence interval for the mean miles per gallon in the population of cars runs from 17.2 to 23.0 miles per gallon."

or, somewhat shorter:

> "Based on our sample of 32 cars, we found an estimate for the mean mileage in the population of 20.1 miles per gallon (99% CI: 17.2, 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 population mean has the value 3.38 (the Tanzanian mean). We write that as follows:

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

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]

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.

[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 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.

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 3.55. 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?

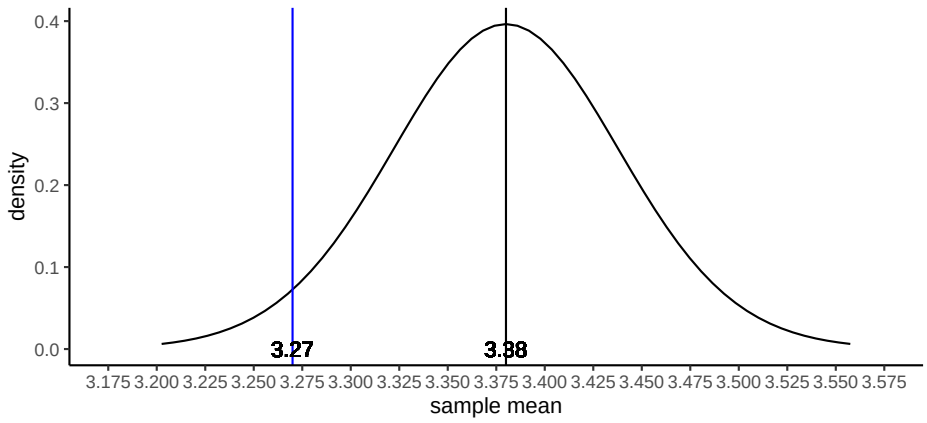


Figure 2.12: The sampling distribution under the null-hypothesis that the South-African population mean is 3.38. The blue line represents the sample mean for our observed sample mean of 3.27.

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 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 3.38, 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

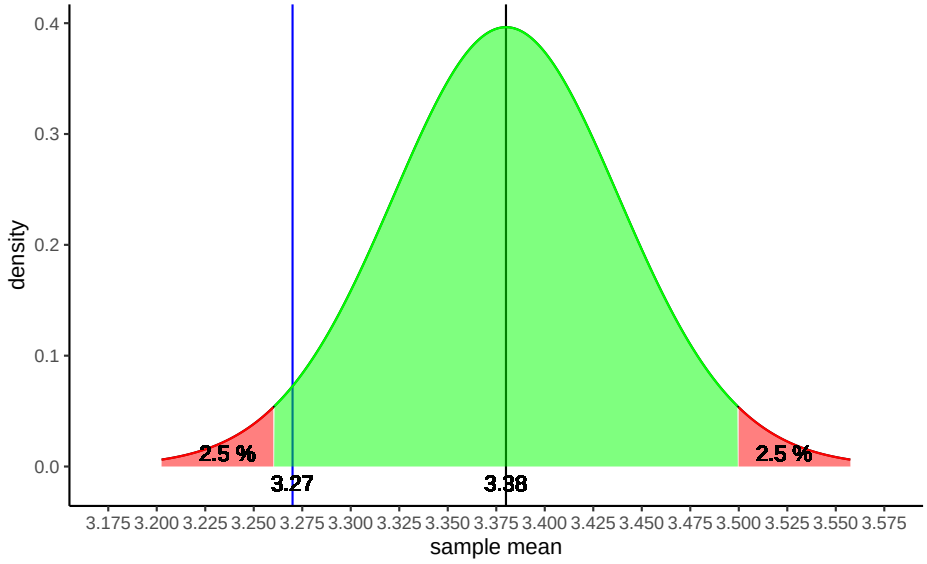


Figure 2.13: The sampling distribution under the null-hypothesis that the South-African population mean is 3.38. The red area represents the range of values for which the null-hypothesis is rejected (rejection region), the green area represents the range of values for which the null-hypothesis is not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

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 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:

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

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.

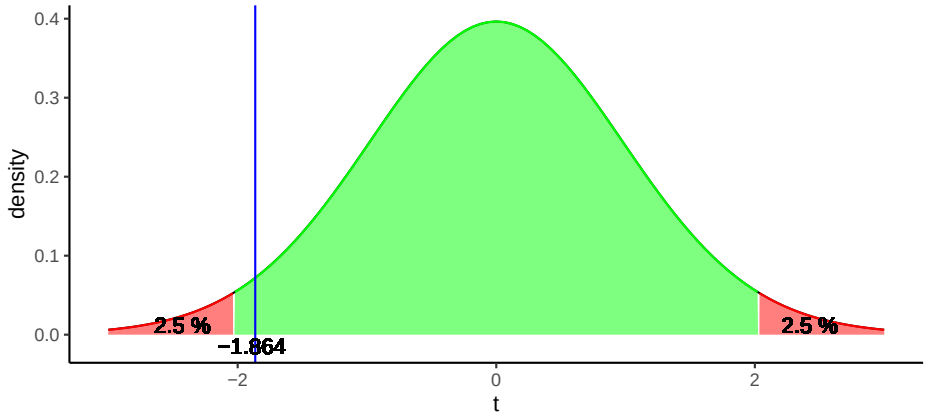


Figure 2.14: A *t*-distribution with 39 degrees of freedom to test the null-hypothesis that the South-African population mean is 3.38. The blue line represents the *T*-score for our observed sample mean of 3.27.

Table 2.3: Quantiles for the standard normal and several t-distributions with varying degrees of freedom.

		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

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 2.3.
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: $\sqrt{\frac{0.15}{10}} = 0.122$
2. Calculate the t-statistic: $t = \frac{3.28-3.38}{0.122} = -0.82$
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 -0.82 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 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?

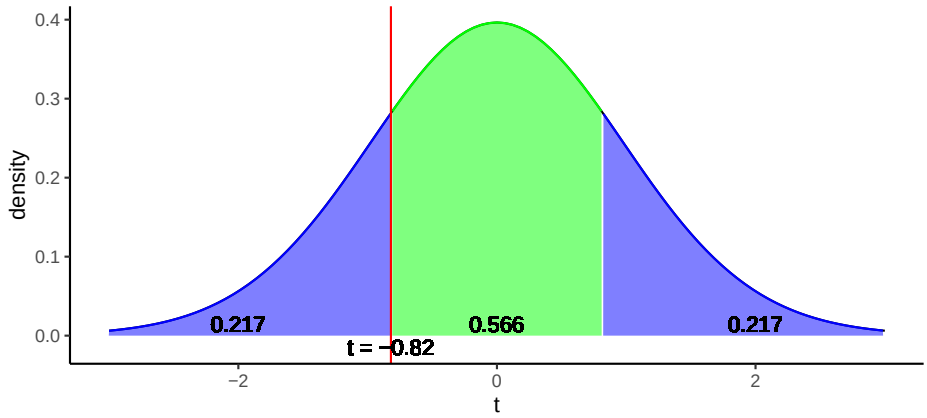


Figure 2.15: Illustration of what a *p*-value is. The total blue area represents the probability that under the null-hypothesis, you find a more extreme value than the *t*-score or its opposite. The blue area covers a proportion of 0.217 + 0.217 = 0.434 of the *t*-distribution. This amounts to a *p*-value of 0.434.

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 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 (i.e., 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.

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:

$$H_0 : \mu = 2.54$$

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 happens to be $s^2 = 0.304$ and we had 48 measures, so we estimate the standard error as $\hat{\sigma}_{\bar{y}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{0.304}{48}} = 0.080$. We then apply standardisation to get a t-value:

$$t = \frac{2.40 - 2.54}{0.080} = -1.75$$

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.

When we report the results from a null-hypothesis test, we often talk about *significance*. When the results show that the null-hypothesis can be rejected, we talk about a *statistically significant* result. When the null-hypothesis cannot be rejected, we call the results non-significant. Note that this says nothing about the importance or size of the results.

In a report, you could state:

> "We tested the null-hypothesis that the mean LH level is 2.54. Taking a sample of 48 measurements, we obtained a mean of 2.40. A t-test showed that this sample mean was not significantly different from 2.54, $t(47) = -1.75, p > .05$."

2.14 Null-hypothesis testing using R

We can do the null-hypothesis testing also with R. 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. The difference is due to our rounding of the sample variance. The sample variance s^2 can be obtained by

```
var(lh)
```

```
## [1] 0.3042553
```

We see that the number of degrees of freedom is 47 $(n - 1)$ and that the p-value equals 0.085. 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.

When reporting a null-hypothesis test, the convention is to report the exact p-value. We can omit the 0 in front of the decimal sign since a p-value is always between 0 and 1 (i.e., .085 instead of 0.085). Three decimals are usually more than enough.

> "We tested the null-hypothesis that the mean LH level is 2.54. Taking a sample of 48 measurements, we obtained a mean of 2.40. A t-test showed that this sample mean was not significantly different from 2.54, $t(47) = -1.758, p = .085$."

2.15 One-sided versus two-sided testing

In the previous sections, 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:

$$H_0 : \mu = 2.54$$
$$H_A : \mu \neq 2.54$$

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:

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

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:

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

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. 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

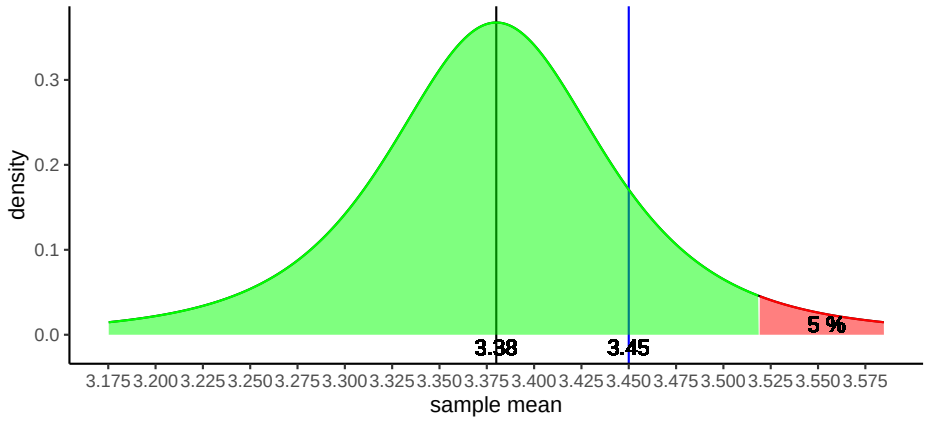


Figure 2.16: The sampling distribution under the null-hypothesis that the South-African population mean is 3.38. In one-tailed testing, the rejection area is located in only one of the tails. The red area represents the range of values for which the null-hypothesis is rejected (rejection region), the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

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

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

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

2.16 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

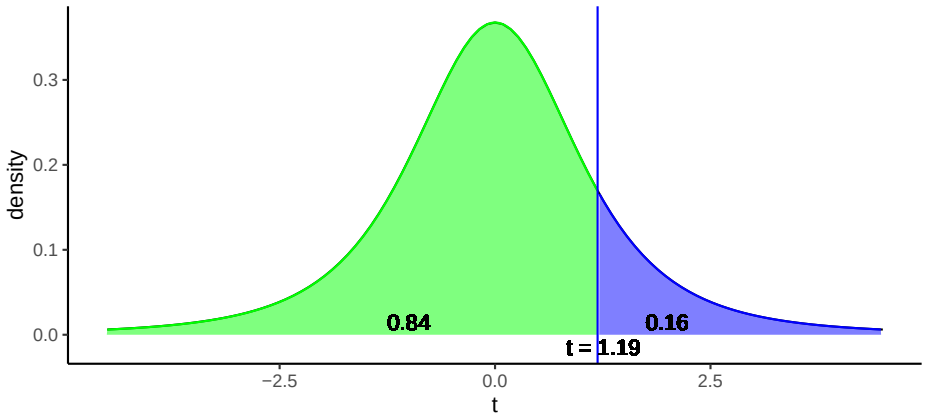


Figure 2.17: The sampling distribution under the null-hypothesis that the South-African population mean is 3.38. For one-tailed testing, the rejection area is located in only one of the tails. The green area represents the probability of seeing a *t*-value smaller than 1.19, the blue are represents the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

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:

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

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

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

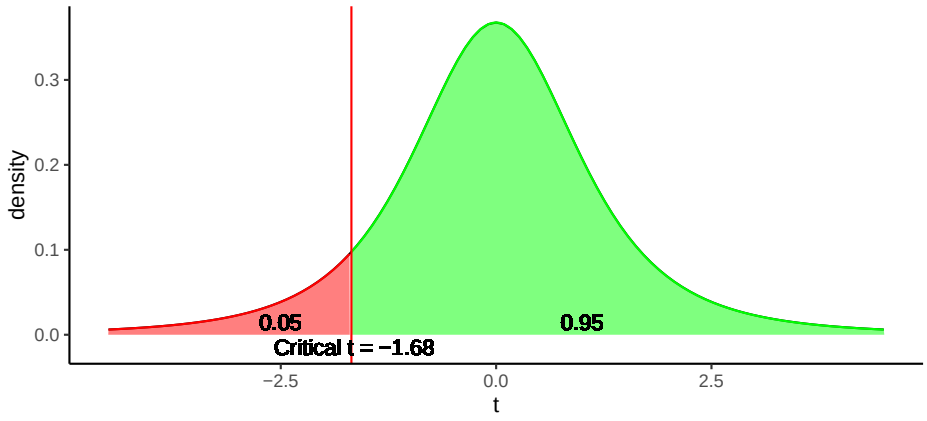


Figure 2.18: One-tailed decision process for deciding whether the average LH level in a woman is too low.

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.

2.17 One-tailed testing using R

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:

```r
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]:

```r
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
```

With one-tailed testing, you can report:

> "We tested the null-hypothesis that the mean LH level is 2.54, against the alternative hypothesis that the mean LH level is less than 2.54. Taking a sample of 48 measurements, we obtained a mean of 2.40. A one-tailed t-test showed that this sample mean was significantly less than 2.54, $t(47) = -1.758, p = .043$."

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 (.04259 vs .08518). Based

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

on the previous sections, you should know why the p-value is halved! (Because the t-distribution is symmetrical and you placed all of the 5% rejection area only on the left of the t-distribution). 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.

2.18 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 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.

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

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. 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 3.26 or larger than 3.50, 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: $1 - 0.101 = 0.899$.

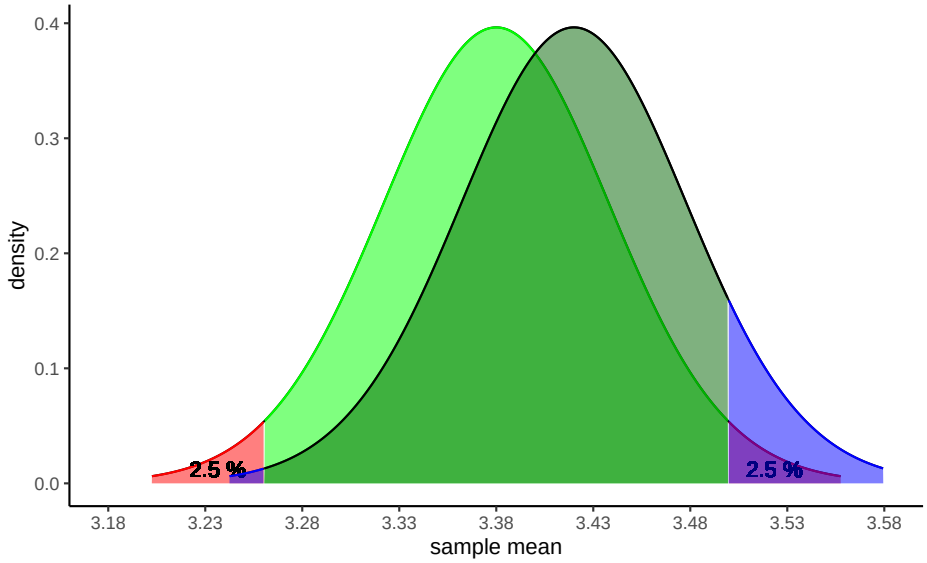


Figure 2.19: Two sampling distributions, one for a population mean of 3.38 (null-hypothesis) and one for a population mean of 3.42 (alternative hypothesis). The red areas represent the probability of a type I error, the dark green area the probability of a type II error. The blue areas represent the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

When we have to draw a conclusion 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

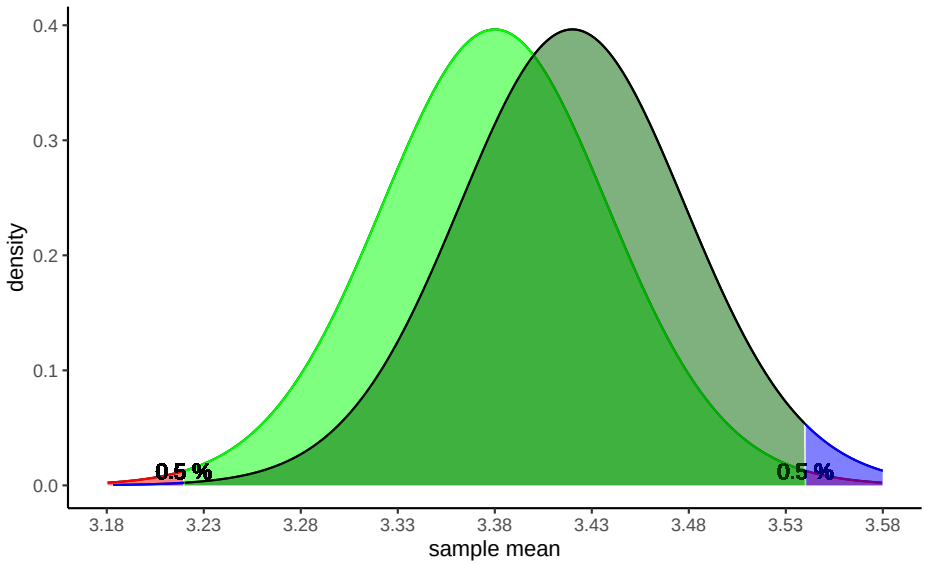


Figure 2.20: Two sampling distributions, one for a population mean of 3.38 (null-hypothesis) and one for a population mean of 3.42 (alternative hypothesis). The red areas represent the probability of a type I error, the dark green area the probability of a type II error. The blue area represents the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

think about this in the context of our luteinising hormone problem.

We saw that if the LH level is low, this could be 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 normal, 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 in this case, and what would be a 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 minimise 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 minimise 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 defend 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.

2.19 Take-away points

- In statistics, *inference* refers to drawing conclusions about population (complete) data on the basis of sample data (a selection of data).
- When you randomly draw equally-sized samples from a population, and for each sample you compute the mean, you can make a histogram of all sample means. This histogram represents the sampling distribution of the sample mean.
- When you randomly draw equally-sized samples from a population, and for each sample you compute the variance, you can make a histogram of all sample variances. This histogram represents the sampling distribution of the sample variance.
- The sample mean is an unbiased estimator of the population mean.
- The sample variance is a biased estimator of the population variance.
- The standard deviation of the sampling distribution is called the standard error.
- The larger the sample size, the smaller the standard error.
- 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.
- 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 shape of the sampling distribution is a t-distribution. The shape of this t-distribution depends on sample size (expressed as degrees of freedom).
- The objective of null-hypothesis testing is that we either reject the null-hypothesis, or not. This is done using the data from a random sample. In the null-hypothesis procedure, we assume that the null-hypothesis is true, and compare the sample data with data that would result if the null-hypothesis were true.

- The p-value represents 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.
- In two-tailed testing, we have rejection regions on both sides of the t-distribution. In one-tailed testing, we have only one rejection region.
- A type I error is the mistake of rejecting the null-hypothesis while it is in fact true.
- A type II error is the mistake of not rejecting the null-hypothesis while it is in fact not true.

Key concepts

- Inference
- Population
- Sample
- Random sample
- Standard error
- Sampling distribution
- Central Limit Theorem
- Confidence interval
- t-statistic
- t-distribution
- Degrees of freedom
- Null-hypothesis testing
- Rejection region
- Acceptance region
- Critical value
- p-value
- One-sided and two-sided testing
- Alternative hypothesis
- Type I and type II errors

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.

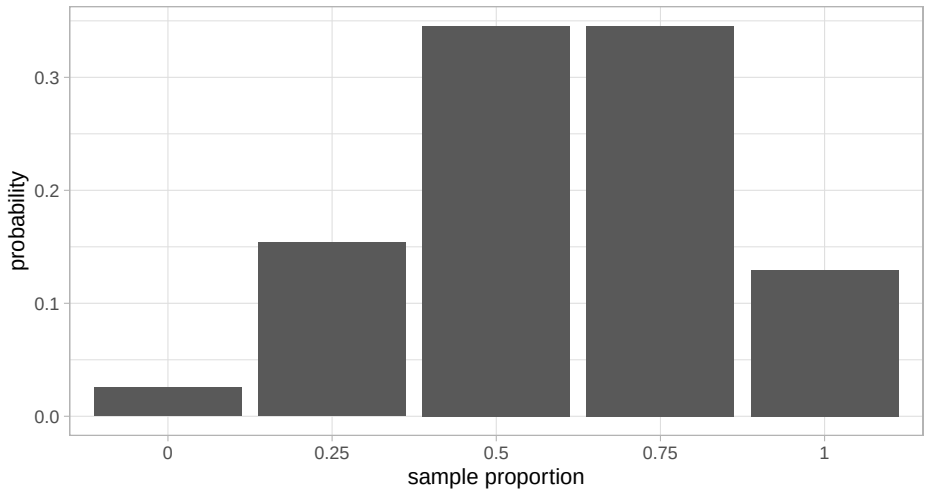


Figure 3.1: Sampling distribution of the sample proportion, when the population proportion is 0.60.

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 (advanced)

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 $0.6 \times 0.6 \times 0.6 \times 0.6 = 0.60^4 = 0.130$. The probability that all 4 elephants are shorter than 3.40 m is equal to $(1 - 0.6)^4 = 0.4^4 = 0.026$.

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 $0.6^2 \times (1-0.6)^2 = 0.058$, 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*:

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

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:

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

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

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

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 $0.6^2 \times (1-0.6)^2 = 0.058$. 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 0.058. 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 $0.058 + 0.058 + 0.058 + 0.058 + 0.058 + 0.058 = 6 \times 0.058 = 0.348$. 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

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	0.058
ABAB	0.6 x 0.4 x 0.6 x 0.4	0.058
ABBA	0.6 x 0.4 x 0.4 x 0.6	0.058
BAAB	0.4 x 0.6 x 0.6 x 0.4	0.058
BABA	0.4 x 0.6 x 0.4 x 0.6	0.058
BBAA	0.4 x 0.4 x 0.6 x 0.6	0.058

$$p(\#A = 2|n = 4, p = 0.6) = \binom{4}{2} \times 0.6^2 \times (1 - 0.6)^2 = 6 \times 0.058 = 0.348$$

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 0.348.

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

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

For example, the probability of obtaining 3 tall elephants when the total number of elephants is 4, is $\binom{4}{3} \times 0.6^3 \times (1 - 0.6)^1 = 4 \times 0.216 \times 0.4 = 0.346$.

When we calculate the probabilities of finding 0, 1, 2, 3, or 4 tall elephants in a 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.

3.3 Confidence intervals (advanced)

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

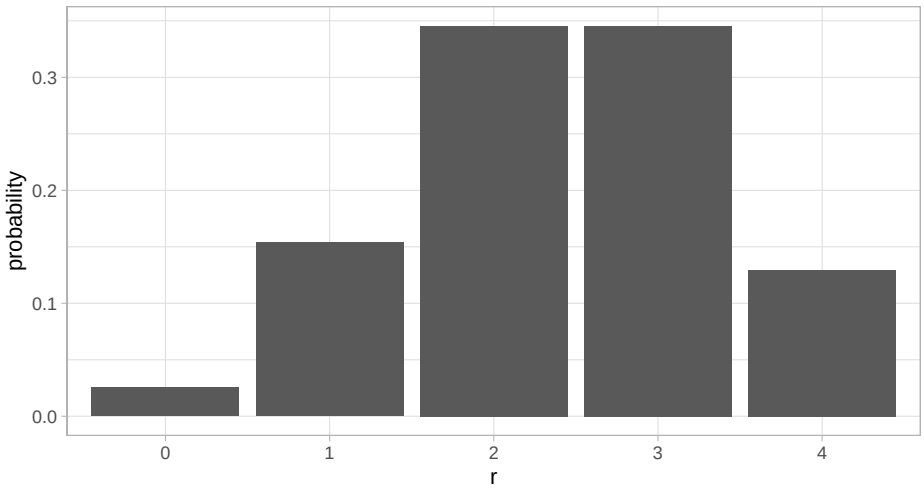


Figure 3.2: Binomial distribution with $n = 4$ and $p = 0.60$.

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):

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

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

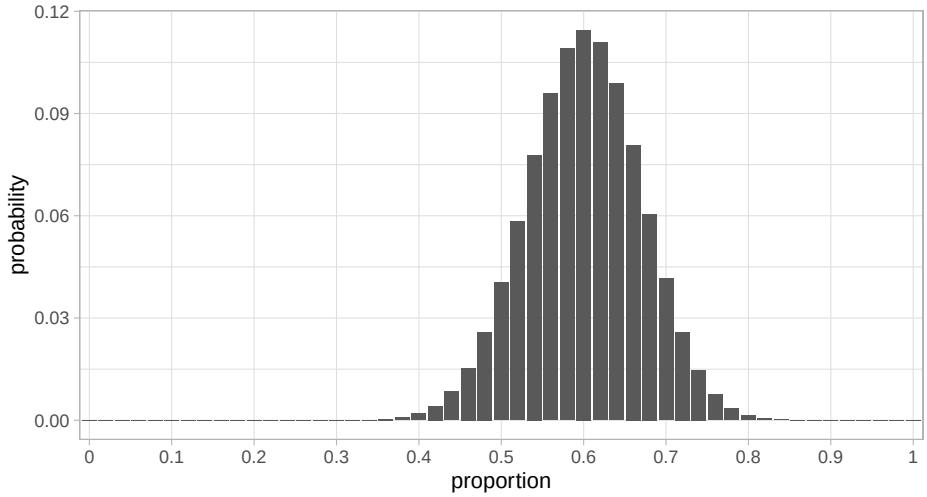


Figure 3.3: Sampling distribution with $n = 50$ and $p = 0.60$.

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 $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.052$.

If we use that value, we get the interval from $0.84 - 1.96 \times 0.052$ to $0.84 + 1.96 \times 0.052$: thus, our 95% confidence interval for the population proportion runs from 0.738 to 0.942.

3.4 Null-hypothesis concerning a proportion using the Central Limit Theorem

To do statistical computations by hand using the binomial distribution is often tedious. Because of the central limit theorem, we know that for large sample sizes, the sampling distribution becomes normal. The mean of the sampling distribution is then equal to the proportion p in the population, and the standard error is equal to $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. If the sample size, n, is large, we can perform null-hypothesis testing using the more familiar normal distribution.

Here we show how to do that with an example using a somewhat larger sample size of 50. 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 zoo elephants bump their head and are therefore taller than 3.40. How can we test the hypothesis that the zoo elephants could be a random 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 of elephants that the zoo elephants were drawn from has a proportion of 0.6 of elephants taller than 3.40 m, and the alternative hypothesis that it this proportion is not equal to 0.60.

$$H_0 : p = 0.60$$
$$H_A : p \neq 0.60$$

Is the proportion of 0.84 that we observe in the sample (the zoo) a probable value to find if the proportion for all elephants 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. 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 zoo elephants were not randomly drawn from the population of Tanzanian elephants.

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 -1.64 and 1.64 (see Table 2.3). If we therefore standardise our proportion, we have a measure that should show a standard normal distribution:

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

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

$$z_p = \frac{0.84 - 0.6}{se} = \frac{0.24}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.24}{0.069} = 3.478$$

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 randomly drawn from the Tanzanian population data.

The decision process is illustrated in Figure 3.4.

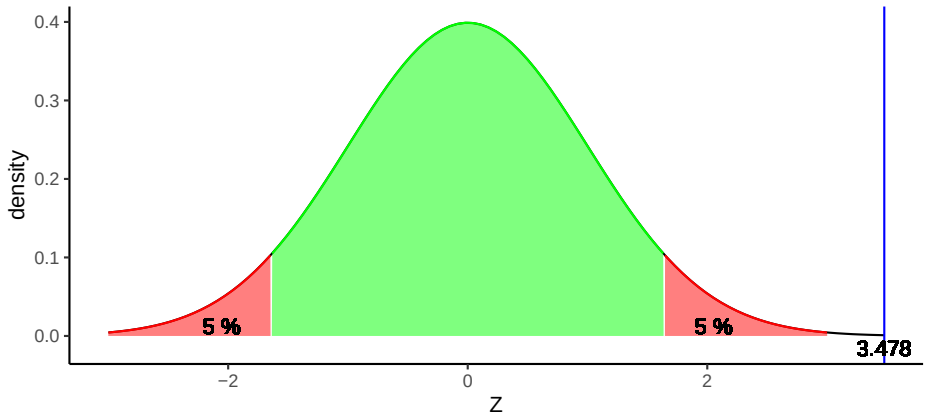


Figure 3.4: A normal distribution to test the null-hypothesis that the population proportion is 0.6. The blue line represents the *z*-score for our observed sample proportion of 0.84.

To get a feel for what critical z-values go together with what type I error rate, try out the interactive app in Figure 3.5. Choose a type I error rate of 10% and check that the critical z-value will be ± 1.64. Next choose a type I error rate of 5% and check that the critical z-value will become ± 1.96. Note that we apply two-tailed (two-sided) hypothesis testing here.

Please Wait

Figure 3.5: [Interactive] The relationship between the type I error rate α, and the critical z-values to reject the null-hypothesis for a proportion. Change the type I error rate, and see how that affects the critical z-value.

3.5 Inference on proportions using R

Using the normal distribution as shown in the previous section is a nice trick when you have to do the calculations by hand. However, the normal distribution

is only a good approximation of the binomial distribution when you have a large sample size. Using the binomial distribution always gives you the most exact answers, but 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 observed 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 most reasonable values for the population proportion are, we look at the 95% confidence interval that runs from 0.71 to 0.93. 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 .0004. Using only three decimals, we can say the p-value is less than .001. With an α of 0.10, this test is significant. We therefore conclude:

> "With a binomial test, we tested the null-hypothesis that the population proportion of elephants taller than 3.40 m is equal to 0.60, with an α of 0.10. Our sample proportion, based on 50 elephants, was 0.84, which is significantly different from 0.6, $p < .001$. We therefore reject the null-hypothesis."

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 0.194 to 0.994. Thus, based on this sample proportion of 0.75, we can state with some degree of confidence that the population proportion is somewhere between 0.194 and 0.994. That's of course not very informative, which makes sense considering we only observe 4 elephants.

We could report:

> "In our sample of 4 elephants, 3 were taller than 3.40 m. The 95% confidence interval for the proportion of elephants in the population that are taller than 3.40 m runs from 0.194 to 0.994."

or, somewhat shorter:

> "Based on a sample of 4 elephants, our estimate for the proportion of elephants in the population that are taller than 3.40 m is 0.75 (95% CI: 0.194, 0.994)."

3.6 Take-away points

- The sampling distribution of a sample proportion is closely related to the binomial distribution.

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

- With increasing sample size, the binomial distribution becomes normal, hence the sampling distribution of a sample proportion also becomes normal (Central Limit Theorem).

Key concepts

- Sampling distribution of a sample proportion
- Binomial distribution
- Binomial coefficient

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 (working with paint all day induces more health problems than working with students all day).

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

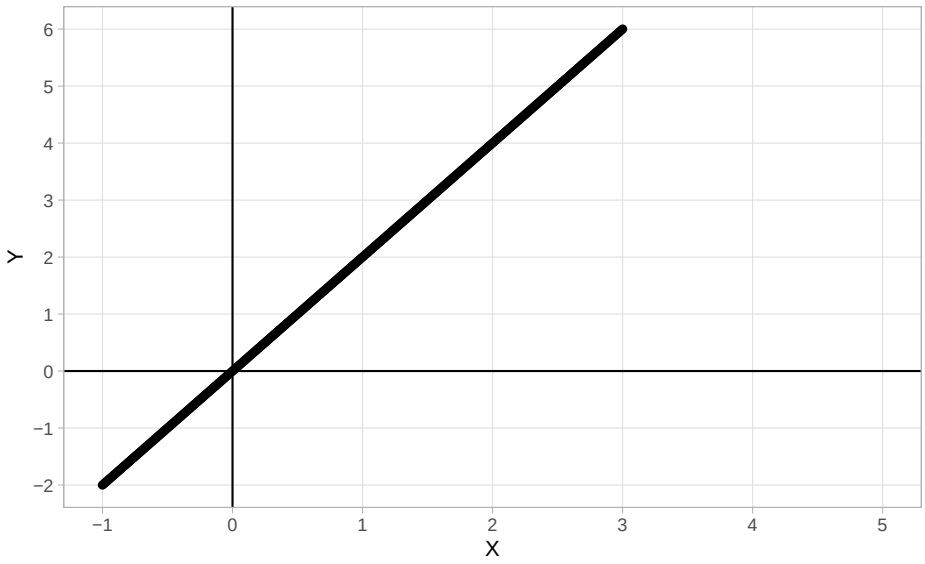


Figure 4.1: Straight line with intercept 0 and slope 2.

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 Figure 4.1, the slope is equal to 2.

The slope is one important feature of a straight line. The second important feature of a straight line is the *intercept*. The intercept is the value 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:

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

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

$$Y = 0 + 2X$$

which can be simplified to

$$Y = 2X$$

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.

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:

$$Y = -2 + 0.5X$$

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:

$$Y = -2 - 2X$$

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.

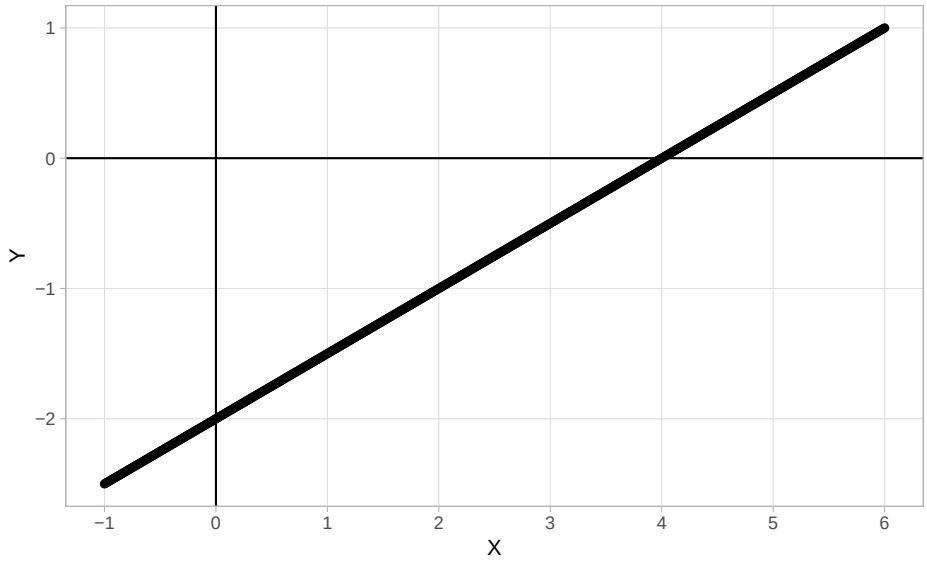


Figure 4.2: Straight line with intercept -2 and slope 0.5.

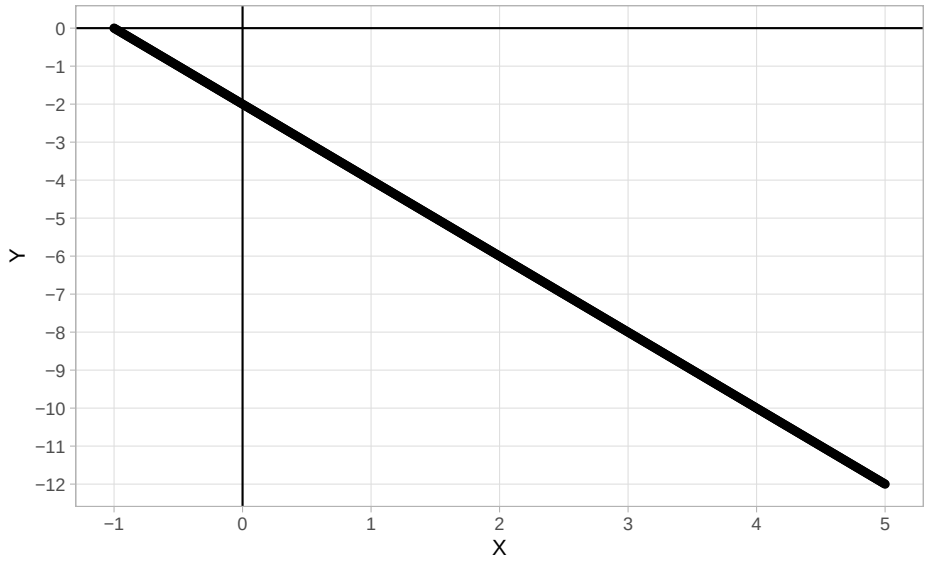


Figure 4.3: Straight line with intercept -2 and slope -2.

In social sciences we hardly ever see such perfectly linear relationships between quantities (variables). For instance, let us plot the relationship between yearly income and the amount of Euros spent on holidays. Yearly income is measured in thousands of Euros (kEuros), 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.

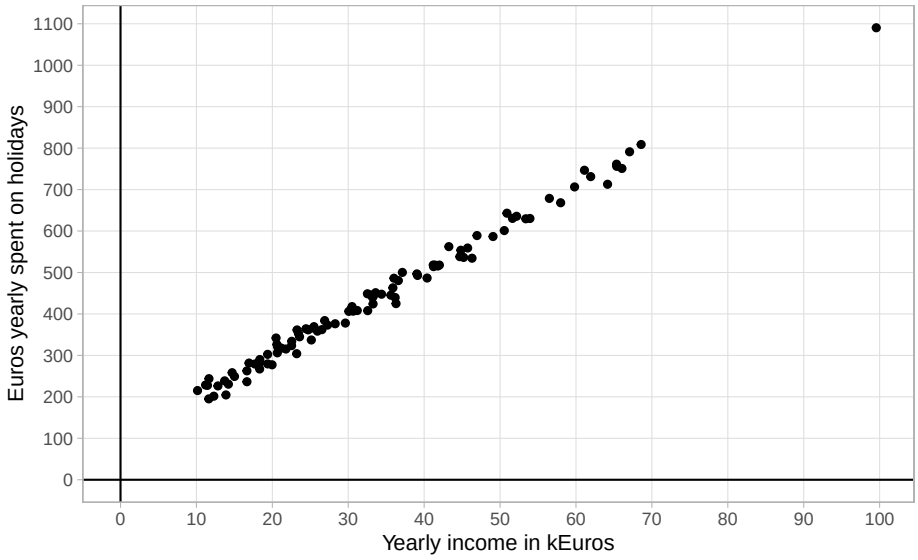


Figure 4.4: Data on holiday spending.

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 around 100.

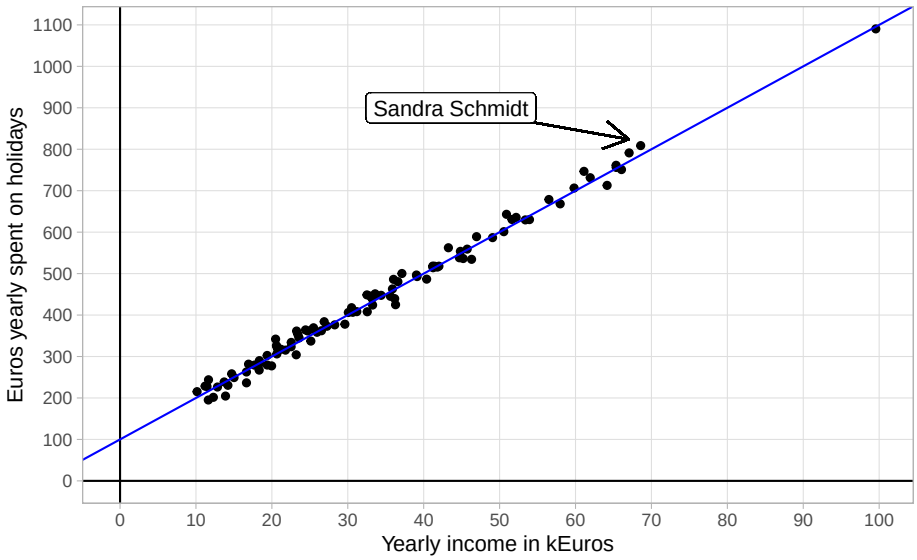


Figure 4.5: Data on holiday spending with an added straight line.

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

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

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 kEuros 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.

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 $\widehat{Y_i}$:

$$e_i = Y_i - \widehat{Y_i}$$

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

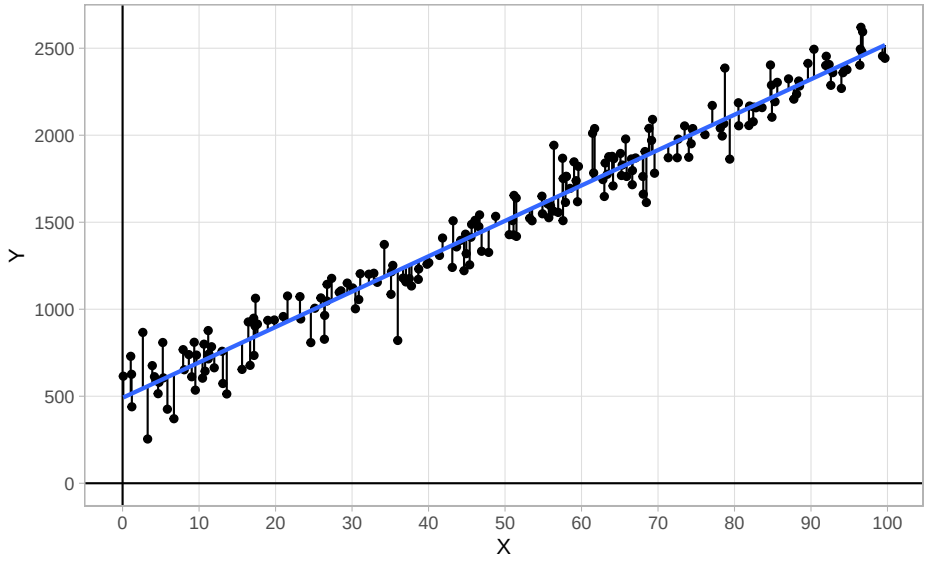


Figure 4.6: Data on variables X and Y with an added straight line.

If we compute residual e_i for all Y-values in the data set, we can plot them using a histogram, with a density plot that smooths the histogram to convey

the general pattern, see Figure 4.7. We see that the residuals are on average 0, and that the histogram is symmetrical. 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. Here, the residuals show a distribution with mean 0 and variance of 13336 (i.e., a standard deviation of 115).

The general assumption of linear models is that the distribution of the residuals is a normal distribution. Sometimes you see a distribution of residuals that is clearly normal, sometimes you don't. It's a bit hard to see from a rough histogram like this, but since the distribution is more or less symmetrical and shows a bell-shaped form, the normality assumption seems reasonable.

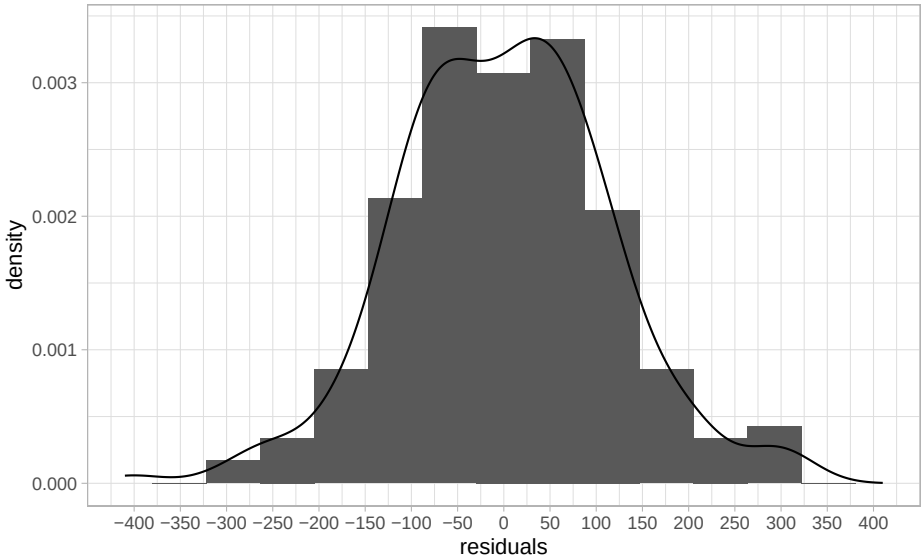


Figure 4.7: Histogram of the residuals (errors).

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.

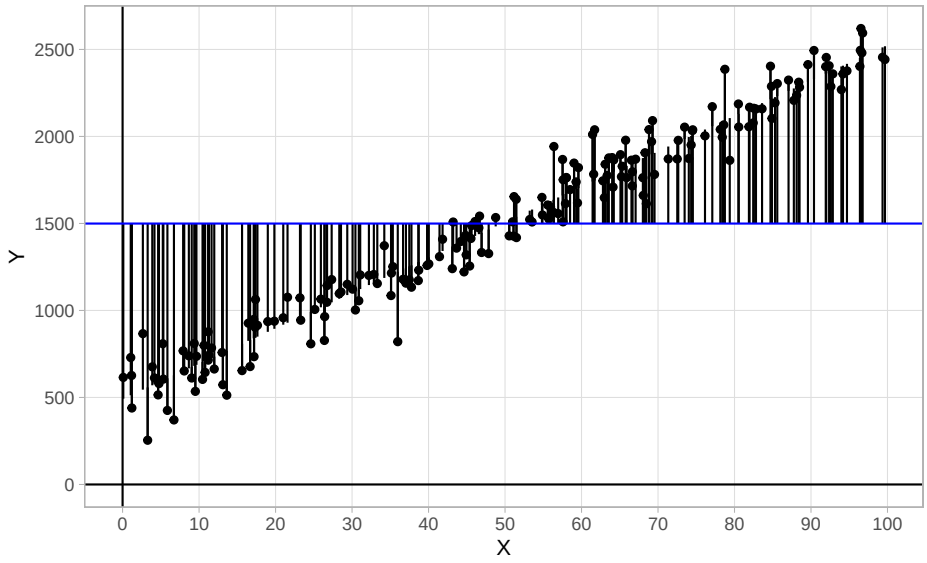


Figure 4.8: Data on variables X and Y with an added straight line. The sum of the residuals equals 0.

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.

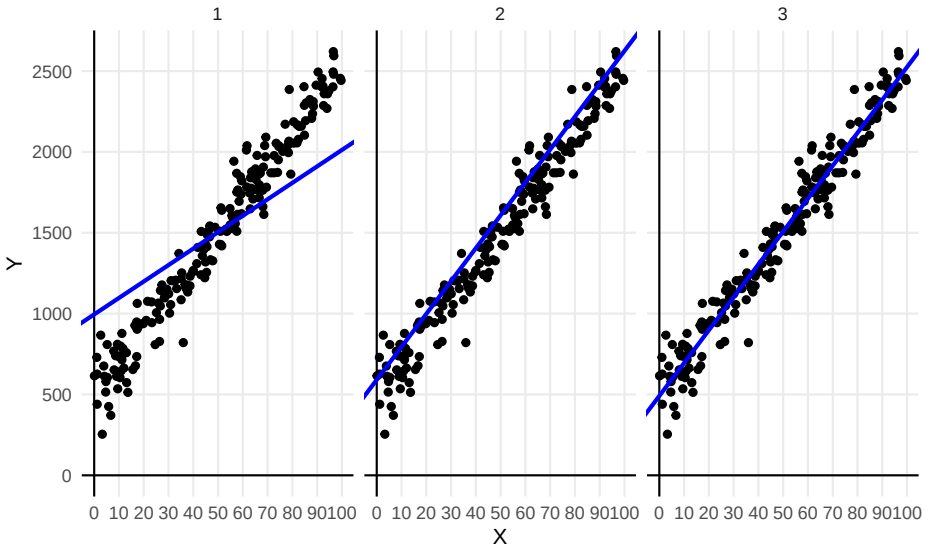


Figure 4.9: Three times the same data set, but with different regression lines.

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

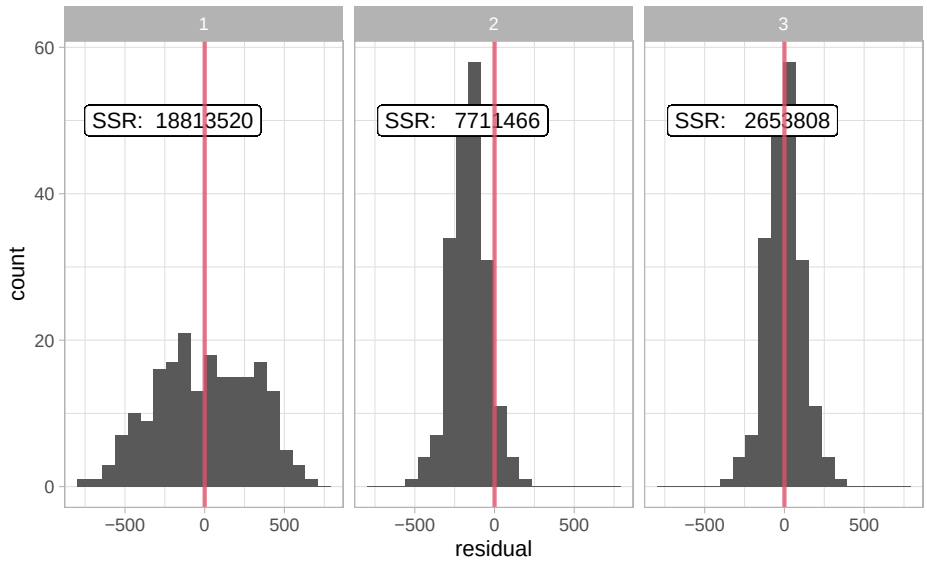


Figure 4.10: Histogram of the residuals (errors) for three different regression lines, and the respective sums of squared residuals (SSR).

least squares or OLS regression line.

Technical details

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:

$$\text{slope} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \tag{4.1}$$

and the Least Squares intercept can be found by:

$$\text{intercept} = \bar{Y} - \text{slope} \times \bar{X}$$

where $\bar{X}$ and $\bar{Y}$ 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.

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

$$\widehat{Y_i} = b_0 + b_1 X_i$$

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

$$Y_i = \widehat{Y_i} + e_i = b_0 + b_1 X_i + e_i$$

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

$$Y_i = b_0 + b_1 X_i + e_i \tag{4.2}$$
$$e_i \sim N(0, \sigma_e^2) \tag{4.3}$$

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.

Note that the linear model prescribes that the residual always comes from the same normal distribution with the same variance parameter $N(0, \sigma_e^2)$. It doesn't matter whether the predicted Y-value equals 10, 100 or -100: we expect that the errors (the residuals) always show a distribution that is normal with the same variation around the predicted variable. The prescription that the residuals show a normal distribution is called the normality assumption. The assumption that the variation around the predicted value is always of the same magnitude is called homogeneity of residual variance. Both are important assumptions when applying linear models. We will come back to them in Chapter 7.

The model prescribes not only that the residuals come from a normal distribution, but also that they are random draws from the normal distribution. That means that whether a residual is positive or negative, and whether it is large or small, is completely unpredictable. The model equation describes the predictable part of the data, the normal distribution of the residuals the unpredictable part. If in some way, there is some systematic pattern in the residuals, the model is incorrect. Any systematic pattern in the residuals is called *dependency*, which will also be discussed further in Chapter 7.

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 random 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 (4.3) 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 with a single variance parameter σ_e^2. 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.

Note on the term *linear*

We saw thus far that linear models and linear regression are about straight lines. But strictly speaking, the term linear in linear model refers to the fact that we use a *linear equation* in the form of $b_0 + b_1 X$. In the linear equation, we see that the increase in Y is predicted by changes in X. For every change in X, we expect to see a change in Y: we *add* something to the predicted Y. For every change of 1 in X, we add the value of the slope to the predicted value of Y. In linear regression, a basic form of a linear model, this simple idea results in a perfect straight line between the predictor X and the dependent variable Y. However, for more sophisticated linear models that we discuss later in this book, we will see relationships that are not straight lines.

Linear models are called linear because they are based on linear equations with an intercept and a slope, not because they result in straight lines for the relationship between X and Y.

4.7 Linear regression in R

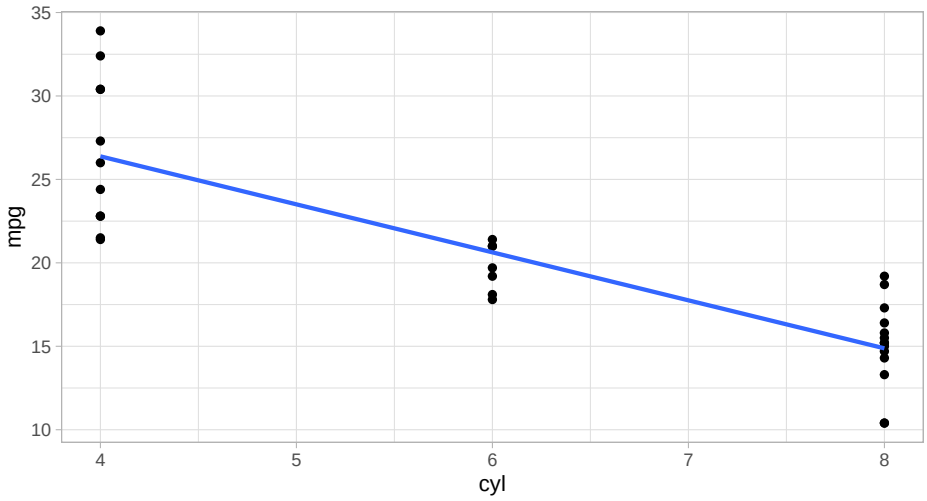


Figure 4.11: Data set on number of cylinders (cyl) and miles per gallon (mpg) in 32 cars.

Figure 4.11 shows the relationship between the number of cylinders (**cyl**) and miles per gallon (**mpg**) in cars, based on a data set available in R under the name `mtcars`. 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 code we first indicate that we start from the `mtcars` data frame. 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 $\sim$ ('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 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 <- mtcars %>%
  lm(mpg ~ cyl, data = .)
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 (variable values vary within a data set, parameter values do not). Here we use the linear model from Equation (4.3) 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 `tidy()` function from 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:

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

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 model and the data can be plotted, as in Figure 4.11, using the code:

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

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: $Y = 100 + 10X$.

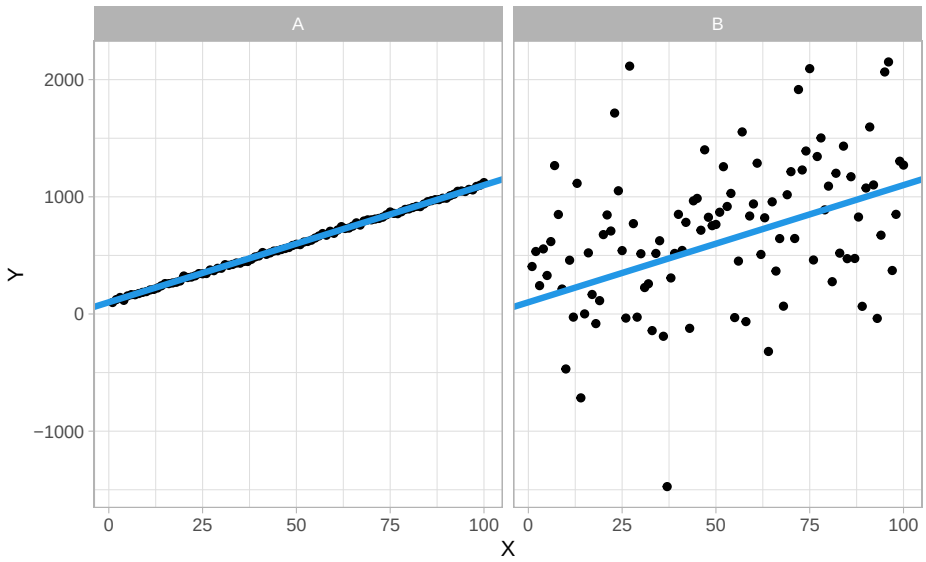


Figure 4.12: Two data sets with the same regression line.

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

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

$$z_X = \frac{x - \bar{X}}{\sigma_X}$$

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

$$z_Y = \frac{y - \bar{Y}}{\sigma_Y}.$$

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

$$\begin{aligned} Z_Y &= 0 + 0.997 \times Z_X \\ Z_Y &= 0 + 0.376 \times Z_X \end{aligned}$$

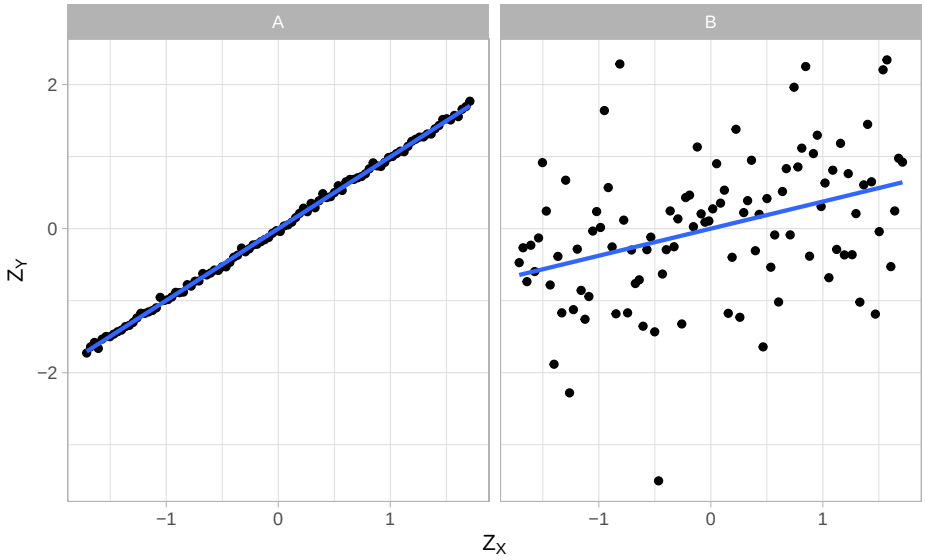


Figure 4.13: Two data sets, with different regression lines after standardisation.

These two slopes, the slope for the regression of standardised Y-values on standardised 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 using a regression line. 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.

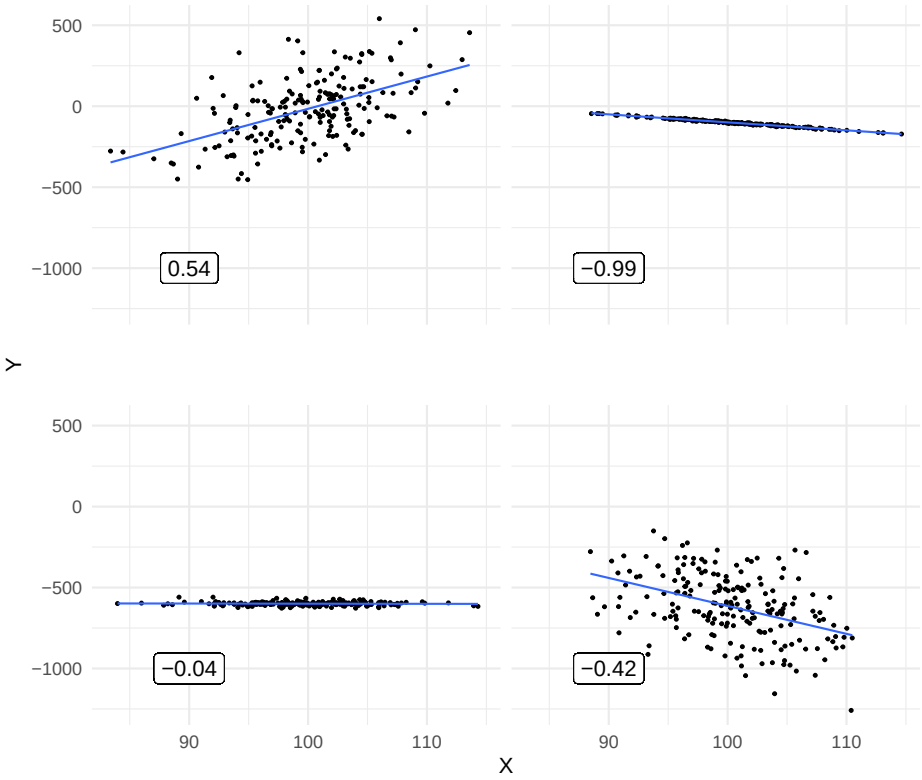


Figure 4.14: Various plots showing different correlations between variables X and Y.

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.

Interactive Figure 4.15 can be used to get an idea of what a correlation means in terms of the relationship between two variables. Change the correlation and see what happens. When the correlation is 1, the dots are on a perfect straight

line; when the correlation is 0, there is complete randomness.

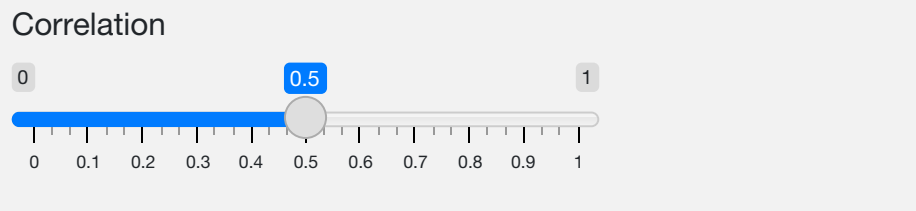


Figure 4.15: [Interactive] How well one variable can be predicted from a second variable can be quantified using the Pearson correlation. It is the slope of the linear regression when both variables are standardised. Change the value of the correlation, and see how that affects the relationship between the 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:

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

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.

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

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

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

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 (4.1). 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 $\sigma_X^2 = \frac{\sum(X_i - \bar{X})^2}{n}$ (see Chapter 1). We could therefore rewrite Equation (4.1) that finds the least squares or OLS slope as:

$$\begin{aligned} \text{slope}_{OLS} &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \\ &= \frac{\sigma_{XY} \times n}{\sigma_X^2 \times n} \\ &= \frac{\sigma_{XY}}{\sigma_X^2} \end{aligned} \tag{4.4}$$

This shows how all three quantities slope, correlation and covariance say something about the linear relationship between two variables. The slope

[1] Again, similar to what was said about the formula for the variance of a variable, on-line you will often find the formula $\frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$. 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 sum of the cross-products not by n but by $n - 1$.

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

X	Y	X - meanX	Y - meanY	Crossproduct
-1	2	-0.6	2.2	-1.32
0	-1	0.4	-0.8	-0.32
1	-2	1.4	-1.8	-2.52
-2	1	-1.6	1.2	-1.92
0	-1	0.4	-0.8	-0.32

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).

Numerical example of covariance, correlation and least square slope

Table 4.1 shows a small data set on two variables X and Y with 5 observations. The mean value of X is -0.4 and the mean value of Y is -0.2. If we subtract the respective mean from each observed value and multiply, we get a column of cross-products. For example, take the first row: $X - \bar{X} = -1 - (-0.4) = -0.6$ and $Y - \bar{Y} = 2 - (-0.2) = 2.20$. 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: 1.02 and 1.47. Now we can calculate the correlation on the basis of the covariance as $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{-1.28}{1.02 \times 1.47} = -0.85$.

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 4.16 together with the regression line. The standardised data and the corresponding regression line are plotted in Figure 4.17. Note that the slopes are different, and that the slope of the regression line for the standardised data is equal to the correlation.

4.10 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

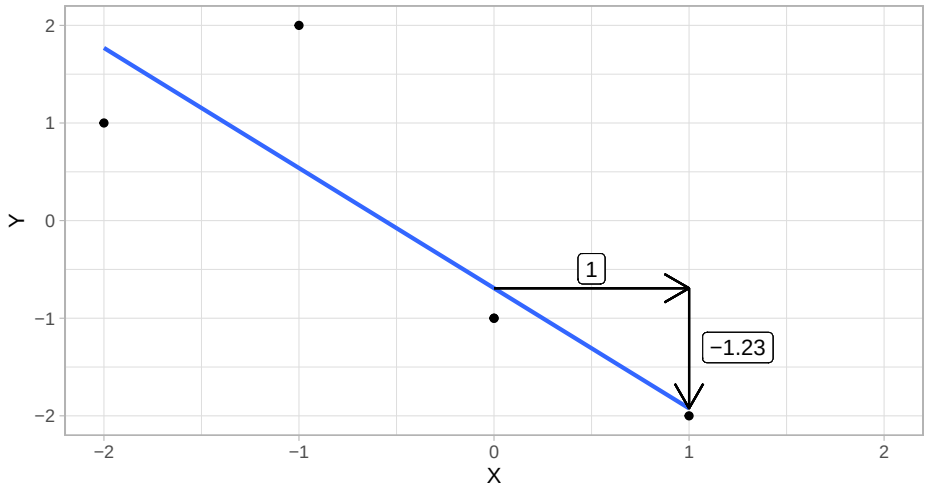


Figure 4.16: Data example and the regression line.

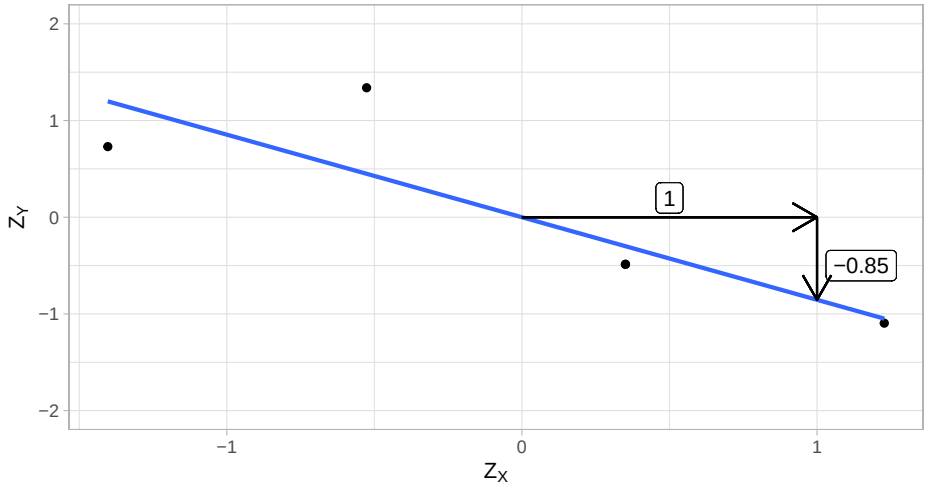


Figure 4.17: Data example (standardised values) and the regression line.

function `cor()`:

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

```
##            cyl       mpg
## 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 -0.85. 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 -9.17. On the diagonal we see the variances of **cyl** and **mpg**.

To determine the least squares slope for the regression line of **mpg** on **cyl**, we divide the covariance by the variance of **cyl** (Equation (4.4)):

```
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 same slope coefficient can of course also be obtained with a linear regression analysis using the standardised values:

```
tibble(z_mpg, z_cyl) %>%
  lm(z_mpg ~ z_cyl, data = .)
```

```
##
## Call:
## lm(formula = z_mpg ~ z_cyl, data = .)
##
## Coefficients:
##            (Intercept)                   z_cyl
##  0.0000000000000006083  -0.8521619594266127695
```

In research practice one often wants to do inference on a correlation: Is the correlation that we find in a sample indicative of a correlation in the population, and if so, how large would that population correlation be?

In the above example, the correlation between the number of cylinders in a car and the miles per gallon was -0.85. This is the sample correlation, based on the sample data available in the `mtcars` dataset. If we want to test the null-hypothesis that the population correlation between the number of cylinders and miles per gallon is 0, we can do that in R with `cor.test()`:

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

```
##
##  Pearson's product-moment correlation
##
## data:  mtcars$cyl and mtcars$mpg
## t = -8.9197, df = 30, p-value = 0.0000000006113
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9257694 -0.7163171
## sample estimates:
##       cor
## -0.852162
```

At the end of the output, you find the correlation again, -0.85. In addition, we see a null-hypothesis test in the form of a t-test and a 95% confidence interval. Based on this output we can report:

> "The correlation between the number of cylinders in a car and the miles per gallon is significantly different from 0, $t(30) = -8.92, p < 0.001$. The 95% confidence interval for the correlation in the population is between -0.93 and -0.72. "

4.11 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:

$$Y = b_0 + b_1 X$$

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 $\widehat{Y}$):

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

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:

$$Y = b_0 + b_1 X + e$$

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

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

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

$$e \sim N(0, \sigma_e^2)$$

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.

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

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:

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

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.

4.12 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 following 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:

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

In the data set, the variance of the weight, Var(`weight`) is equal to 72274. Since we also know the variance of the residuals (15362), we can solve for the variance explained by **volume**:

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

So the proportion of explained variance is equal to $\frac{56912}{72274} = 0.79$. 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:

$$\begin{aligned} \texttt{weight} &= 22.4 + 0.71 \times \texttt{volume} + 0.5 \times \texttt{area} + e \\ e &\sim N(0, 6031) \end{aligned}$$

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 $\frac{66243}{72274} = 0.92$. 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*.

4.13 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 $(200 - 80)/200 = 0.60$.

$$\begin{aligned} R^2 &= \sigma^2_{explained}/\sigma^2_Y \\ &= (\sigma^2_Y - \sigma^2_{unexplained})/\sigma^2_Y \\ &= (\sigma^2_Y - \sigma^2_e)/\sigma^2_Y \end{aligned} \tag{4.5}$$

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.

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. 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 residual 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!

This is a good thing, since we want to have a model that makes the smallest errors possible. However, there is also a danger that we are too optimistic drawing conclusions about the population based on only a limited data sample. Maybe the best linear regression in the sample is not the same as the best linear regression in the population data.

Because in any least squares regression analysis based on a sample of data, the explained variance is always maximised, we may overestimate the variance that is explained in the population data. In regression analysis, we therefore very often use an *adjusted R-squared* that takes this possible overestimation (*inflation*) into account. The adjustment is based on the number of independent variables and sample size.

The formula is

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where n is sample size and p is the number of independent variables.

The more independent variables you have (p), and the smaller your data set (n), the larger the difference between the R-squared and the adjusted R-squared.

Computational example

For example, if R^2 equals 0.60 and we have a sample size of 100, and 2 independent variables, the adjusted R^2 is equal to $1-(1-0.60)\frac{100-1}{100-2-1} = 1-(0.40)\frac{99}{97} = 0.59$. Thus, the estimated proportion of variance explained at population level, corrected for inflation, equals 0.59. Because R^2 is inflated, the adjusted R^2 is never larger than the unadjusted R-squared.

$$R^2_{adj} \leq R^2$$

4.14 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 code 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. Below we see the code and 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:

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

We can use this equation to make predictions about the weight of a particular book. For example, 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.

```
model %>% summary() %>%
  pluck("r.squared")
```

```
## [1] 0.9284738
```

```
model %>% summary() %>%
  pluck("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 (actually, the standard deviation sigma) of the residuals can also be found in the summary object, where it is called `sigma`. If you square that value, you get the variance of the residuals.

```
model %>% summary() %>%
  pluck("sigma") %>%
  .^2
```

```
## [1] 6031.052
```

4.15 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 for the relationship between the weight of a book and its volume and area:

Table 4.2: Part of Cape Fur Seal data.

age	weight	heart
33	27.5	127.7
10	24.3	93.2
10	22.0	84.5
10	18.5	85.4
12	28.0	182.0
18	23.8	130.0

$$
\begin{aligned}
\texttt{weight} &= b_0 + b_1 \times \texttt{volume} + e \\
\texttt{weight} &= b_0 + b_1 \times \texttt{area} + e \\
\texttt{weight} &= b_0 + b_1 \times \texttt{volume} + b_2 \times \texttt{area} + e
\end{aligned}
$$

If we carry out these three analyses, we obtain an R^2 of 0.8026 if we only use **volume** as predictor, and an R^2 of 0.1268 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 $0.8026 + 0.1268 = 0.9294$. However, if we carry out the multiple regression with **volume** and **area**, we obtain an R^2 of 0.9285, 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)
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
```

```
sigma <- out1 %>% summary() %>%
  pluck("sigma") %>%
  .^2 # taking the square
```

resulting in the equation:

$$\begin{aligned} \texttt{age} &= 11.4 + 0.82 \times \texttt{weight} + e \\ e &\sim N(0, 200) \end{aligned}$$

From the data we calculate the variance of **age**, and we find that it is 1090.86. The variance of the residuals is 200, so that the proportion of explained variance is $(1090.86 - 200)/1090.86 = 0.82$.

Since we also have data on the weight of the heart alone, we could try to predict the age from the weight of the heart. Then we get output

```
out2 <- 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
```

```
sigma <- out2 %>% summary() %>%
  pluck("sigma") %>%
  .^2 # taking the square
```

that leads to the equation:

$$\begin{aligned} \texttt{age} &= 20.6 + 0.11 \times \texttt{heart} + e \\ e &\sim N(0, 307) \end{aligned}$$

Here the variance of the residuals is 307, so the proportion of explained variance is $(1090.86 - 307)/1090.86 = 0.72$.

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
```

```
sigma <- out3 %>% summary() %>%
  pluck("sigma") %>%
  .^2 # taking the square
```

with model equation:

$$\begin{aligned} \texttt{age} &= 10.3 - 0.03 \times \texttt{heart} + 0.99 \times \texttt{weight} + e \\ e &\sim N(0, 204) \end{aligned}$$

Here we see that the regression parameter for **weight** has increased from 0.82 to 0.99. At the same time, the regression parameter for **heart** has decreased, has even become negative, from 0.11 to -0.03. From this equation we see that there is a strong relationship between the total weight and the age of a seal, but on top of that, for every unit increase in the weight of the heart, there is a very small decrease in the expected age. The slope for **heart** has become practically negligible, so we could say that on top of the effect of total weight, there is no remaining relationship between the weight of the heart and age. In other words, once we can use the total weight of a seal, there is no more information coming from the weight of the heart.

This is because the total weight of a seal and the weight of its heart are strongly correlated: heavy seals generally have heavy hearts. Here the correlation turns out to be 0.96, almost perfect! This means that if you know the total weight of a seal, you practically know the weight of its heart. This is logical of course, since the total weight is a composite of all the weights of all the parts of the animal: the total weight variable *includes* the weight of the heart.

Here we have seen, that if we use multiple regression, we should be aware of how strongly the independent variables are correlated. Highly correlated predictor variables do not add extra predictive power. Worse: they can cause problems in obtaining regression parameters because it becomes hard to tell which variable is more important: if they are strongly correlated (positive or negative), then they measure almost the same thing!

A situation where there are correlations among the independent variables is called *multicollinearity*. When two predictor variables are perfectly correlated, either 1 or -1, regression is no longer possible, the software stops and you get a warning. But also if the correlation is much smaller, you should be very careful interpreting the regression parameters. With strongly correlated variables, select the variable that makes most theoretical sense and consider to omit the other one.

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 theoretically 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. It also makes sense in terms of explained variance: **weight** on its own explains more variance (82%) than **heart** on its own (72%).

4.16 Simpson's paradox

With multiple regression, you may uncover very surprising relationships between two variables, that can never be found using simple regression. In the previous section, we saw that there is a general positive relationship between the weight of a seal's heart and its age. However, when we included the overall weight of the seal into the regression, we saw that suddenly the regression coefficient for the heart variable became small and changed direction from positive to negative. It is therefore not a trivial matter to decide which variables to add to a linear equation.

Here's an example from Paul van der Laken[2], who simulated a data set on the topic of Human Resources (HR). It nicely illustrates that you have to be very careful interpreting the regression coefficients in multiple regression, and linear models in general.

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).

[2] https://paulvanderlaken.com/2017/09/27/simpsons-paradox-two-hr-examples-with-r-code/

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:

$$\texttt{salary} = 45543 + 4912 \times \texttt{Neuroticism} + e$$

Figure 4.18 shows the data and the regression line. From this visualisation it looks like neuroticism relates *positively* to 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.

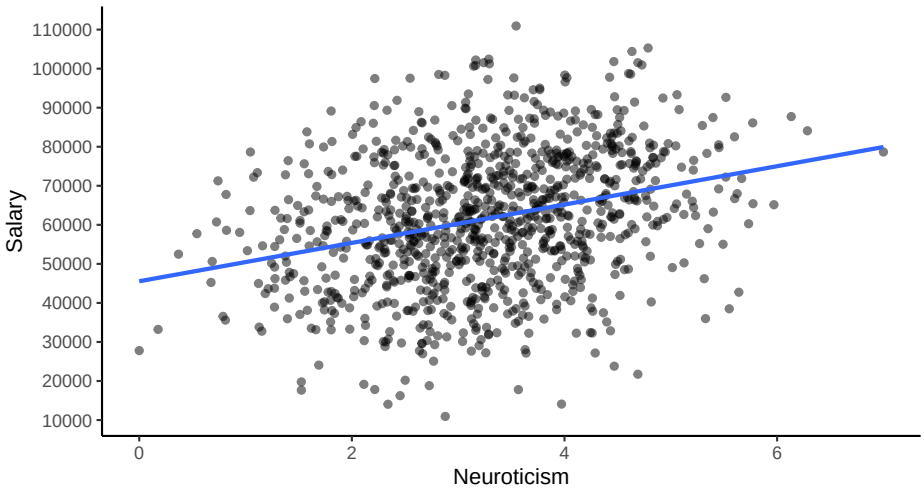


Figure 4.18: Simulated HR data set.

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:

$$\texttt{salary} = 50935 - 3176 \times \texttt{Neuroticism} + 20979 \times \texttt{Education} + e$$

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 pieces of 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 4.19: 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**.

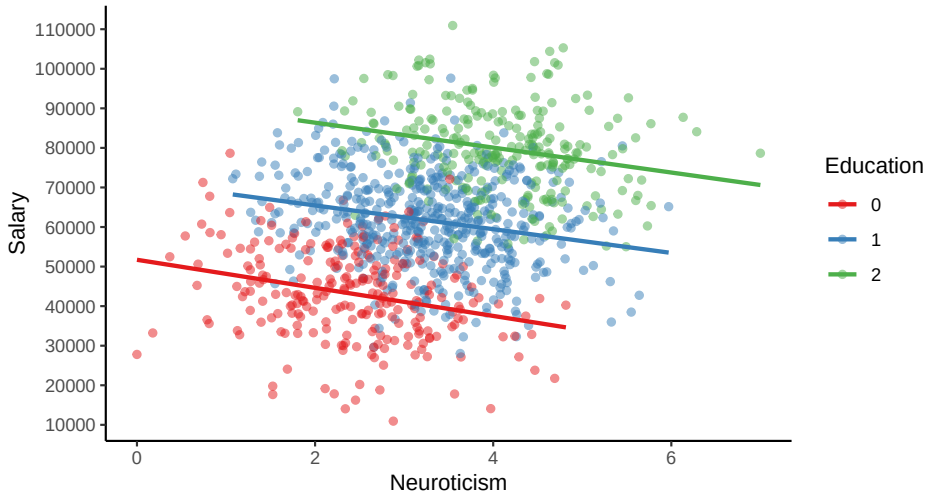


Figure 4.19: Same HR data, now with markers for different education levels.

The paradox is resolved when we look more closely at Figure 4.19. Look at the red dots in the bottom left of the graph: most of them have relatively low scores on the Neuroticsm scale (most of them less than 4). Now look at the green dots at the top right: most of them have relatively high scores on the Neuroticism scale (practically all of them more than 2). The blue dots show average Neuroticism scores. Thus, there seems to be a positive correlation between the level of education and neuroticism: employees with

higher education levels are also more neurotic. Again, similar to the previous section, we see that a correlation between independent variables can cause problems with the interpretation. Yes, there is a general tendency that neurotic people earn more money. But that is not due to their neuroticism. It is the educational background that explains the differences in salary, and because neuroticism tends to be correlated with educational background, there seems to be a positive relationship between neuroticism and salary. However, if you take into account education (i.e., put **Education** into the model as an extra predictor), the positive correlation disappears, even becomes negative. This negative correlation means that within the educational levels, there is a tendency that neurotic people earn less than non-neurotic people.

Diving deeper into multiple regression: confounders, mediators and colliders

Figure 4.20 shows the conceptual model of how the world might work: education has an effect on neuroticism: the higher the education level, the more neurotic people are. But neuroticism in and of itself also influences salary: given a certain education level, more neuroticism leads to a lower salary. Note that this is only a theory, you might have a better idea about how these variables interact in the world.

When noting the positive correlation between neuroticism and salary, you might conclude that more neurotic people earn more money than less neurotic people. But the figure shows that education completely explains that positive relationship: because high education people have higher neuroticism scores (an arrow pointing from education to neuroticism), and because high education people have bigger salaries (an arrow pointing from education to salary), there appears to be positive relationship between neuroticism and salary. The positive relationship is however completely explained by the effects of education. Education is called here a *confounder* variable: a variable that induces a correlation between two other variables. In general, it is always good to add a confounder variable to your multiple regression. Looking at the effect of neuroticism on salary and including education, we see that we get to the truth: that there is actually a *negative* correlation between neuroticism and salary when you look at the education levels separately.

However, when you are interested in establishing the effect of education on salary, we should not use the neuroticism variable in the regression. That is because neuroticism is a *mediator* variable in the model: it is in the middle between education and salary and therefore *mediates* the effect of education on salary. To some extent neuroticism explains the correlation between education and salary: education makes people more neurotic, and neuroticism has in turn a negative effect on salary, hence it *mediates* a negative effect of education on salary. In addition there is non-mediated or direct effect of education on salary that is positive and strong. Taken together, the weak negative relationship

mediated by neuroticism and the strong direct positive relationship result in a positive correlation. In general, in multiple regression, one should not add mediating variables in a regression analysis. Thus, when analysing the effect of education on salary, and you believe that the model in Figure 4.20 represents the state of the world, you should not include neuroticism into the regression analysis.

So far we talked about confounder and mediator variables. There is also a third kind: a *collider* variable. An example of a collider is when you are interested in the effect of education on neuroticism. If the model in Figure 4.20 is the correct one, both education and neuroticism have an effect on a third variable: salary. If you are interested in the effect of education on neuroticism, you should not add a variable that is affected by both these variables. An easy explanation of why you should not add collider variables into your regression analysis is based on a basketball example. If we look at the top 100 best professional basketball players, we might see that the smaller players are also the more skilled ones. This is not necessarily true for all people in the world that play basketball. In everyday life, there might be no connection between a person's height and their skill. It is because taller people have a higher chance of becoming a professional than shorter people, and because more skilled players have a higher chance of becoming a professional than less skilled ones, you end up with a *negative* correlation between height and skill when you only look at the professional players. That's because people who are not exceptionally tall, will have to be exceptionally skilled, whereas people who are not exceptionally skilled, will have to be exceptionally tall to become drafted as a professional and become one of the best. Returning to our employee data: if we are interested in the effect of education on neuroticism, and we control for salary, we will most likely see a correlation that we are not interested in. That's because people end up in the lowest pay scale either because they have no or very low levels of education, or if they have mental problems (high neuroticism score). People within each pay-scale therefore might show a correlation between neuroticism and education level that is not particularly informative about causation.

Thinking about *confounders*, *colliders* and *mediators* helps us think about the problem, and helps us decide in what cases to add a variable in our regression analysis, and when it is best not to. For more on these concepts, look into the literature on causal inference, for instance *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, by Morgan and Winship (2014), or *Causal Inference in Statistics: A primer*, by Pearl et al. (2016).

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

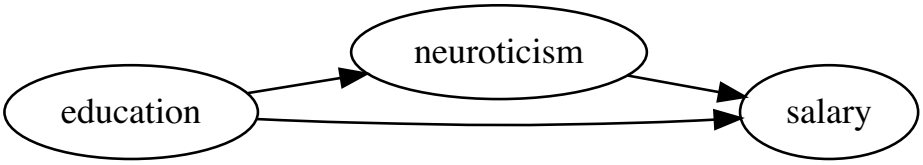


Figure 4.20: A path diagram for the assumed relationship between education, salary and neuroticism in the real world.

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 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. The spurious correlation is induced by the confounder variable temperature.

This spurious correlation is plotted in Figure 4.21. 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 the dark blue data points (low temperatures).

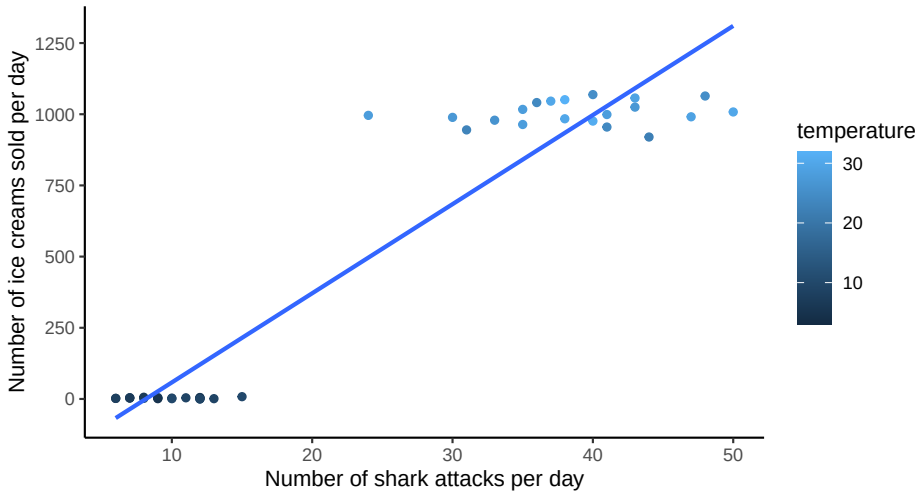


Figure 4.21: A spurious correlation between the number of shark attacks and ice cream sales.

4.17 Take-away points

- A simple linear equation represents a straight line.
- A straight line has an intercept and a slope.
- The intercept is the value of Y given that $X = 0$.
- Data usually do not show a straight line, but a straight line might be a good approximation (prediction model) for the data.
- Finding a reasonable straight line is called regression.
- A residual is the difference between an observed Y-value and the predicted Y-value.
- Finding the best regression line is usually based on the least squares principle.
- Correlation stands for the co-relation between two variables. It tells you how well one variable can be predicted from the other using a linear equation.
- Correlation is standardised to be between -1 and 1. It is the slope of the regression line for standardised X- and Y-values.
- Covariance is an unstandardised measure for the co-relation.

- Unexplained variance is the variance of the residuals.
- Explained variance is total variance of the dependent variable minus the residual variance.
- R-squared is the proportion of explained variance.
- It is possible to include more than one independent variable in a linear model. We then talk about multiple regression.
- It is not wise to include two independent variables that are highly correlated with each other. Having correlations among independent variables is called collinearity.
- The relationship between one independent variable and the dependent variable can change, depending on what other independent variables are included in the regression model. An extreme example of this is Simpson's paradox. Whether to include or exclude variables from a regression analysis should be based on the theoretical model that lies behind the analysis.

Key concepts

- Dependent and independent variables
- Intercept
- Slope
- Regression
- Residual
- Ordinary Least Squares
- Correlation coefficient
- Explained and unexplained variance
- R^2 (R-squared)
- Multiple regression
- Multicollinearity
- Simpson's paradox

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 all bottles*.

In this chapter we will see how to do inference in the case of a linear model.

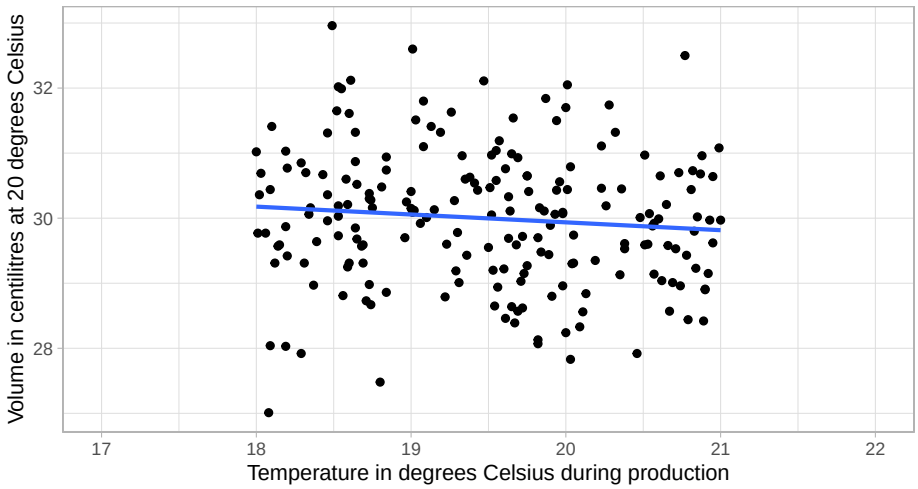


Figure 5.1: The relationship between temperature and volume in a sample of 200 bottles.

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.

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

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

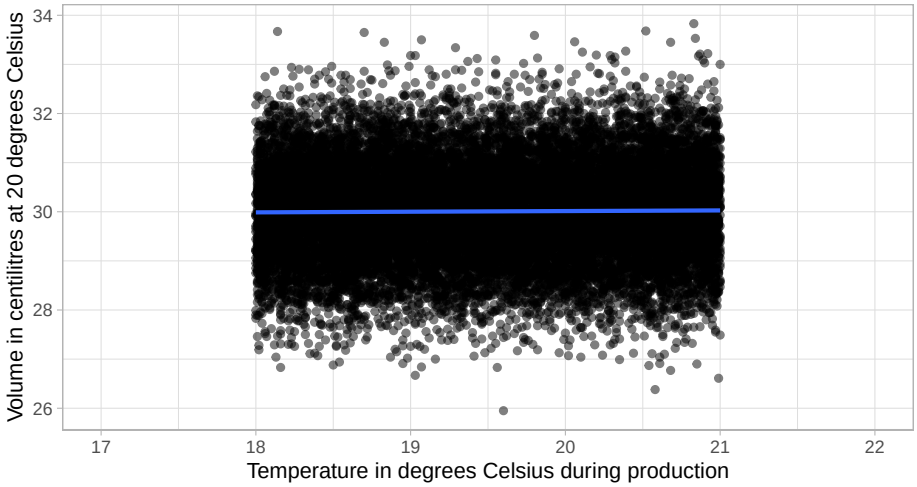


Figure 5.2: The relationship between temperature and volume in all 80,000 bottles.

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 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?

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

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).

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

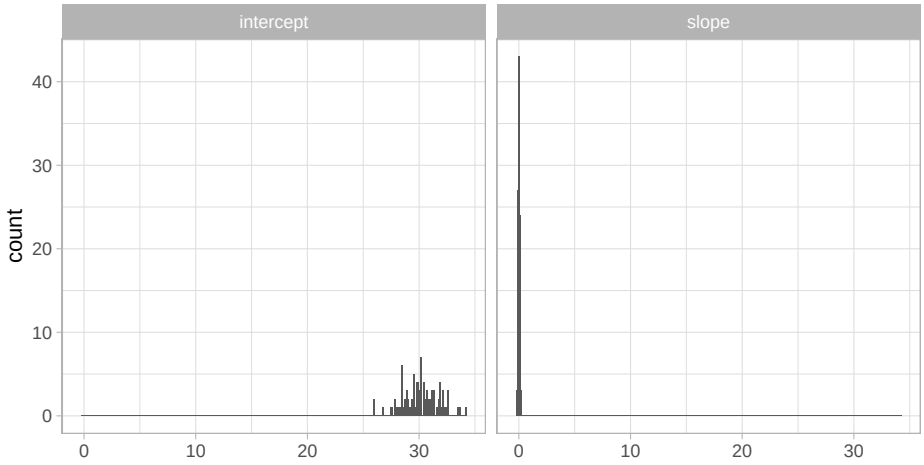


Figure 5.3: Distribution of the 100 sample intercepts and 100 sample slope.

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.

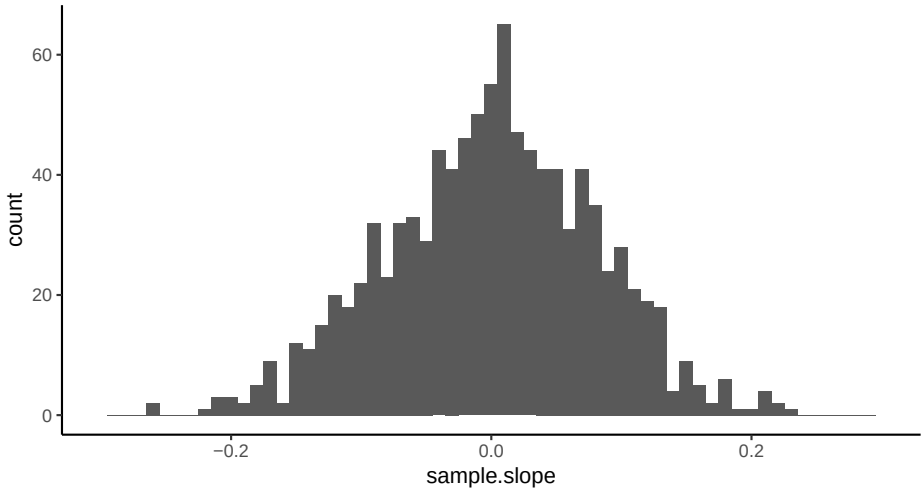


Figure 5.4: Distribution of 1000 sample slopes.

Remember from Chapter 2 that the standard deviation of the sampling 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 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 5.5 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.

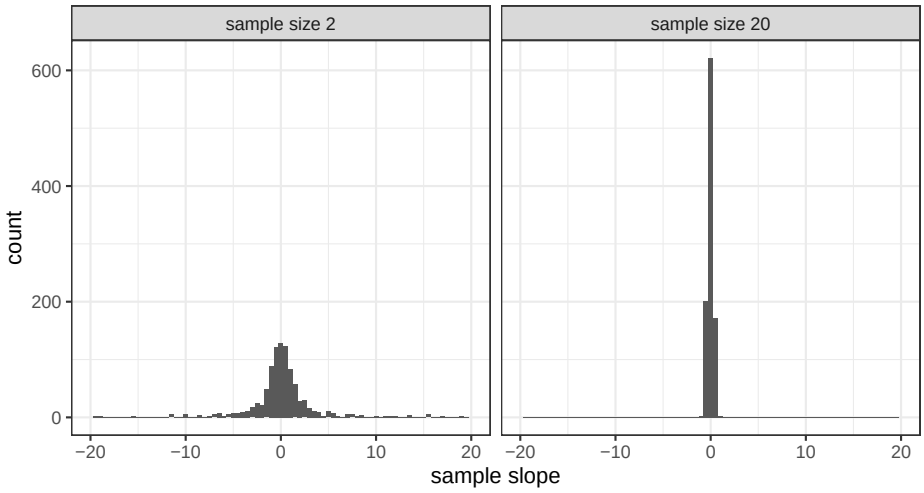


Figure 5.5: Distribution of the sample slope when sample size is 2 (left panel) and when sample size is 20 (right panel).

In Figure 5.5 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

[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.

population slope. The dependence of a sample slope on chance and sample size is also illustrated in Figure 5.6.

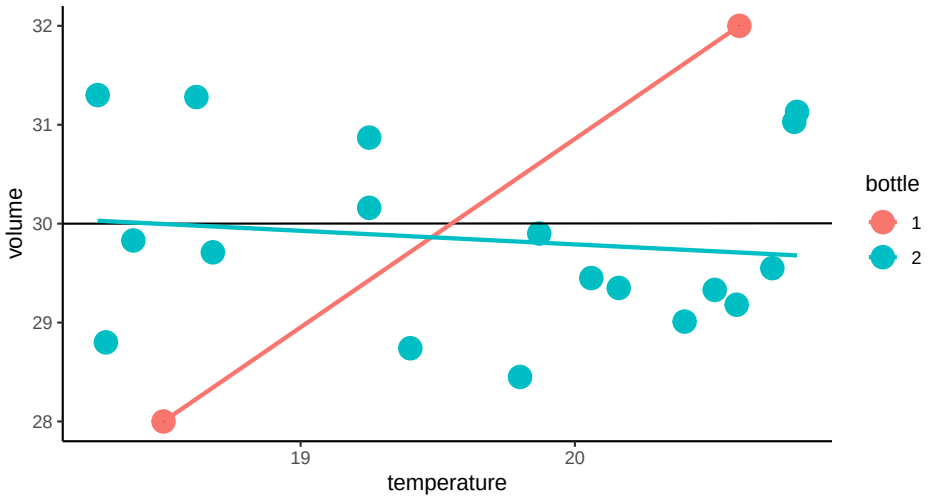


Figure 5.6: The averaging effect of increasing sample size. The scatter plot shows the relationship between temperature and volume for a random sample of 20 bottles (the dots); the first two bottles in the sample are marked in red. The red line would be the sample slope based on these first two bottles, 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.

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.

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:

$$\begin{aligned}\sigma_{\widehat{b_1}} &= \sqrt{\frac{s_R^2}{s_X^2 \times (n-1)}} \\ &= \sqrt{\frac{\frac{\sum_i (Y_i - \widehat{Y_i})^2}{n-2}}{\frac{\sum_i (X_i - \bar{X})^2}{n-1} \times (n-1)}}\end{aligned}$$

where b_1 is the slope coefficient in the sample, n is sample size, s_R^2 is the sample variance of the residuals, and s_X^2 the sample variance of 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.

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

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.

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

$$t = \frac{b_1 - \beta_1}{\widehat{\sigma_{b_1}}} = \frac{b_1 - \beta_1}{\sqrt{\frac{s_R^2}{s_X^2 \times (n-1)}}} \tag{5.1}$$

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 (s_R^2) is a little bit different, and also the variation in the predictor variable (s_X^2). 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 s_R^2 and s_X^2 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 (5.1) 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

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

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:

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

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:

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

In other words, if we believe Martha, our sample slope is only about 1.5 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.

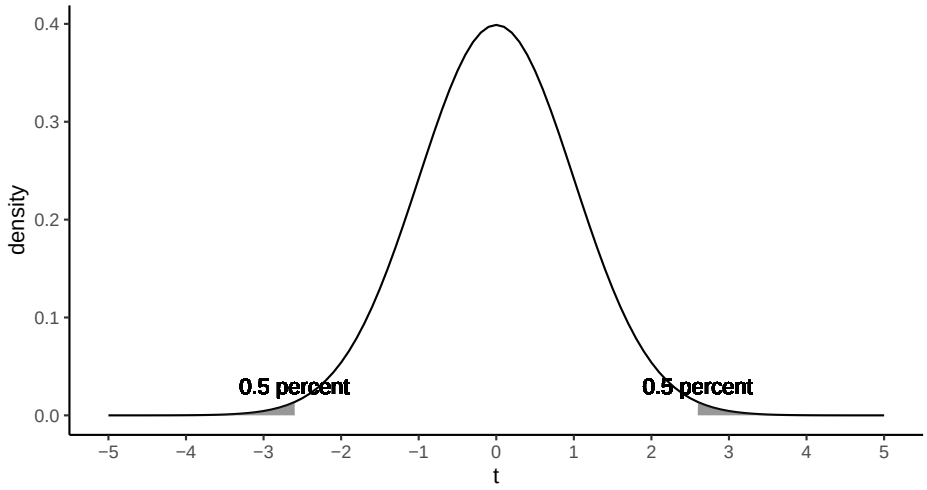


Figure 5.7: The t-distribution with 198 degrees of freedom.

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 $-0.121-2.6\times0.08$ to $-0.121+2.6\times0.08$ contains reasonable values for the population slope. If we do the calculations, 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.

We could report about this in the following way, mentioning sample size, the sample slope, the standard error, and the confidence interval.

> "Based on 200 randomly sampled bottles, we found a slope of -0.121 (SE = 0.08, 95% CI: -0.28, 0.04)."

Luckily, this interval contains the true value; 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.

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

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.

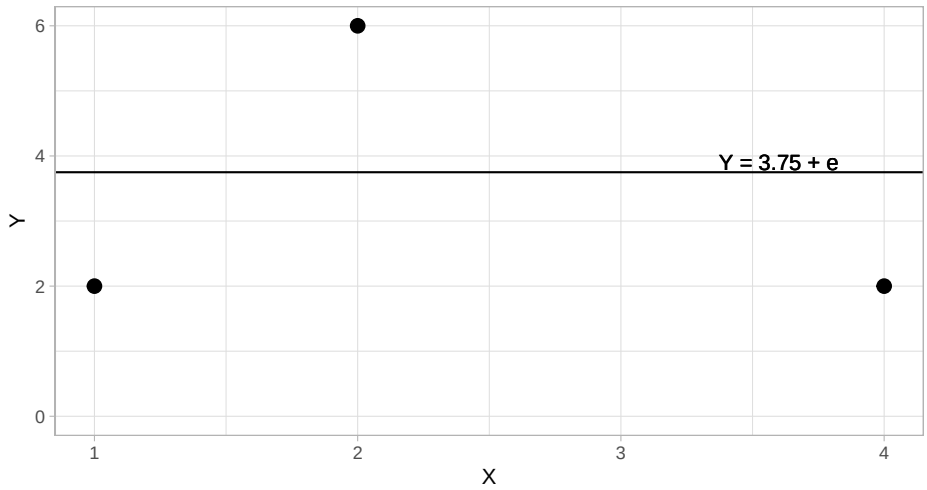


Figure 5.8: Illustration of residual degrees of freedom in a linear model, in case there is no slope and the intercept equals 3.75.

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 Chap. 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 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

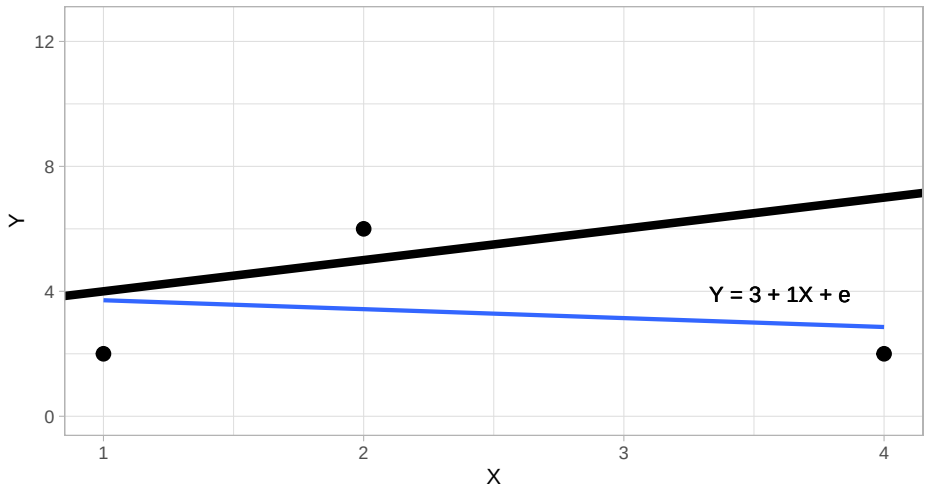


Figure 5.9: Illustration of residual degrees of freedom, in case of a linear model with both intercept and slope for four data points (black line). The blue line is the regression line only using the three known data points.

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 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.

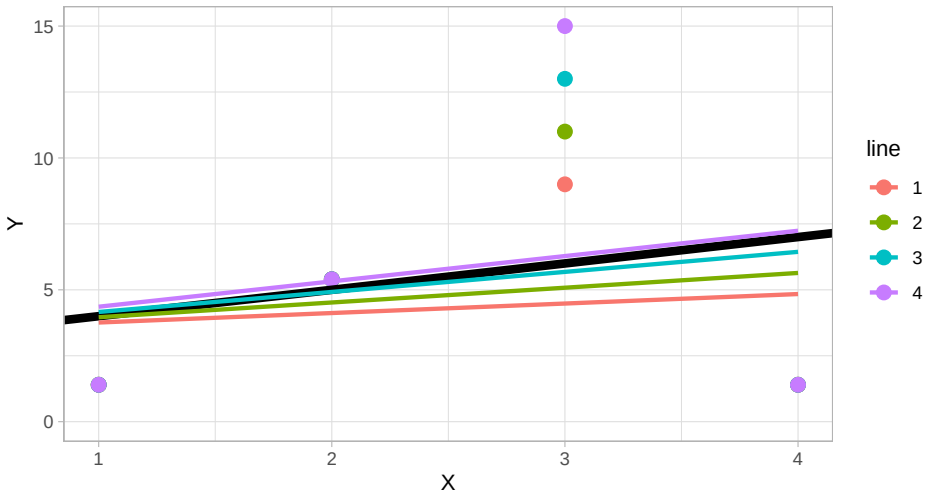


Figure 5.10: Different regression lines for different values of Y if $X = 3$.

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 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

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

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, from a population where the slope is 0, 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.

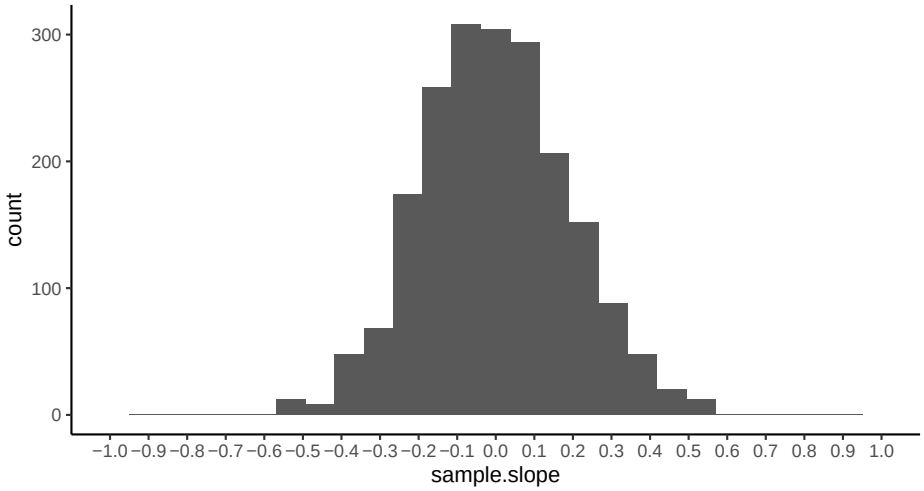


Figure 5.11: Distribution of the sample slope when the population slope is 0 and sample size equals 40.

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*. We say that the slope is *significantly different from 0*, or simply that *the slope is significant*.

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 or larger, 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.

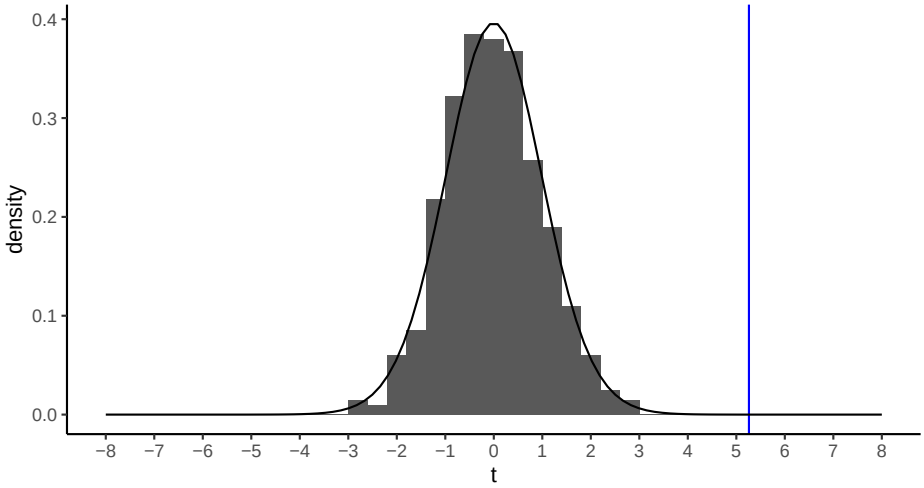


Figure 5.12: The histogram of 1000 sample slopes and its corresponding theoretical t-distribution with 38 degrees of freedom. The vertical blue line represents the t-value of 5.26.

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.

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

```
## [1] 2.939069e-06
```

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 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.

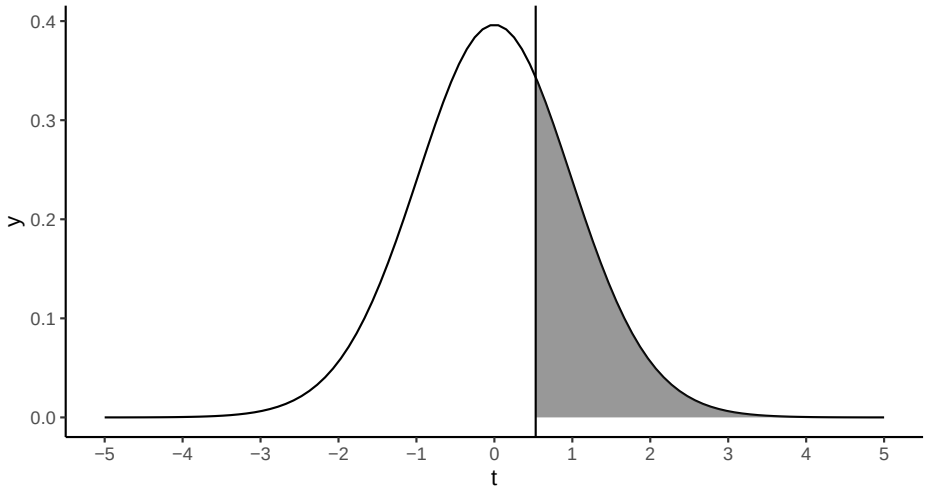


Figure 5.13: Probability of a t-value larger than 0.53.

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.

```r
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
```

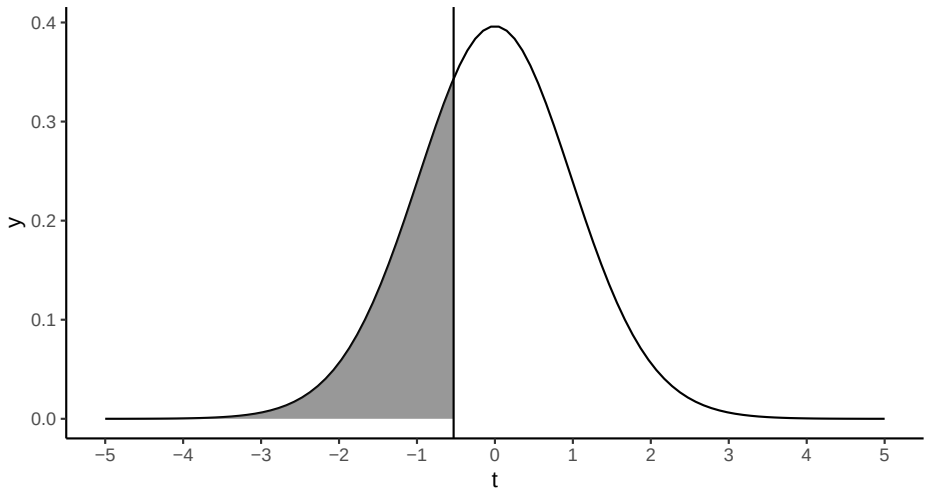


Figure 5.14: Probability of finding a t-value smaller than -0.53.

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

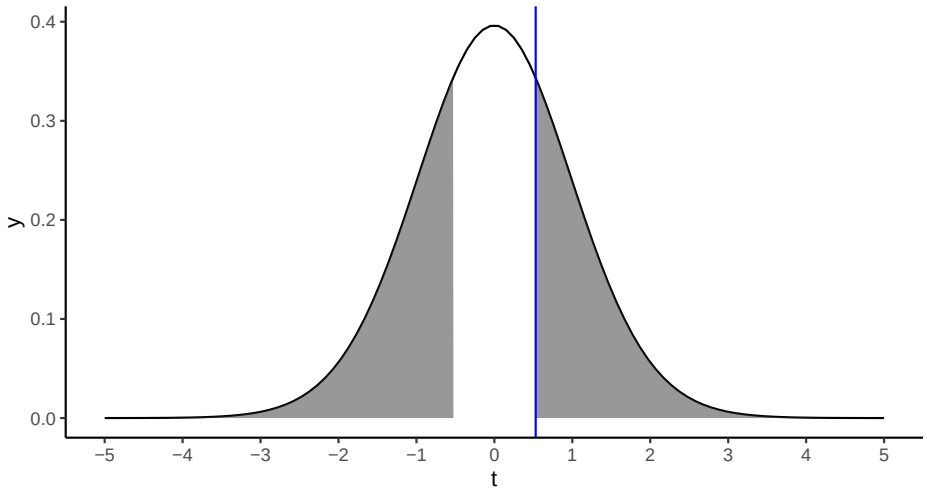


Figure 5.15: The blue vertical line represents a t-value of 0.53. The shaded area represents the two-sided p-value: the probability of obtaining a t-value smaller than -0.53 or larger than 0.53.

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 (i.e., the population slope could well be 0). One could say there is no empirical evidence for the existence of a slope not equal to 0. This leaves the possibility that there is a slope in the population, but that our method of research failed to find evidence for it. Formally, the null-hypothesis testing framework only allows refuting a null-hypothesis by some empirical evidence. It does not allow you to *prove* that the null-hypothesis is true, only that the null-hypothesis being true is a possibility.

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).

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

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 kilometres between university and their home. There are two research questions. The 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:

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

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:

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

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

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

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:

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

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 = TRUE, conf.level = 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   12.5      14.1
## 2 price.index    -0.309    0.0263    -11.8  6.92e-14   -0.381    -0.238
## 3 income.level    0.196    0.0291      6.74 7.20e- 8    0.117     0.276
```

```
freeny$market.potential %>% length()
```

```
## [1] 39
```

In the last two columns we see for example that the 99% confidence interval for the `price.index` slope runs from -0.381 to -0.238.

We can report:

> "In a multiple regression of market potential on price index and income level ($N = 39$), we found a slope for price index of -0.309 (SE = 0.026, 99% CI: -0.381, -0.238)."

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. The table below gives an overview.

Table 5.3: Four different scenarios for hypothesis tests.

	Test conclusion	
	do not reject H_0	reject H_0
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. The table below gives an overview.

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

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

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 $t = 1/2 = 0.5$).

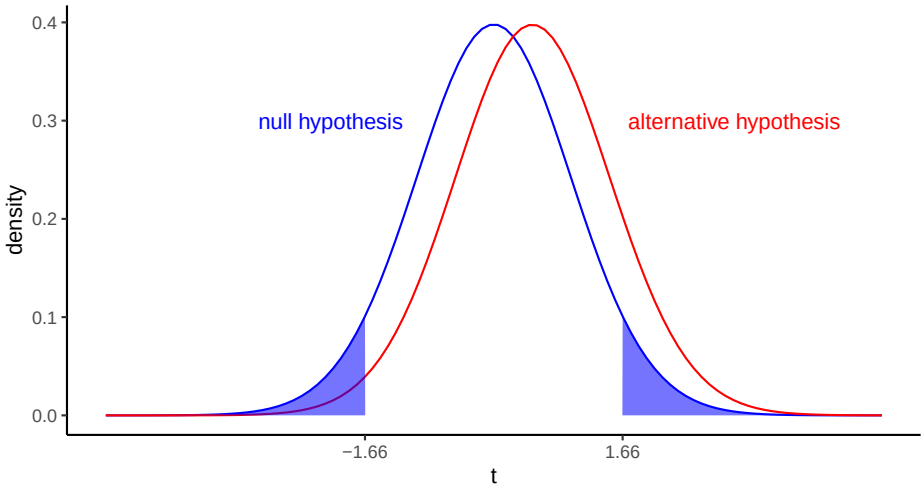


Figure 5.16: Different t-distributions of the sample slope if the population slope equals 0 (left curve in blue), and if the population slope equals 1 (right curve in red). Blue area depicts the probability that we find a p-value value smaller than 0.10 if the population slope is 0 (α).

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 @(fig:inf_21).

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 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

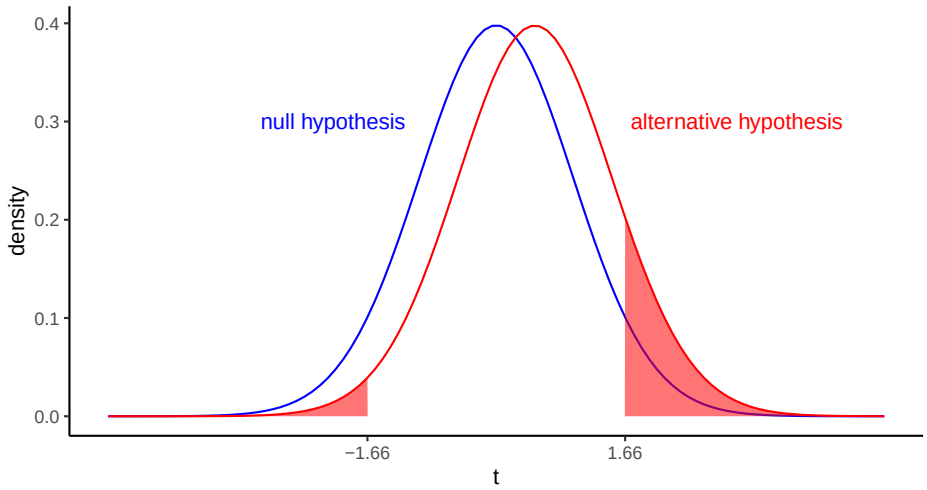


Figure 5.17: Different t-distributions of the sample slope if the population slope equals 0 (left curve in blue), and if the population slope equals 1 (right curve in red). Shaded area depicts the probability that we find a p-value value smaller than 0.10 if the population slope is 1 $(1 - \beta)$.

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.

The table below 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.

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

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

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).

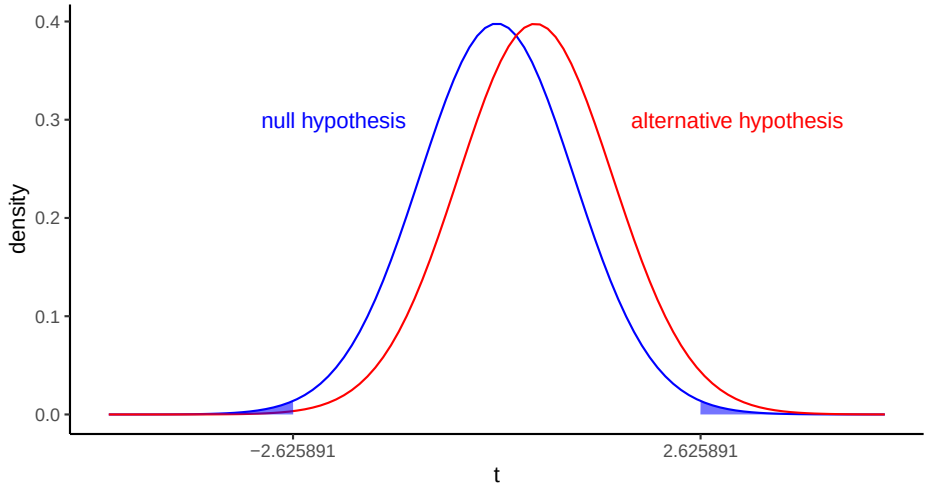


Figure 5.18: Different t-distributions of the sample slope if the population slope equals 0 (left curve), and if the population slope equals 1 (right curve). Blue area depicts the probability that we find a p-value value smaller than 0.01 if the population slope is 0.

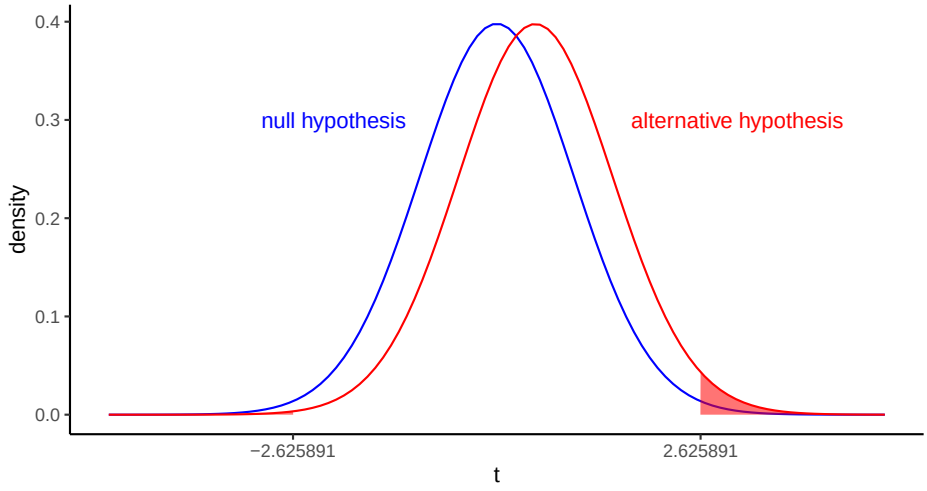


Figure 5.19: Different t-distributions of the sample slope if the population slope equals 0 (left curve in blue), and if the population slope equals 1 (right curve in red). Red area depicts the probability that we find a p-value value smaller than 0.01 if the population slope is 1: $1 - \beta$.

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.

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.

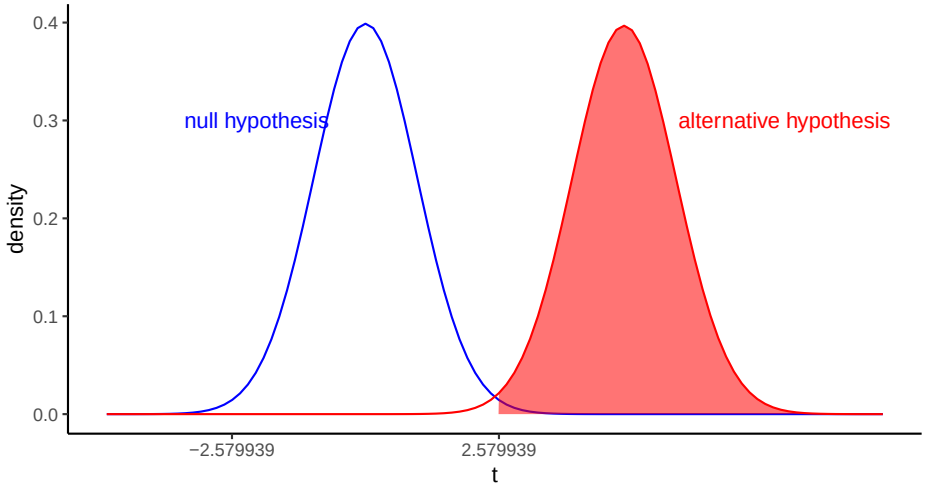


Figure 5.20: Different t-distributions of the sample slope if the population slope equals 0 (left curve in blue), and if the population slope equals 1 (right curve in red). Now for a larger sample size. Shaded area depicts the probability that we find a p-value value smaller than 0.01 if the population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

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 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 $\frac{b}{\sigma_{\hat{b}}} = 10/2 = 5$, 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.

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

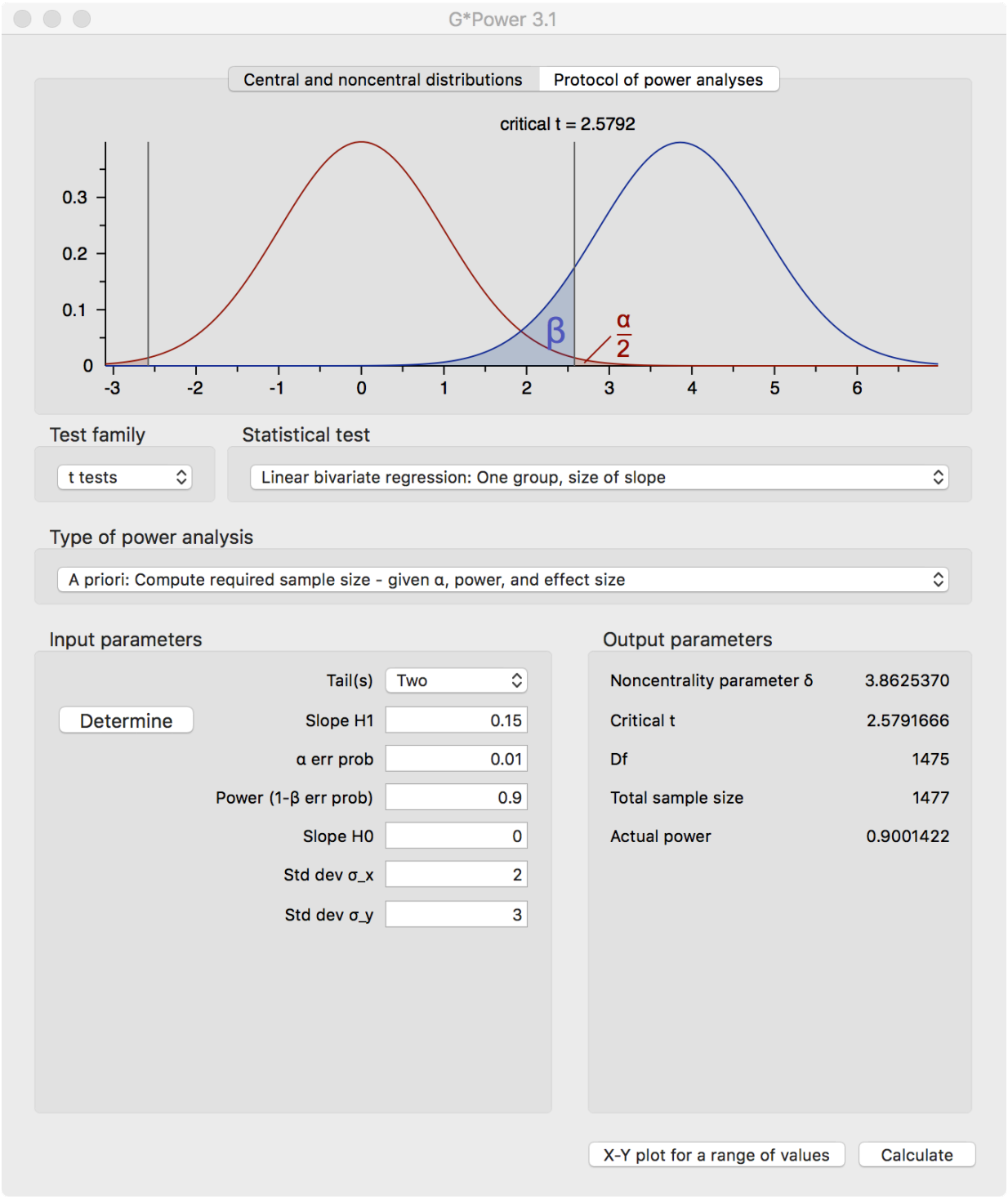


Figure 5.21: The G*power app interface.

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 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.

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

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

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.

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + e$$

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 $\sqrt{\frac{s^2}{n}} = 0.080$ (Section 2.13), 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.

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

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 β_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 = TRUE)
```

```
## # 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.

5.16 Take-away points

- You can use *sample* data to perform inference on *population* data.
- The linear model coefficients that are based on sample data have uncertainty due to the random sampling.
- Uncertainty of model parameters is quantified using standard errors.
- Standard errors become smaller with increasing sample size, and coefficients become more precise.
- Model parameters have t distributions, which can be used to construct confidence intervals.

- In a model with K independent variables, the residual degrees of freedom is $n - K - 1$.
- With null-hypothesis testing, usually the hypothesis is tested that a particular slope in the linear model is equal to 0 in the population.
- Statistical power refers to the probability of finding a significant result, given that the population coefficient is a certain non-zero value.
- Statistical power depends on the population value (if it is large, then you have a higher probability of finding a significant result than when it is close to 0).
- Statistical power increases with increasing sample size.
- Power analysis can give you insight about how many datapoints you need in order to have reasonable statistical power.
- Null-hypothesis is beginning to get outdated. Always consider if an alternate approach like reporting confidence intervals is more appropriate to answer your research question.
- A null-hypothesis can be done by checking whether a certain confidence interval includes 0. If 0 is not in the interval, the null-hypothesis that the population value is 0 can be rejected.
- With an intercept-only model, you can test the null-hypothesis that the population mean of the dependent variable equals 0, and compute a confidence interval.

Key concepts

- Statistical power
- Power analysis
- Intercept-only 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

Table 6.1: Bus trips to Paris.

person	seat	price
001	aisle	57
002	aisle	59
003	window	68
004	window	60
005	aisle	61

Table 6.2: Bus trips to Paris.

person	seat	window	price
001	aisle	0	57
002	aisle	0	59
003	window	1	68
004	window	1	60
005	aisle	0	61

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'.

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.

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:

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

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

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

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 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**.

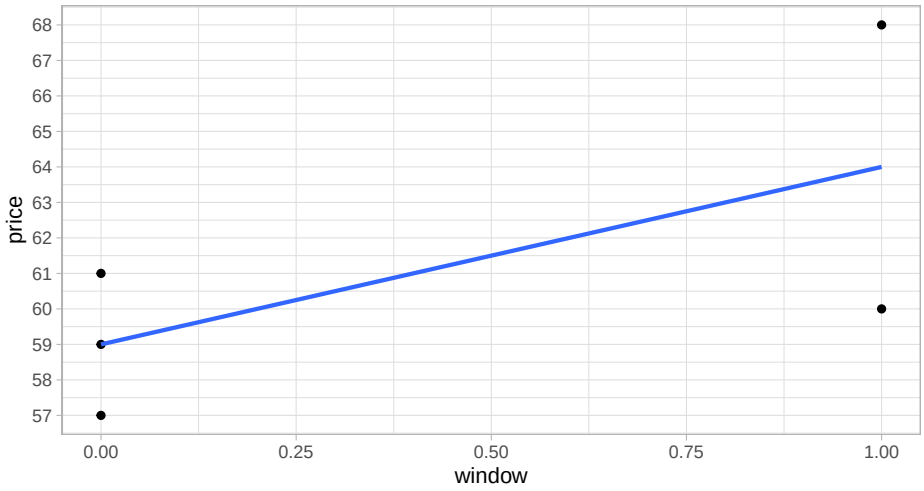


Figure 6.1: Relationship between dummy variable window and price.

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**.

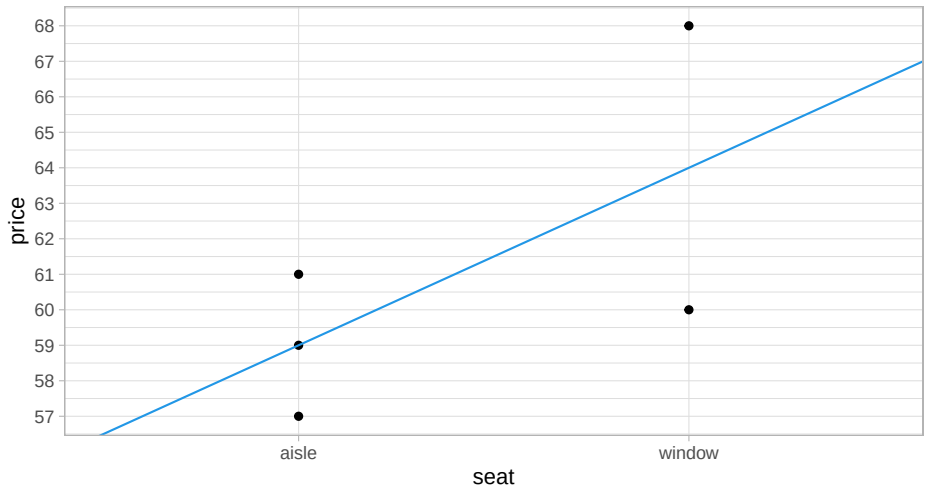


Figure 6.2: Relationship between type of seat and price.

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	57	-2	4
002	aisle	0	59	0	0
003	window	1	68	4	16
004	window	1	60	-4	16
005	aisle	0	61	2	4

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

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:

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	57	59.1	-2.1	4.41
002	aisle	0	59	59.1	-0.1	0.01
003	window	1	68	63.9	4.1	16.81
004	window	1	60	63.9	-3.9	15.21
005	aisle	0	61	59.1	1.9	3.61

40.05.

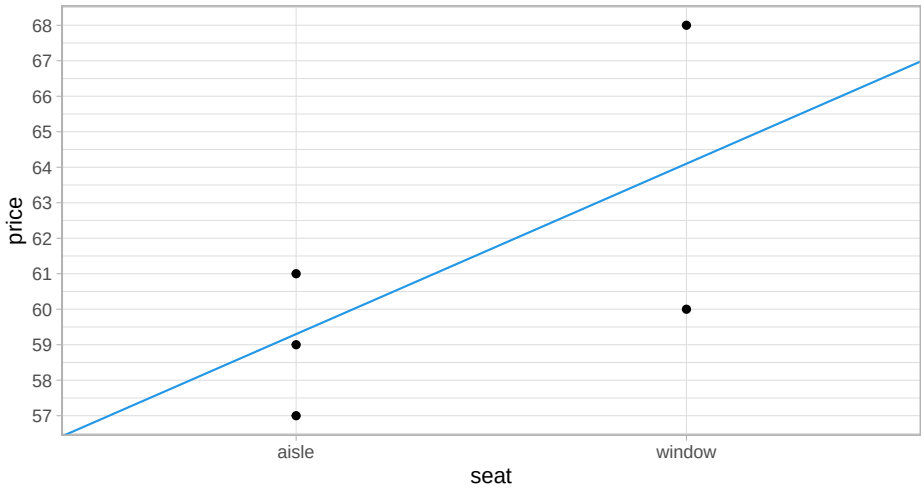


Figure 6.3: Relation between type of seat and price, with the regression line being not quite the least squares line.

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 conditions (see Figure 6.4).

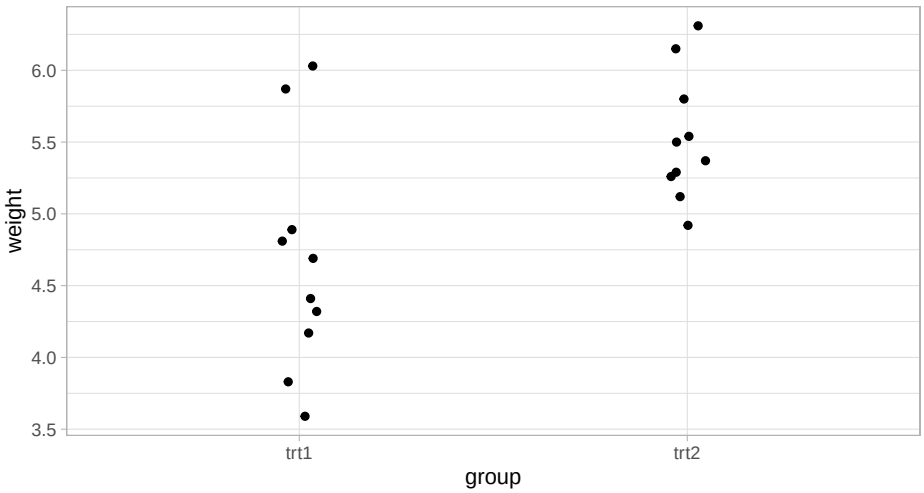


Figure 6.4: Data on yield under two experimental conditions: treatment 1 and treatment 2.

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:

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

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

$$\begin{aligned} 4.661 &= b_0 + b_1 \times 0 = b_0 \\ 5.526 &= b_0 + b_1 \times 1 = b_0 + b_1 \end{aligned}$$

So from this, we know that intercept $b_0 = 4.661$, and if we fill that in for the second equation above, we get the slope:

$$b_1 = 5.526 - b_0 = 5.526 - 4.661 = 0.865.$$

Thus, we get the linear equation

$$\widehat{\texttt{weight}} = 4.661 + 0.865 \times \texttt{treatment} \tag{6.1}$$

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

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	0.86	0.29	3.01	0.01	0.26	1.47

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 (6.1), 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 $\alpha = 0.05$.

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:

> "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, data = .) %>%
  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 %>%
  dplyr::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, data = .)
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. It is a dummy variable, coding treatment group 2 as 1 and treatment group 1 as 0. The slope value of 0.865 now refers to the effect of this dummy variable for 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 have to make sure that the treatment 2 group is coded as the first level of the group factor:

```
PlantGrowth <- PlantGrowth %>%
 mutate(group = factor(group, levels = c("trt2", "trt1", "ctrl")))
```

We can check the order of the categories using the following code:

```
PlantGrowth %>%
  pull(group) %>%
  levels()
```

```
## [1] "trt2" "trt1" "ctrl"
```

You see that "trt2" comes first, so this will by default be the reference category from now on.

Next, you can run a linear model with this slightly altered data set:

```
out <- PlantGrowth %>%
  filter(group != "ctrl") %>%
  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 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 1. 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 2 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 by default. If that however makes the output hard to interpret, think of the best way to change the order of the levels (categories). For experimental designs, it makes sense to code control conditions as the first level, and experimental conditions as second level (you're often interested in the effect of the experimental condition *compared* to the control condition, so the control condition is the reference

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

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 majority group as the reference group (the first level) and the social minority as the second group. In educational studies, it makes sense to code an old teaching method as the reference group and a new method as the second group. 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:

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

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**.

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:

$$\widehat{\texttt{price}} = 41.5 + 5 \times \texttt{window} + 0.25 \times \texttt{legroom} \tag{6.2}$$

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 centimetres 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.

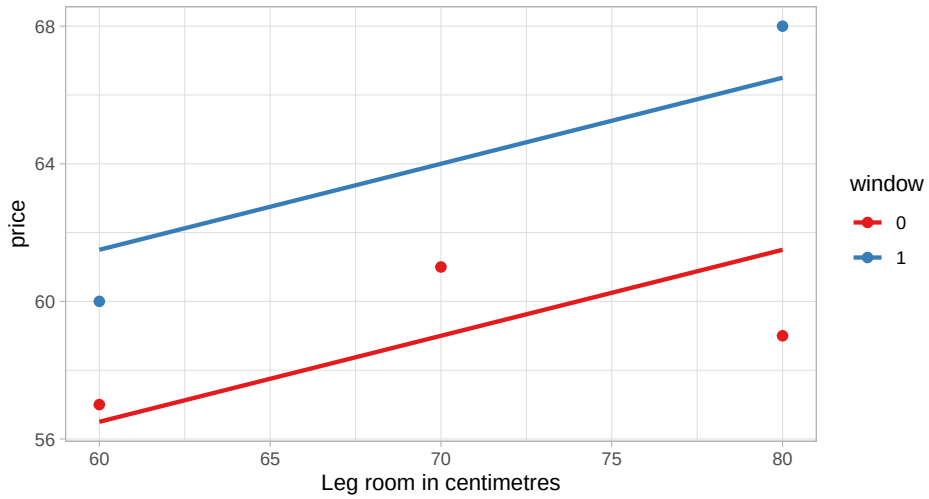


Figure 6.5: The bus trip to Paris data, with the predictions from a linear model with legroom and window as independent variables.

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 (6.2) and make predictions for aisle seats, we fill in `window = 0` and we get the following linear equation:

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

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

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

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 steeper 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 regression 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 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, given that the lines must be parallel.

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

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,

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

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 now have 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. We discuss how to do that in R. We create factors in such a way that we control in what order the categories are coded. The first category will always be the reference group.

R creates dummy variables automatically when we use factor variables. The first category (in terms of internally numbered category) by default 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. First, let's check whether the **group** variable is indeed a factor variable.

In a previous section we changed the data set, so let's start with the original data set again.

```
data("PlantGrowth")
```

If we look at the structure of it, we can see the types of variables.

```
PlantGrowth %>%
  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 ...
```

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. Let's check the order in which the levels/categories are coded:

```
PlantGrowth %>%
  pull(group) %>%
  levels()
```

```
## [1] "ctrl" "trt1" "trt2"
```

We see the control condition comes first, so that will be the reference category when we run a linear model.

```
out <- PlantGrowth %>%
  lm(weight ~ group, data = .)
out %>%
  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 grouptrt1     -0.371     0.279     -1.33 1.94e- 1
## 3 grouptrt2      0.494     0.279      1.77 8.77e- 2
```

We see the effect of treatment 1 and the effect of treatment 2 against the reference category ctrl.

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

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):

$$\widehat{\texttt{weight}} = 5.03 - 0.37 \times \texttt{grouptrt1} + 0.49 \times \texttt{grouptrt2}$$

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

Control group specimens score 0 on variable **grouptrt1** and 0 on variable **grouptrt2**. 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 **grouptrt1** variable but 0 on the **grouptrt2** 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 **grouptrt1** variable but 1 on the **grouptrt2** variable. Therefore, their predicted weight equals:

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

6.8 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 **grouptrt1** dummy variable is the estimated population difference between the mean for treatment 1 minus the mean for the control condition. Similarly, the slope for the **grouptrt2** 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;

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 changing the order of the levels in a different way, where either treatment 1 or 2 is the reference group.

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*.

6.9 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 population of the treatment 2 group, and μ_c be the mean yield in the population of the control group. Then we can specify the null-hypothesis using symbols in the following way:

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

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, where we make use of the `Anova()` function from the `car` package:

```
library(car)
PlantGrowth %>%
  lm(weight ~ group, data = ., contrasts = list(group = contr.sum)) %>%
  Anova(type = 3) %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term         sumsq    df statistic   p.value
##   <chr>        <dbl> <dbl>     <dbl>     <dbl>
## 1 (Intercept) 772.       1   1987.    8.02e-27
```

```
## 2 group         3.77     2      4.85  1.59e- 2
## 3 Residuals    10.5     27     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.

Note that we estimate the linear model in a slightly different way. Instead of running `lm(weight ~ group, data = .)` we add the extra code `contrasts = list(group = contr.sum)`. This has to do with how you want the categorical variables to be coded in a linear model analysis (in this case the variable **group**). R uses dummy coding by default (using 0 and 1). Instead of dummy coding, we here use coding using -1 and 1 (called sum-to-zero coding, explained in Ch. 10). For the simple ANOVA here it does not make a difference, but it will make a difference once you want to include more independent variables (see Ch. 9). We include it here to be consistent throughout the book. Also, the code `type = 3` is not strictly necessary here, but it will be once you move to multiple independent variables.

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.

F-values are test statistics and are used in the same way as t-statistics. Under the null-hypothesis they have a known distribution (an F-distribution) with an average of 1. If the F-value is significantly larger than 1, you can reject the null-hypothesis. Whether or not the null-hypothesis can be rejected, depends not only on the p-value, reported in the last column (`p.value`), but also on the degrees of freedom. The degrees of freedom determine the shape of the F-distribution, in the same way as the degrees of freedom for the t-distribution.

The only difference between F-tests and t-tests is that there are two different kinds of degrees of freedom. There are the residual degrees of freedom, which corresponds to the degrees of freedom for a t-test: the larger your sample size, the larger the degrees of freedom. For the F-test there is also the degrees of freedom for the effect you are interested in. Here we compare 3 groups, and we see that we have 2 degrees of freedom for this effect of group (you see that in the `df` column in the ANOVA output. In general, when comparing K groups, the degrees of freedom is $K - 1$ for the group effect.

In the output we see that the F-value for the `group` effect is larger than 1, and that the associated p-value `p.value` is smaller than 0.05. We then report that the group means are significantly different, $F(2, 27) = 4.85$, $p = 0.016$. Note that the effect degrees of freedom comes first, and then the residual degrees of freedom.

6.10 Computing and testing the F-statistic

The F-statistic is constructed on the basis of Sums 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 was minimised to get the least squares estimator of regression coefficients. Actually, the sum of squares that we see in the ANOVA table here in the row named `Residuals` is exactly the same 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) and report the t-test. 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).

In many ANOVA tables we also see Mean Squares. They form the basis for the F-statistic. The Mean Squares are not shown in the ANOVA table here, but they can easily be computed based on the Sums of Squares and the degrees of freedom. These Mean Squares the sums of squares (`sumsq`) divided by the respective degrees of freedom (`df`), $MS = \frac{SS}{df}$. For instance, in the row for Residuals, the Sum of Squares equals 10.5, the degrees of freedom equals 27, and the Mean square error therefore equals $\frac{10.5}{27} = 0.389$. For the effect of group, the Mean Square is $\frac{3.77}{2} = 1.885$.

The F-value is in turn computed based on these Mean squares values.

$$F = \frac{MS_{effect}}{MS_{residuals}}$$

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 we computed as 1.885, and the mean square of the residuals (error), which we computed as 0.389. Thus, the F-value for country is computed as $\frac{1.885}{0.389} = 4.85$. 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? As we said, it 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 group was 2 (3 groups so 2 degrees of freedom), and the one for the residual mean square was 27 (27 residual degrees of freedom). We therefore can 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

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).

	Model degrees of freedom							
	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

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.

Of course, instead of using the table, you can also simply report the p-value plotted in your R output, and call it significant when it is smaller than your pre-set α.

6.11 Difference between ANOVA and regular linear model output

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 = TRUE)
```

```
## # 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 test-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 ($t = 1.77, p = 0.088$), but we actually need one p-value for the null-hypothesis of three equal means.

The ANOVA table looks 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:

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

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 group). 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 group). 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.

> The general 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.

6.12 The logic of the F-statistic (advanced)

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:

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

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:

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

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.

Although both models are equivalent, the one with the three means is more helpful when understanding the logic of ANOVA and the F-statistic, since they are about the estimated group means in the population.

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. Chapter 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 (the sample means differ from the population mean because of random sampling).

This variation in the 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, the distribution of these many sample means 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 $\sigma_{\bar{Y}} = \sqrt{\frac{s^2}{n}}$ (Chapter 2). If we take the square, we see that the variance of the sample means is equal to

$$\sigma^2_{\bar{Y}} = \frac{s^2}{n} \tag{6.3}$$

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:

$$\widehat{\sigma^2} = s^2 = \sigma^2_{\bar{Y}} \times n = \frac{SS}{2} \times n \tag{6.4}$$

or in words: our estimate of the residual variance σ^2 in the population is the estimated variance of the group means in the population times the number of observations per group. Here, note that SS refers to the sum of squared differences between the sample means and the population mean. To compute the variance, we divide by the number of groups minus 1 (i.e., 2).

So the numerator of F 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 (we computed the sum of squared differences for the

group means). 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 $SSR/(n-3)$.

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.

6.13 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 $\bar{Y} = 2.11$, denoted in black in Figure 6.6. In red, we see the sample group means. For group A, that is equal to $(4+2+1)/3 = 2.33$, for group B this is $(4+2+2)/3 = 2.67$, and for group C this equals $(2+1+1)/3 = 1.33$.

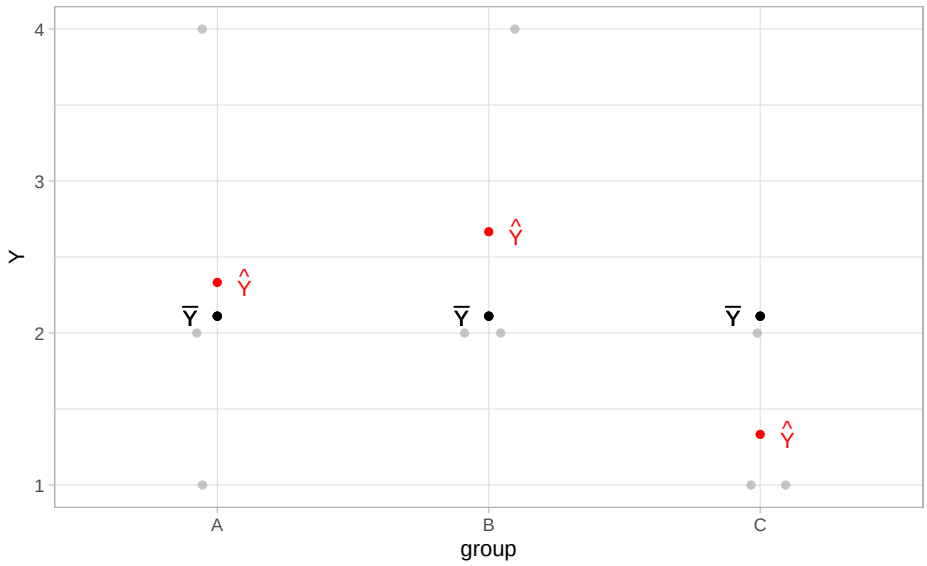


Figure 6.6: Illustration of ANOVA using a very small data set. In grey the raw data, in black the overall sample mean (grand mean), and in red the sample group means.

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

$$Y = b_1 \texttt{dummy}_A + b_2 \texttt{dummy}_B + b_3 \texttt{dummy}_C + e \qquad (6.5)$$
$$e \sim N(0, \sigma^2)$$

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

$$\widehat{Y} = 2.33 \texttt{dummy}_A + 2.67 \texttt{dummy}_B + 1.33 \texttt{dummy}_C$$

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 (6.6)) and the predicted values. We see that the predicted value for each observed value is equal to the sample group mean.

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 $SSR = 8.00$. To obtain the Mean Squared Error (MSE or `meansq` for Residuals), we divide the SSR by

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

Table 6.11: ANOVA table for small data example.

term	sumsq	df	statistic	p.value
(Intercept)	40.11	1	30.08	0.002
group	2.89	2	1.08	0.397
Residuals	8.00	6	NA	NA

the degrees of freedom. Because the linear model with 2 dummy variables (3 groups) has $n - K = 9 - 3 = 6$ residual degrees of freedom (see Chapter 5), we also have only 6 residual degrees of freedom. Thus we get $MSE = 8/6 = 1.33$. We can see the values for SSR (`sumsq`) and the degrees of freedom (`df`) 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 2.33, so the squared deviation equals 0.05. The sample mean for group B was 2.67, so the squared deviation equals 0.31. Lastly, the sample mean for group C was 1.33, so the squared deviation equals 0.31. Adding these squared deviations gives a sum of squares of 0.96. 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 = 0.48. This we must multiply by the sample size per group (3) to obtain an estimate of σ^2 (see Equation (6.3)), thus we obtain 1.44.

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 sum the squared deviations, we no longer have to multiply by sample size. Thus we have as the sum of squares 2.89. Then we only have to divide by the number

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

of groups minus 1, so we have $2.89/2 = 1.44$. This sum of squares and the degrees of freedom of 2 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,

$$F = \frac{MSB_{\texttt{group}}}{MSE} = \frac{1.44}{1.33} = 1.08$$

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 randomly drawn 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 1.08, 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 1.08 or larger. R output tells us exactly how often that happens by reporting the p-value.

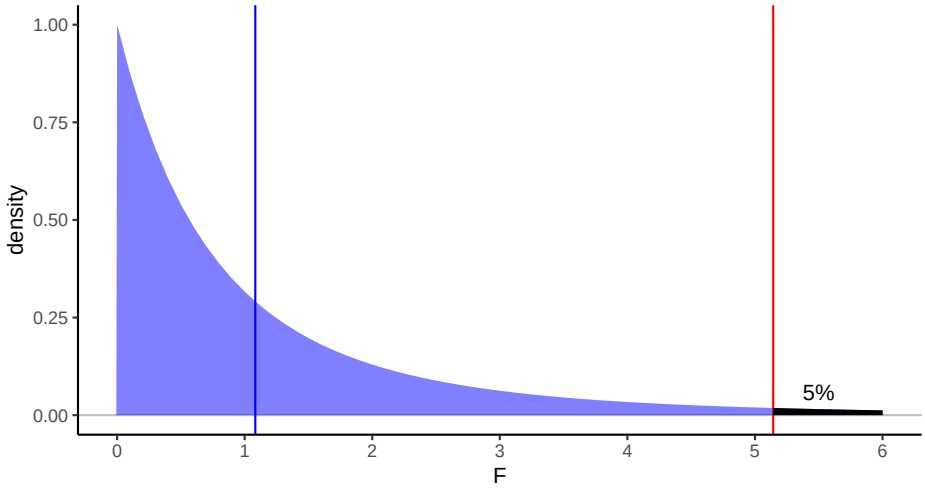


Figure 6.7: Density plot of the F-distribution with 2 and 6 degrees of freedom. In blue the observed F-statistic in the small data example, in red the critical value for an α of 0.05. The blackened area under the curve is 5%.

6.14 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, $F(2,\ 6) = 1.08, p = .40$."

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

variables that 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.

6.15 Relationship between F- and t-distributions (advanced)

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 $\alpha = 0.05$.

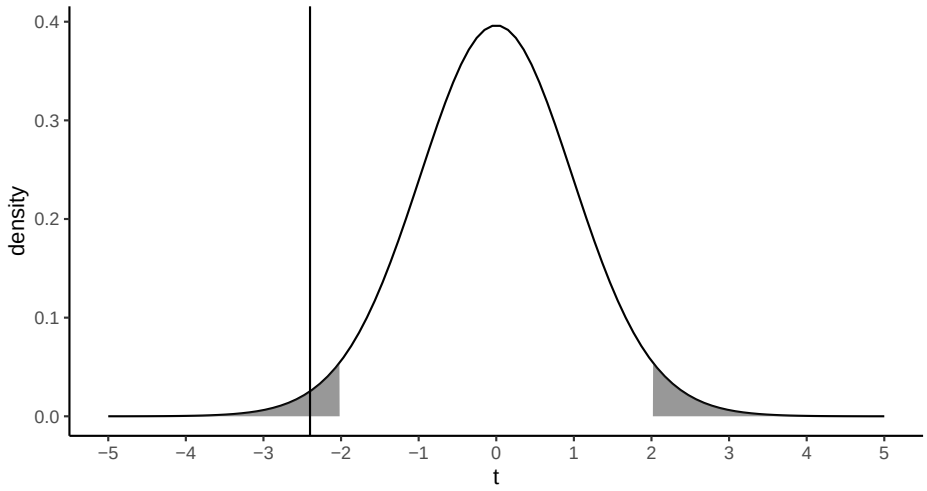


Figure 6.8: The vertical line represents a t-value of -2.40. The shaded area represents the extreme 5% of the possible t-values.

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 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.

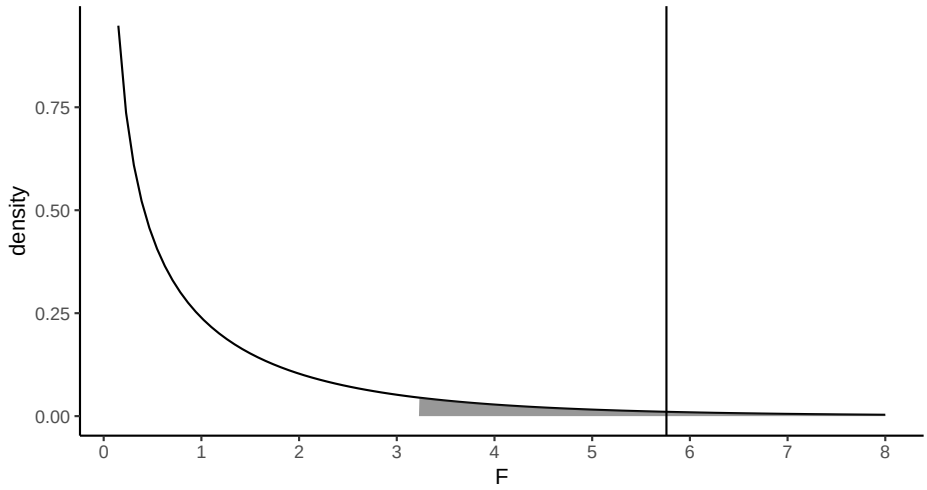


Figure 6.9: The F-distribution with 1 model degrees of freedom and 40 error degrees of freedom. The shaded area is the upper 5% of the distribution. The vertical line represents the square of -2.40: 5.76

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 (one-sided) p-value associated with an F-value that is the square of the t-value.

$$F(1, x) = t^2(x)$$

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.

6.16 Take-away points

- A categorical independent variable can be added to a model by creating a quantitative variable that represents the categorical one.
- A categorical variable with only two classes (groups/categories) can be recoded into a dummy variable.
- In regression with a dummy variable, the slope is equal to the difference in group means.
- A categorical variable with more than two classes (groups/categories) can be recoded into a *set* of dummy variables.
- With dummy coding there is always a reference group (the one that is coded as 0).
- When testing a null-hypothesis about the equality of more than two group means, an analysis of variance (ANOVA) should be carried out, reporting the F-statistic.
- When the null-hypothesis is true, then you expect to see an F-value of around 1.
- An F-value much greater than 1 indicates evidence for rejecting the null-hypothesis. How much evidence depends on the degrees of freedom (model and residual degrees of freedom).
- The F-distribution is closely related to the t-distribution.

Key concepts

- Dummy coding
- Analysis of variance (ANOVA)
- F-statistic
- Within and between-groups sums of squares
- Model degrees of freedom
- Residual degrees of freedom
- Mean squares

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.

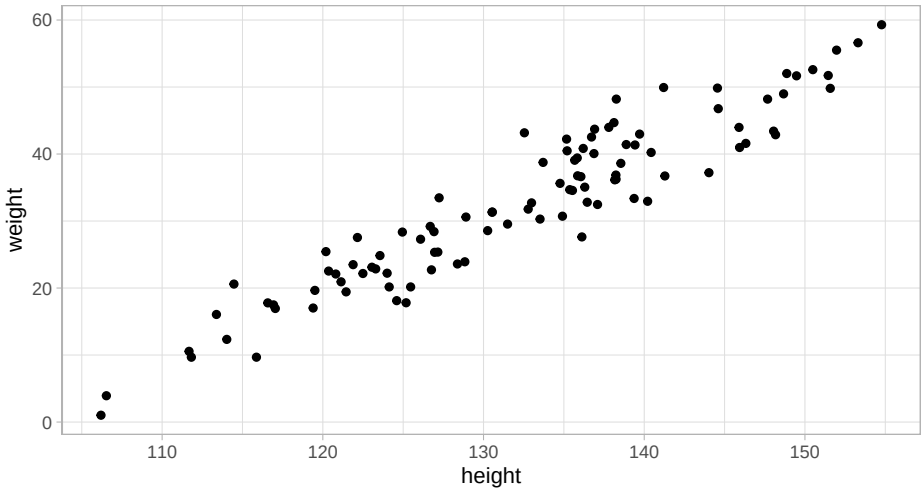


Figure 7.1: Data set on height and weight in 100 children.

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:

$$\texttt{weight} = -104.83 + 1.04 \times \texttt{height} + e$$
$$e \sim N(0, \sigma = 4.04)$$

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.

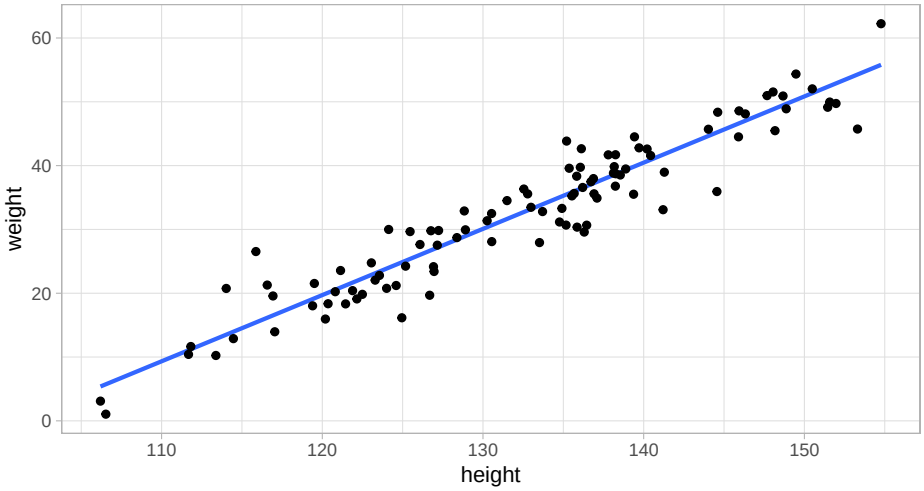


Figure 7.2: Data set on height and weight in 100 children and the least squares regression line.

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 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 systematic 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.

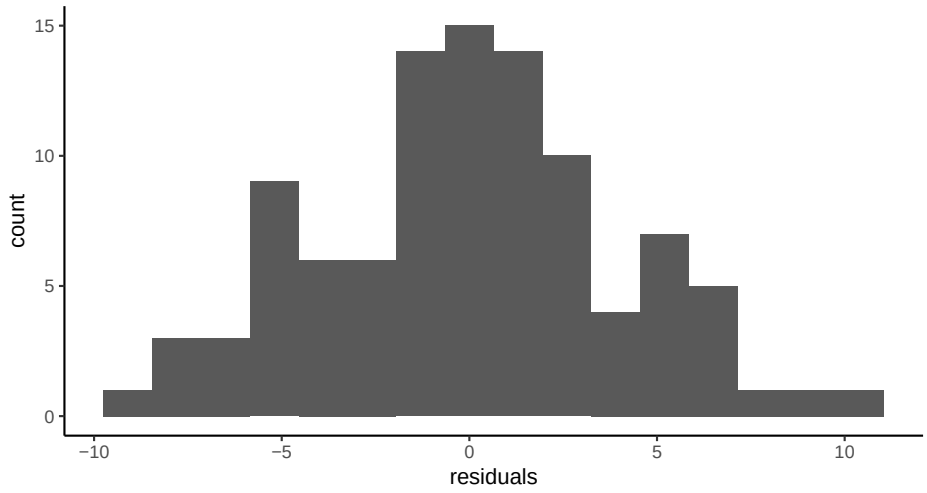


Figure 7.3: Histogram of the residuals after regressing weight on height.

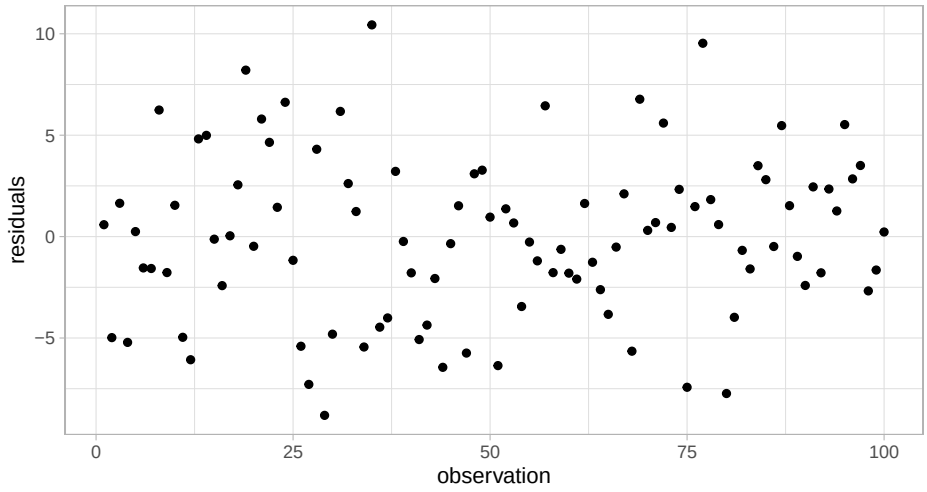


Figure 7.4: Residual plot after regressing weight on height.

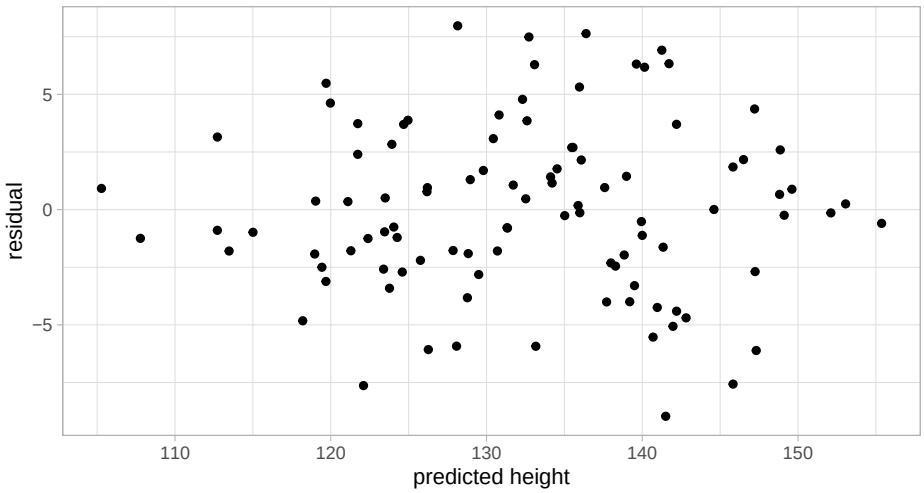


Figure 7.5: Residuals as a function of height.

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:

$$\texttt{height} = 100 + 5 \times \texttt{age} + e \tag{7.1}$$

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.

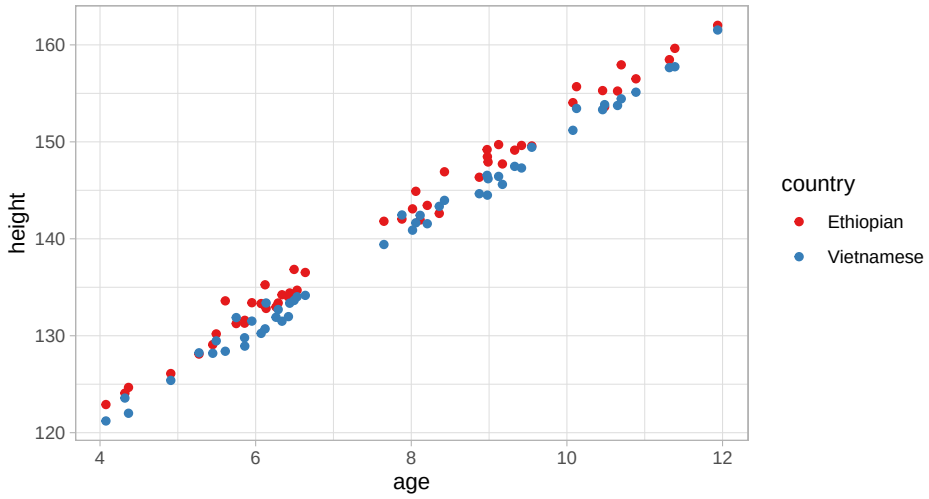


Figure 7.6: Data on age and height in children from two countries.

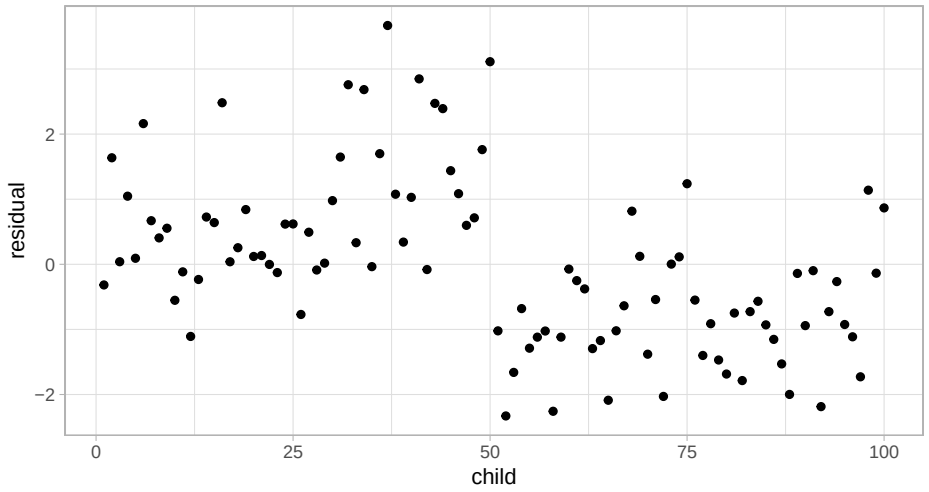


Figure 7.7: Residual plot after regressing height on age.

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: 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.

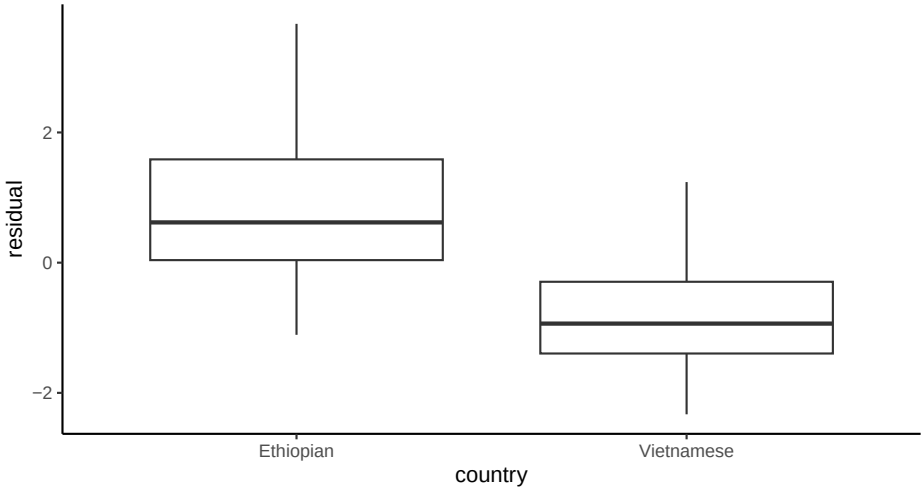


Figure 7.8: Residual plot after regressing height on age.

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:

$$\widehat{\texttt{height}} = 102.641 + 5.017 \times \texttt{age} - 1.712 \times \texttt{countryViet} \qquad (7.2)$$

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.

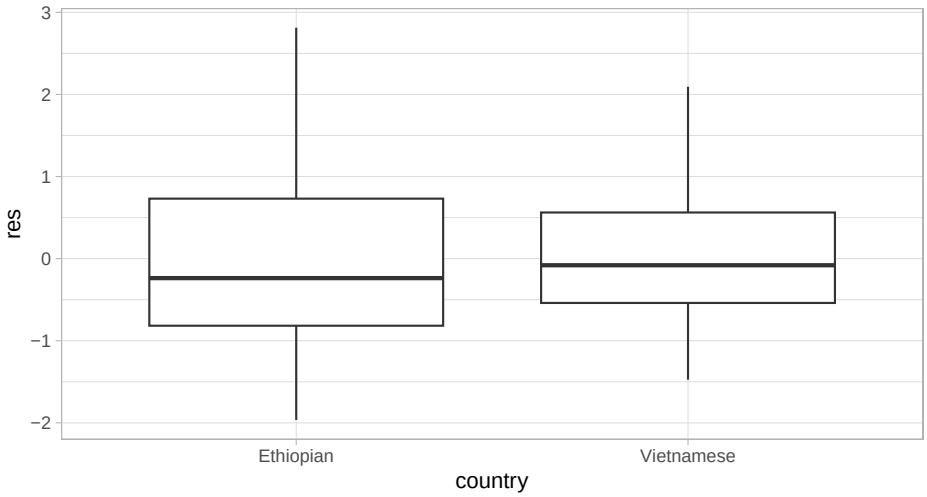


Figure 7.9: Residual plot after regressing height on age and country.

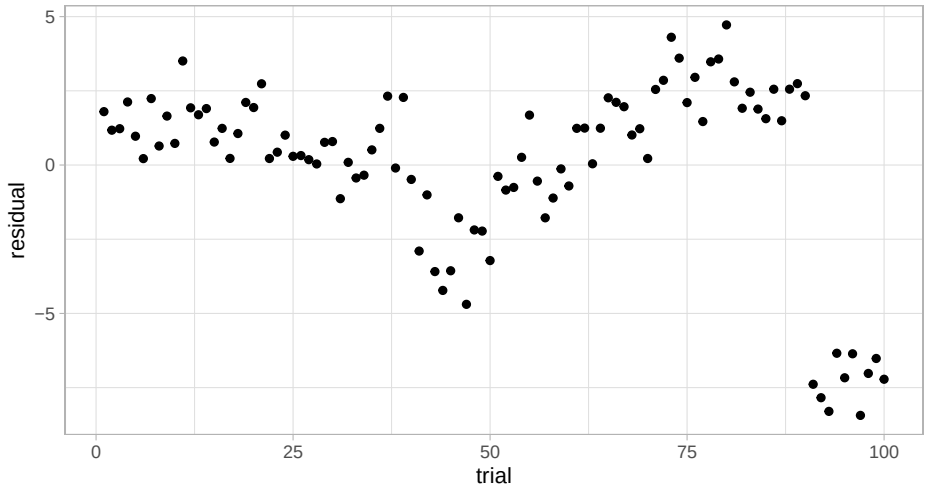


Figure 7.10: Residual plot after regressing reaction time on IQ.

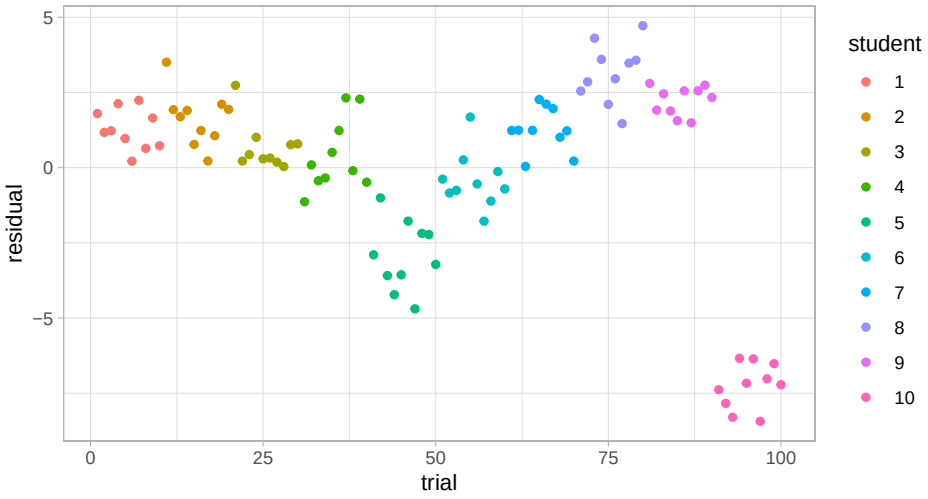


Figure 7.11: Residual plot after regressing reaction time on IQ, with separate colours for each student.

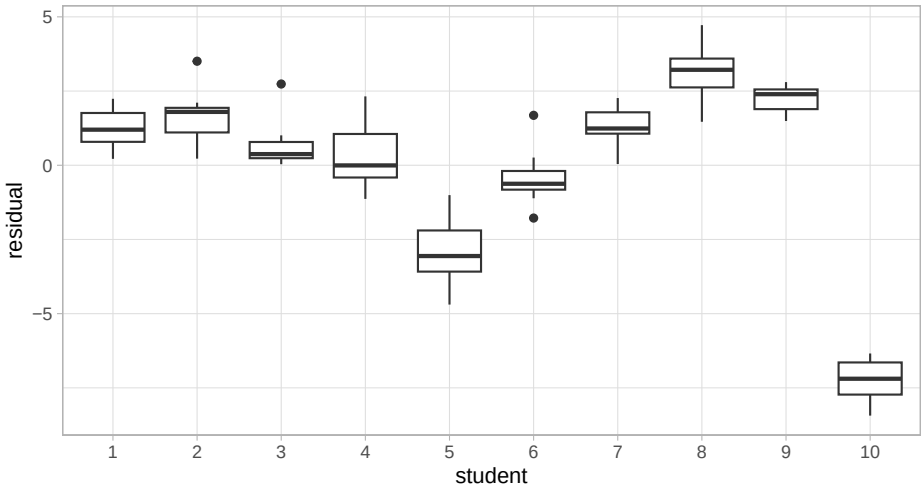


Figure 7.12: Box plot after regressing reaction time on IQ.

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 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 12.

The assumption of independence is the most important assumption in linear models. Only 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, resulting in a much wider confidence 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

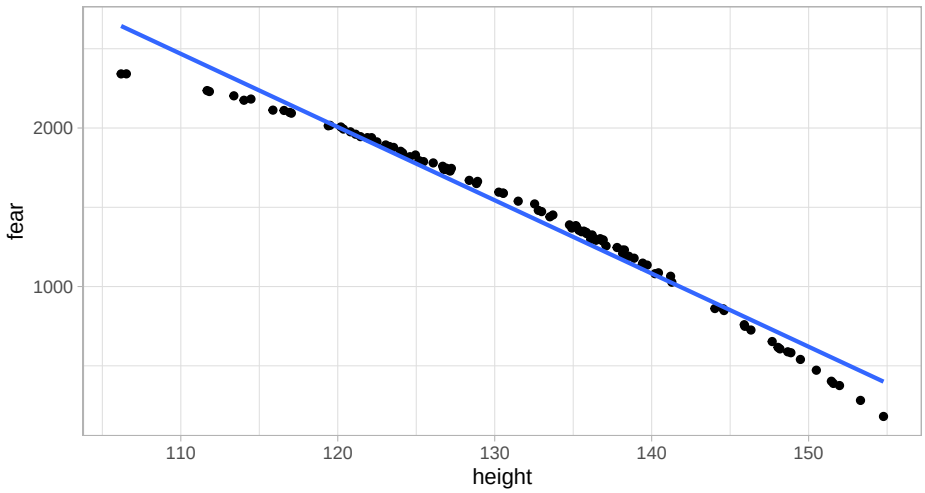


Figure 7.13: Least squares regression line for fear of snakes on height in 100 children.

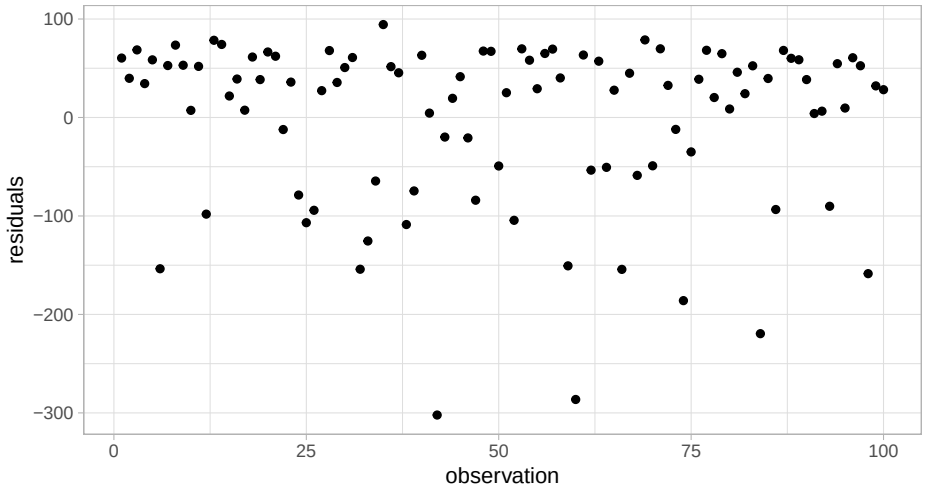


Figure 7.14: Residual plot after regressing fear of snakes on height.

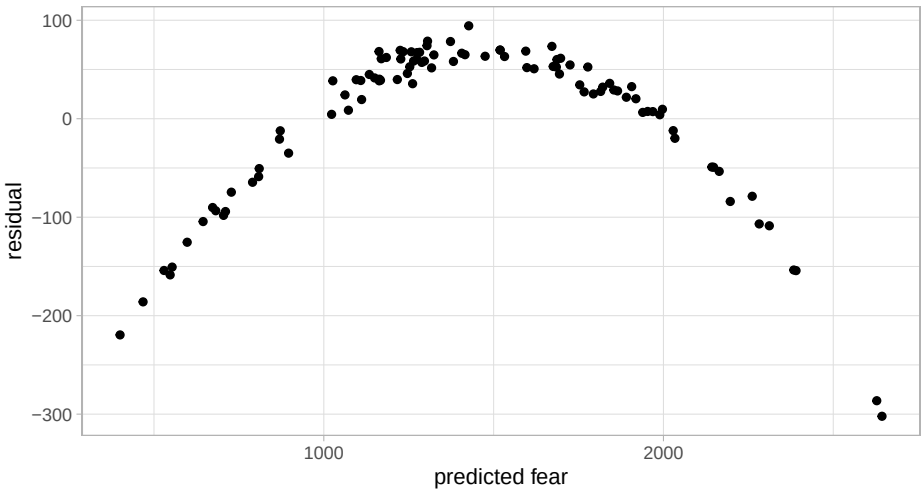


Figure 7.15: Residual plot after regressing fear of snakes on height.

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 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:

$$\widehat{\texttt{fear}} = -2000 + 100 \times \texttt{height} - 0.56 \times \texttt{height2} \tag{7.3}$$

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 (7.3) 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

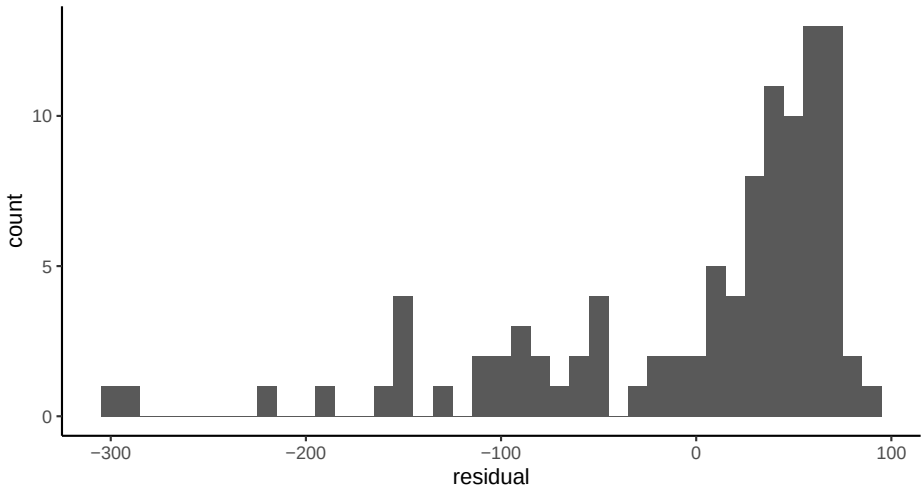


Figure 7.16: Histogram of the residuals after regressing fear of snakes on height.

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.

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

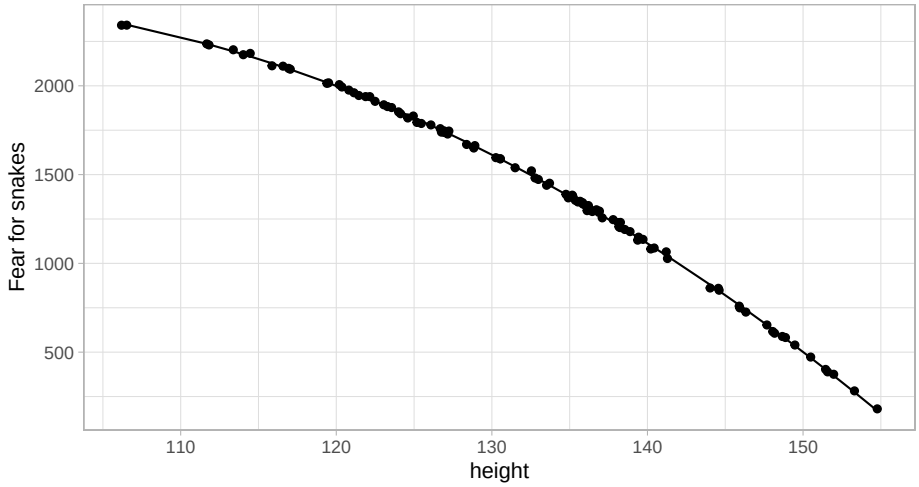


Figure 7.17: Observed and predicted fear based on a linear model with height and height squared

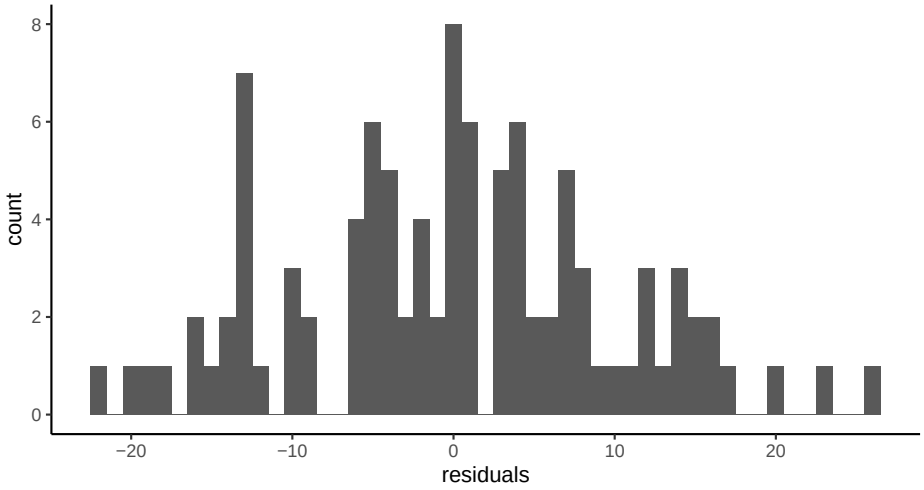


Figure 7.18: Histogram of the residuals of the fear of snakes data with height squared introduced into the linear model.

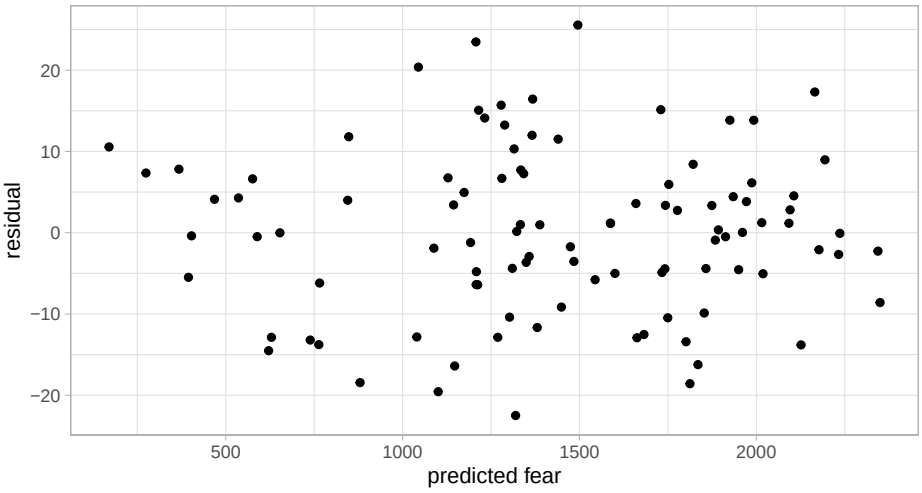


Figure 7.19: Residuals plot of the fear of snakes data with height squared introduced into the linear model.

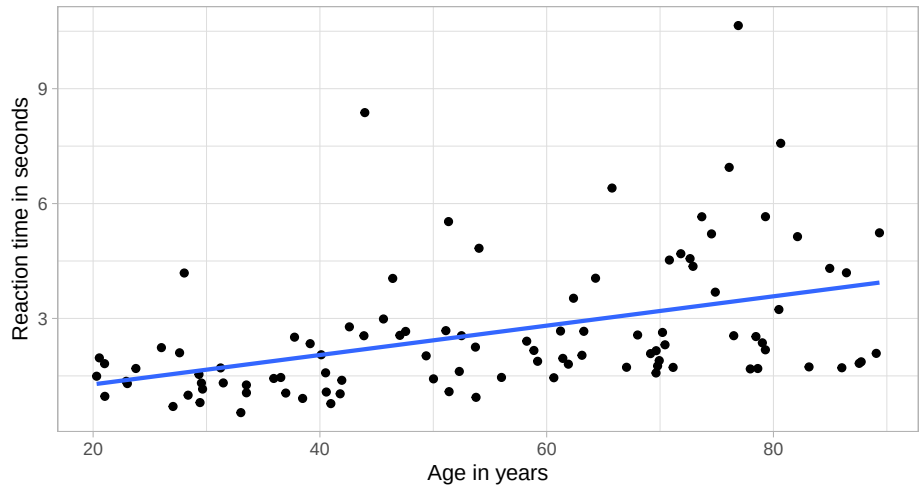


Figure 7.20: Least squares regression line for reaction time on age in 100 adults.

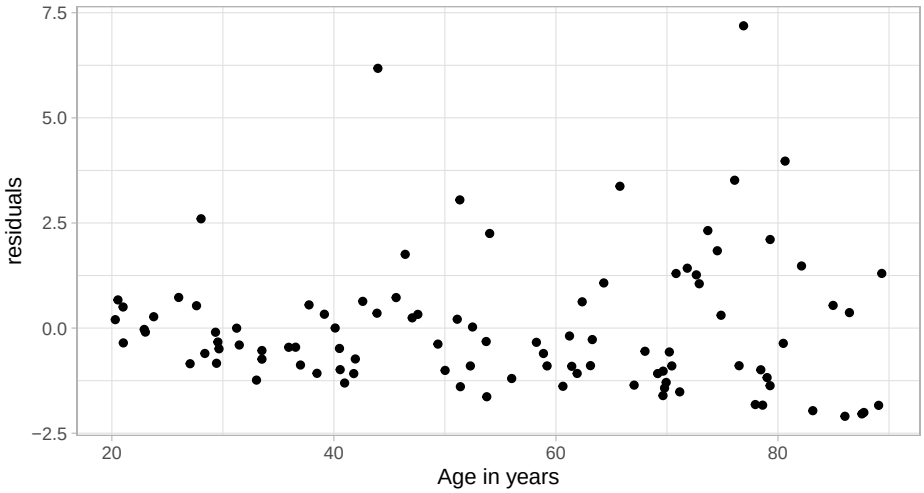


Figure 7.21: Residual plot after regressing reaction time on age.

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 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

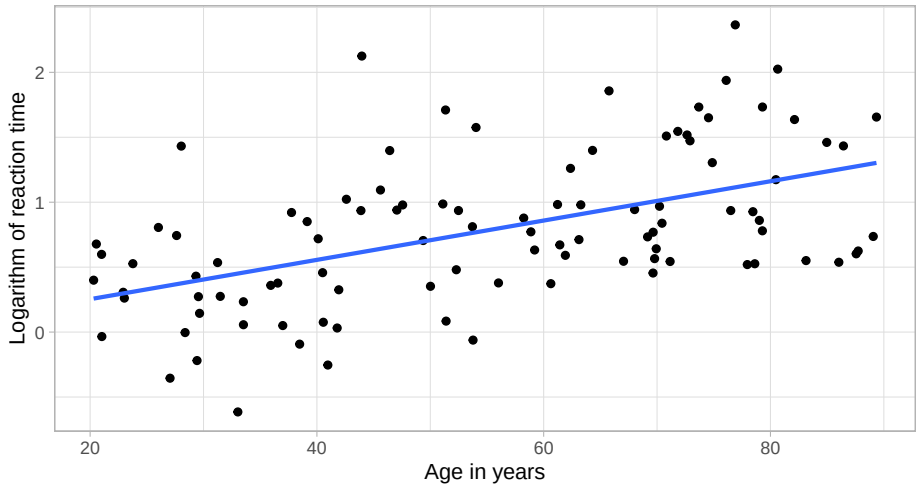


Figure 7.22: Least squares regression line for log reaction time on age in 100 adults.

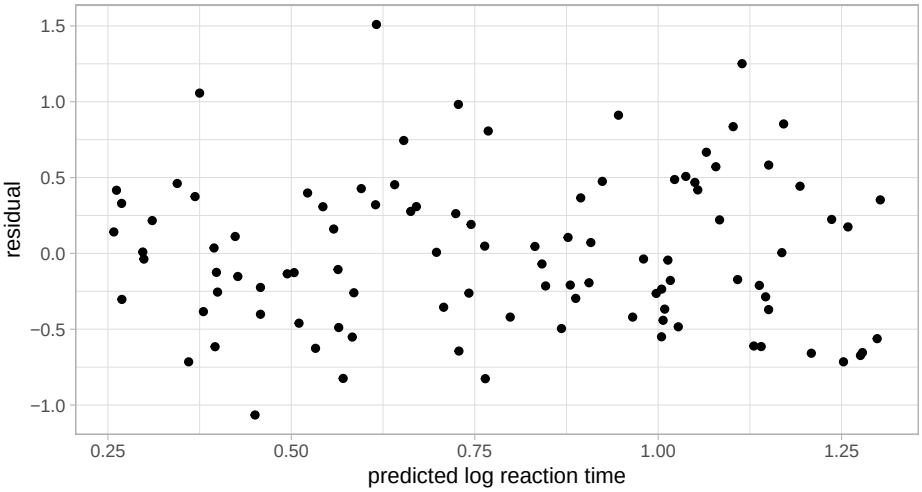


Figure 7.23: Residual plot after regressing log reaction time on age.

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 what the histogram of the residuals and the density look 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.

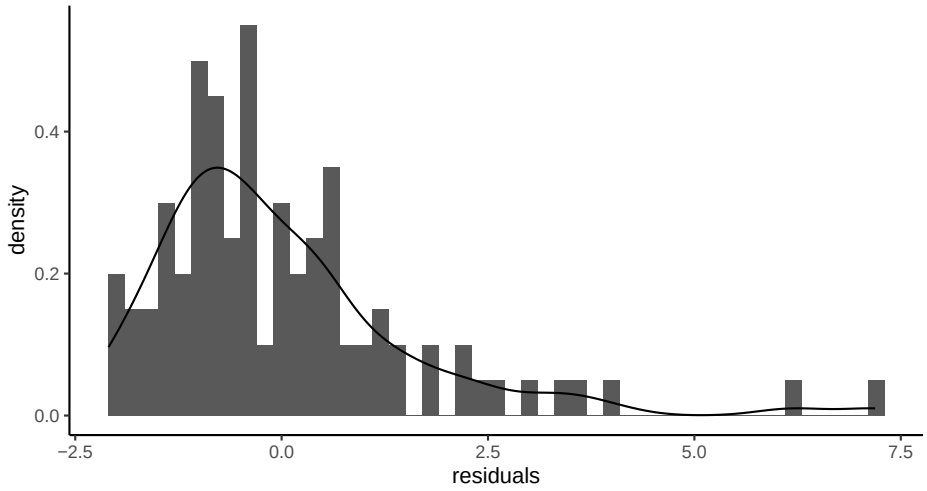


Figure 7.24: Histogram of the residuals after a regression of reaction time on age.

After a logarithmic transformation of the reaction times, we get the histogram and the density in Figure 7.25, which look 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

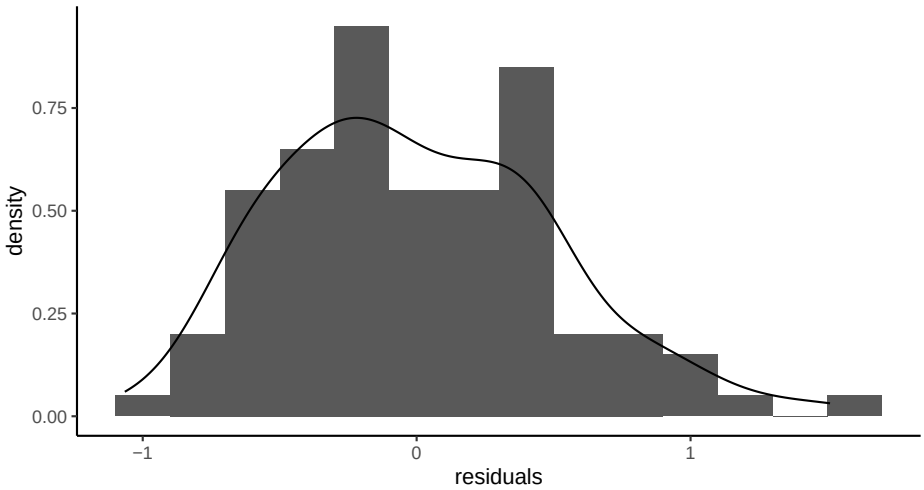


Figure 7.25: Histogram of the residuals after a regression of log reaction time on age.

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.

Judging normality from histograms is always a bit tricky. An often used alternative is a quantile-quantile plot, or qq-plot for short. It is based on the quantiles of a distribution. Remember quantiles from Chapter 1: If a value x is the 70th quantile of a distribution, it means that 70% of the values is smaller than x. When a student's performance is in the 90th percentile, it means that 90% of the other students show worse performance.

As we saw in Chapter 1, the quantiles of the normal distribution are known. In the qq-plot, we plot the observed quantiles of a given distribution (y-axis) against the expected quantiles of the normal distribution (x-axis). If the distribution is normal, then the data points should be on a straight line.

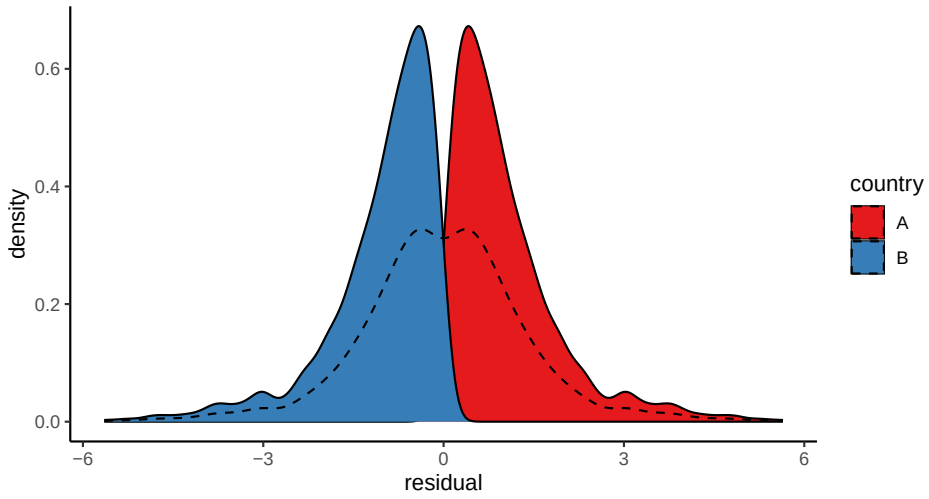


Figure 7.26: Two distributions might each be very non-normal, but the density of the combined data might look normal nevertheless (the dashed line). Normality should therefore always be checked for each subgroup separately.

Figure 7.27 shows the qq-plot based on the linear regression of response times (rt) on age. When the linear model is applied to the original response time data, the residuals are clearly not normally distributed, whereas when the linear model is applied to the response time after a logarithmic transformation, the data show a more or less normal distribution.

It should be noted that the assumption of normally distributed residuals is the least important assumption. Even when the 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 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. A histogram can be rather coarse if your sample size is limited. Add a density to get the general idea of the distribution. A qq-plot is also very powerful. If the histogram/density plot/qq-plot does not look symmetric/normal 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

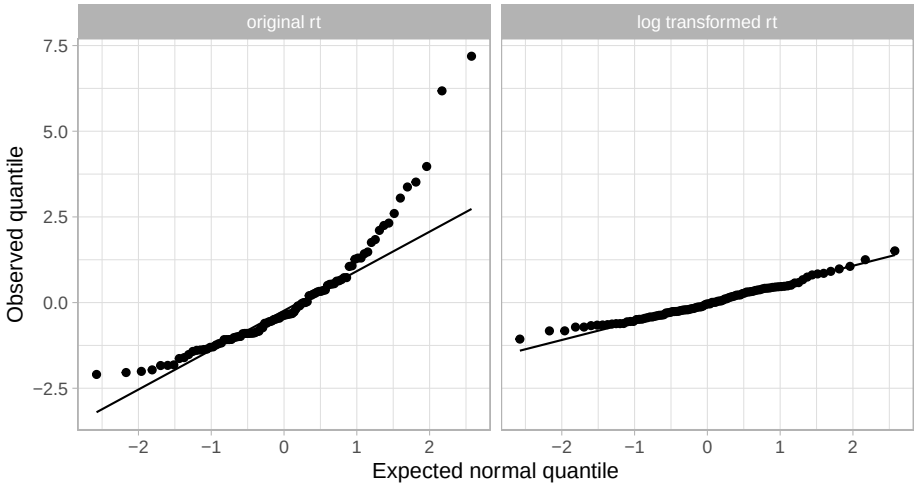


Figure 7.27: The qq-plot for the residuals based on the original response time data shows clear nonnormality (left panel), whereas the qq-plot for the residuals based on the log transformed data are close to the line, indicating normality

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.

7.7 Checking assumptions in R

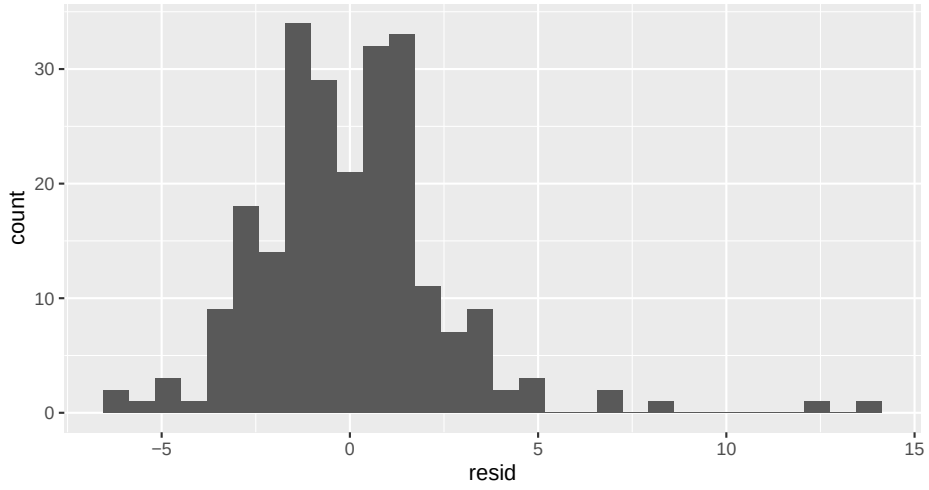
In this section we show the general code for making residual plots in R. We will look at how to make the three types of plots of the residuals to check the four assumptions.

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 = .)
```

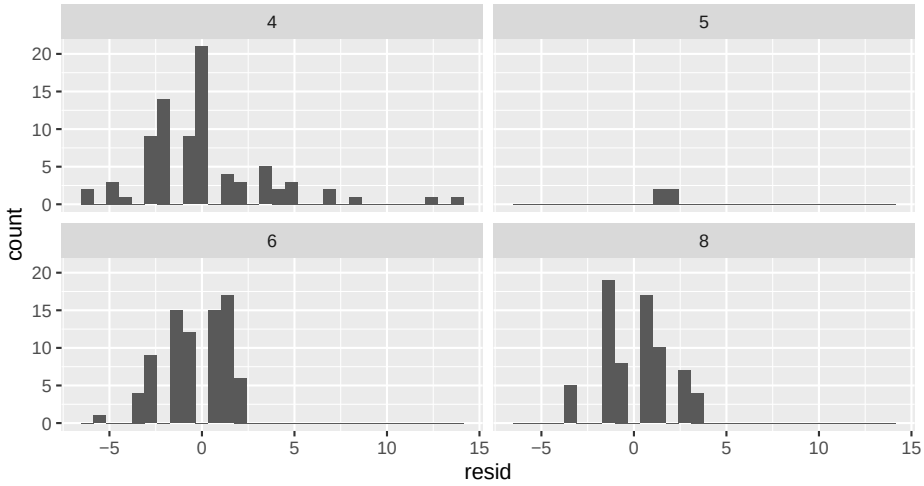
Next, we use the function `add_residuals()` from the `modelr` package to add residuals to the data set and plot a histogram.

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



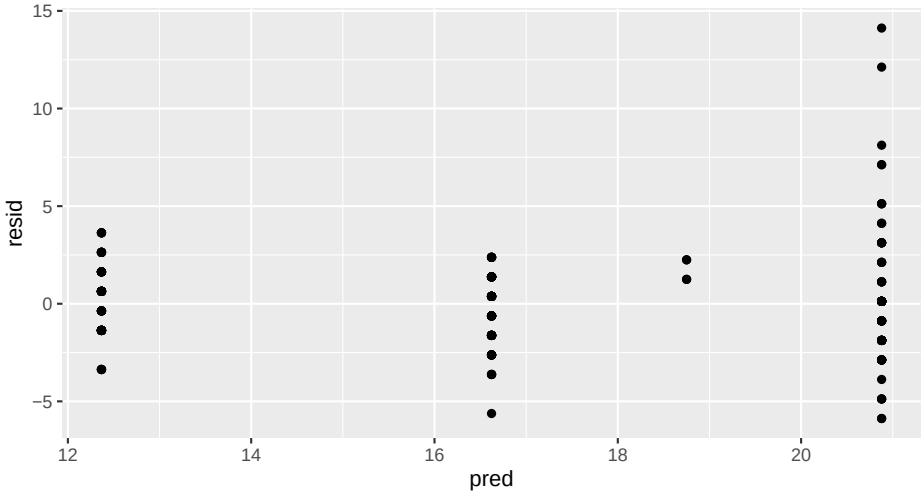
As stated earlier, it's even better to do this for the different subgroups separately:

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



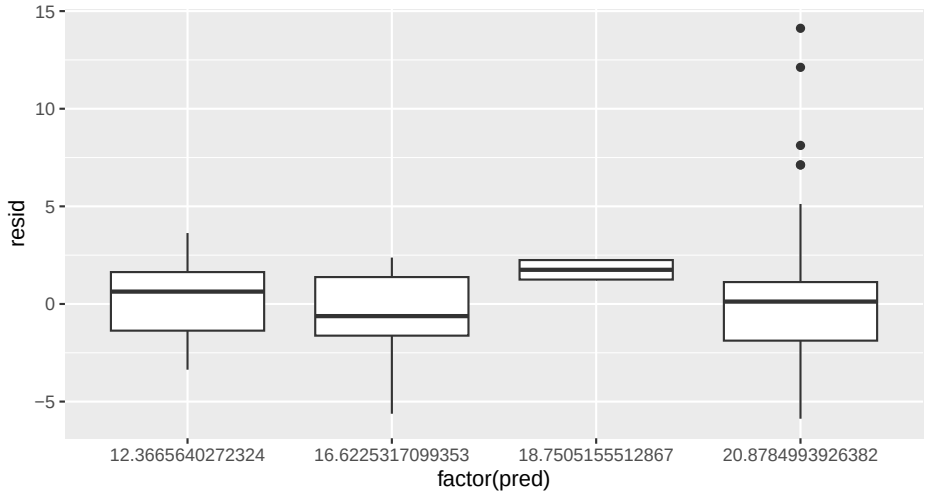
For the second type of plot, we use two functions from the `modelr` package to add predicted values and residuals to the data set, and use these to make a residual plot:

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



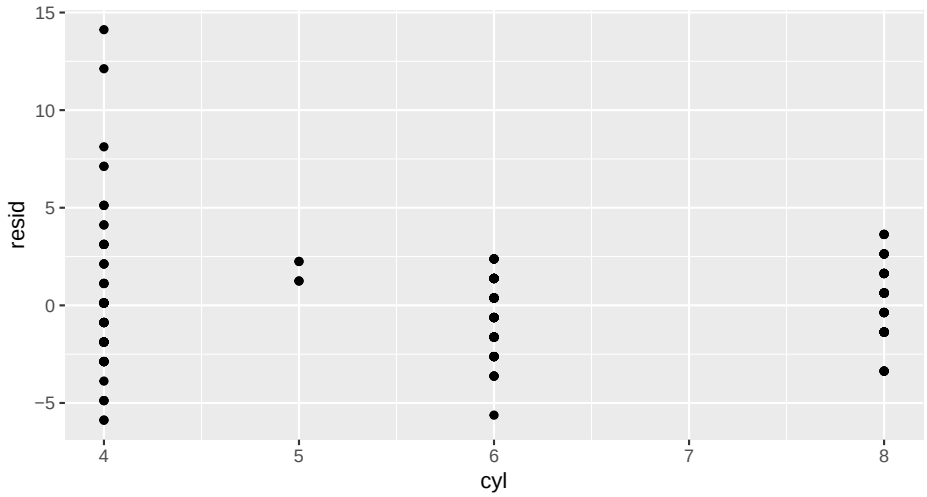
When there are few values for the predictions, or when you have a categorical predictor, it's better to make a boxplot:

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



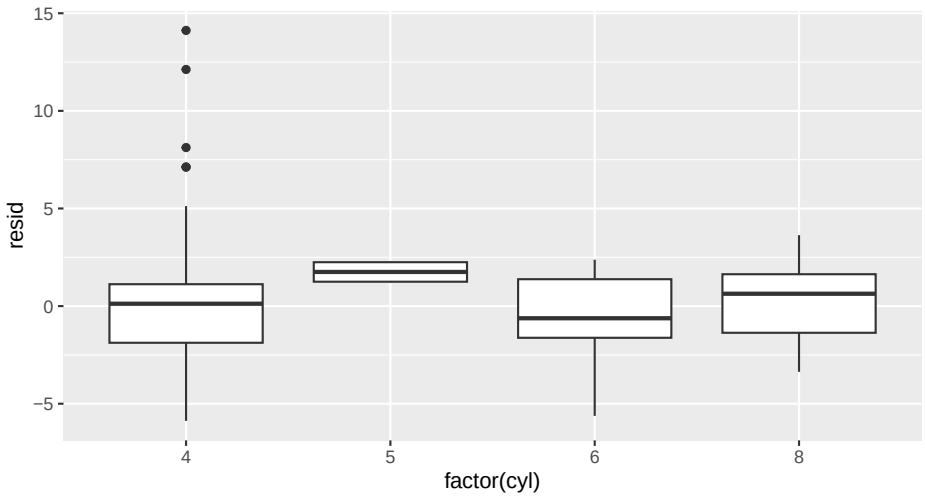
For the third type of plot, we put the predictor on the x-axis and the residual on the y-axis.

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



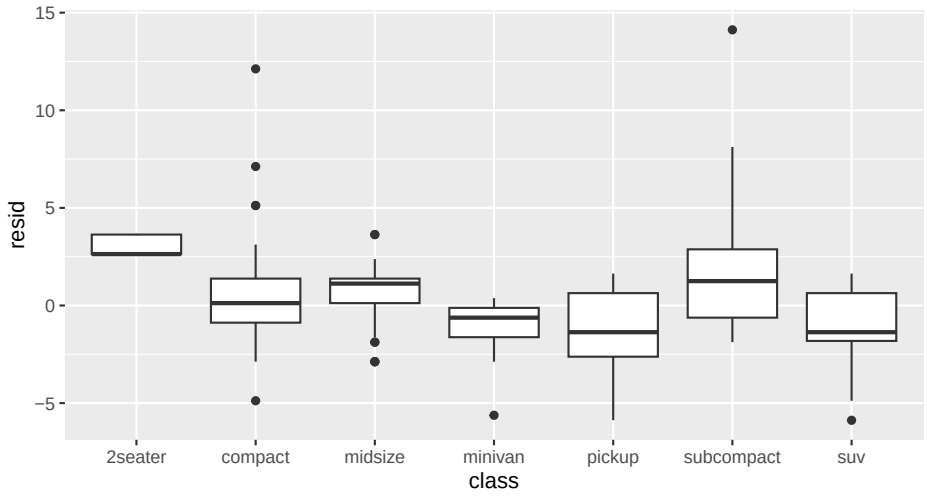
Again, with categorical variables or variables with very few categories, it is sometimes clearer to use a boxplot:

```
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()
```



7.8 Take-away points

- The general assumptions of linear models are *linearity* (additivity), *independence*, *normality* and *homogeneity of variance*.
- Linearity refers to the characteristic that the model equation is the *summation of parameters*, e.g. $b_0 + b_1X_1 + b_2X_2 + ...$
- Normality refers to the characteristic that the residuals are *drawn from a normal distribution*, i.e. $e \sim N(0, \sigma^2)$.
- Independence refers to the characteristic that the residuals are completely *randomly* drawn from the normal distribution. There is no systematic pattern in the residuals.
- Homogeneity of variance refers to the characteristic that there is only one normal distribution that the residuals are drawn from, that is, with *one specific variance*. Variance of residuals should be the same for every meaningful subset of the data.
- Assumptions are best checked visually.
- Problems can often be resolved by some transformation of the data, for example taking the logarithm of a variable, or computing squares.
- Inference is generally robust against violations of these assumptions, except for the independence assumption.

Key concepts

- Linearity
- Homogeneity of variance
- Heteroscedasticity
- Homoscedasticity
- Independence

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

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

AgeMother	Aggression
32	Sometimes aggressive
31	Often aggressive
32	Often aggressive
30	Not aggressive
31	Sometimes aggressive
30	Sometimes aggressive
31	Not aggressive
31	Often aggressive
31	Not aggressive
30	Sometimes aggressive
32	Often aggressive
32	Often aggressive
31	Sometimes aggressive
30	Sometimes aggressive
31	Not aggressive

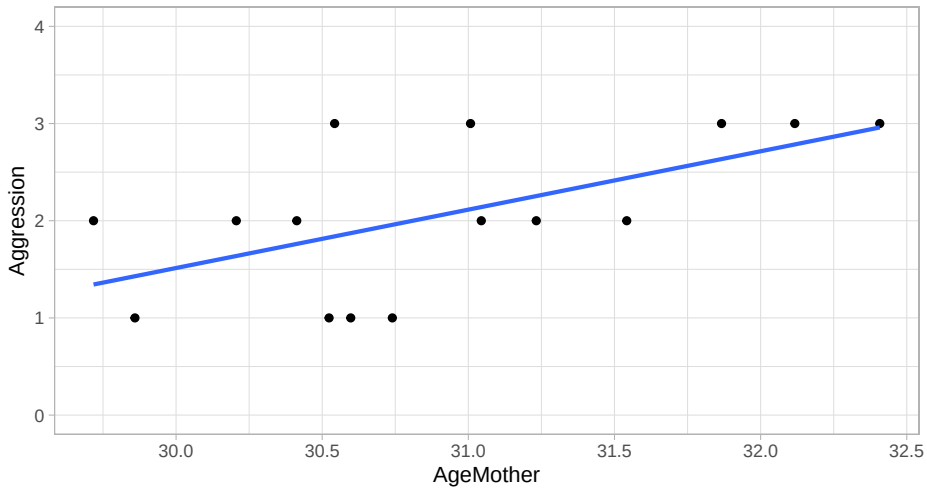


Figure 8.1: Regression of the child's aggression score (1,2,3) on the mother's age.

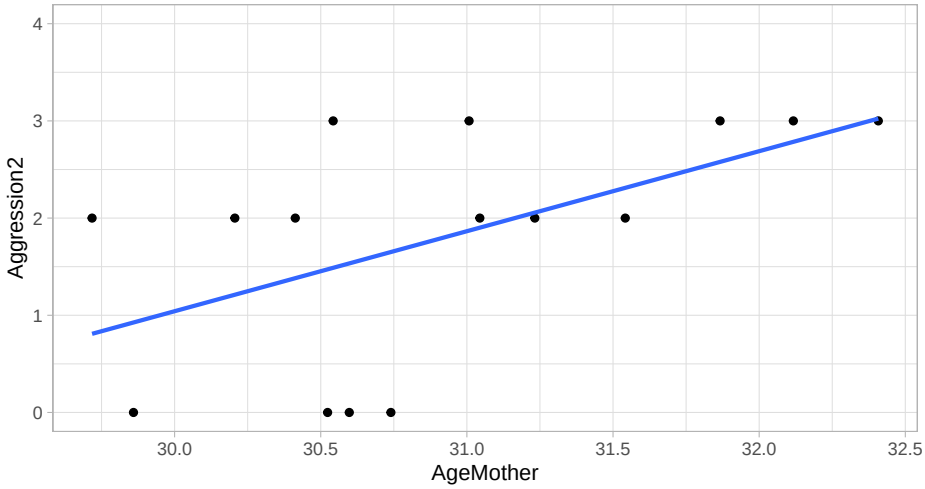


Figure 8.2: Regression of the child's aggression score (0,2,3) on the mother's age.

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 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 skewed, 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

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

AgeMother	Aggression
35	6
32	4
35	6
36	5
33	3
30	1
32	4
32	2
34	4
30	2
32	3
31	2
32	3
31	3
38	7

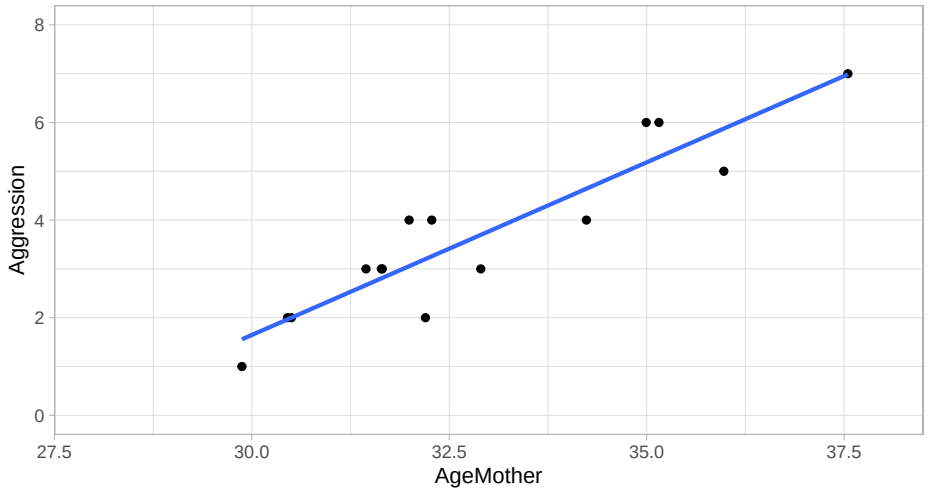


Figure 8.3: Regression of the child's aggression 1 thru 7 Likert score on the mother's age.

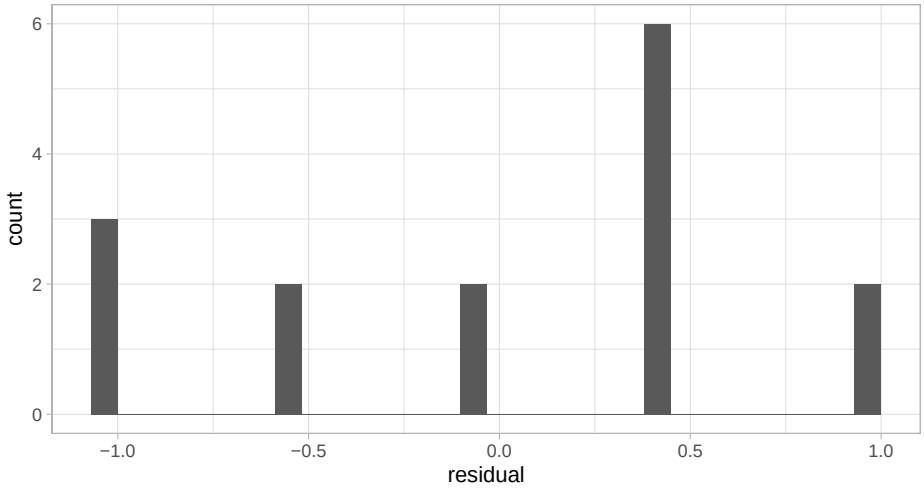


Figure 8.4: Histogram of the residuals after the regression of a child's aggression score on the mother's age.

Table 8.3: Results of the 2023 London marathon

contestant	**time**	**rank**
Sifan Hassan	2.18.33	1
Alema Megertu	2.18.37	2
Peres Jepchirchir	2.18.38	3
Shelia Chepkirui	2.18.51	4

the regression line, and that directly affects the relative sizes of the residuals.

8.2 Analysing ranked data

Many of the non-parametric methods discussed in this book are actually methods strongly related to linear models. Often the only difference is that instead of analysing the original data values, we analyse *ranks*. For instance, let's look at the times in which four runners finished in the 2023 London marathon, see Table 8.3.

Note that by ranking we lose some of the original information. Based on solely the ranks, we don't see anymore that Hassan finished 4 seconds earlier than Megertu, that Megertu and Jepchirchir crossed the finish line almost together, and that Chepkirui lagged somewhat behind, finishing 13 seconds after Jepchirchir.

Table 8.4: 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

As said, many non-parametric methods discussed in this book are simply linear models applied to ranks, rather than the original data. One example is Spearman's rho, an alternative to Pearson's correlation.

First, we will discuss non-parametric alternatives for two numeric variables. We will start with Spearman's ρ (rho, pronouned "roe"), also called Spearman's rank-order correlation coefficient r_s. Next we will discuss an alternative to r_s, Kendall's τ (tau, pronounced "taw"). After that, we will discuss the combination of numeric and categorical variables, when comparing groups.

8.3 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 8.4. We see that student 9 is the brightest student in both geography and history, and student 7 is the dullest student in both subjects.

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

Table 8.5: 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

Table 8.6: 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

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 8.5.

Theoretically, any difference in ranking could be as large as 9 (when one student is ranked first on one subject, and ranked last on the other subject), but here we see a maximum 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 8.6.

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):

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

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:

$$\begin{aligned} r_s &= 1 - \frac{6 \times 10}{10^3 - 10} \\ &= 1 - \frac{60}{990} \\ &= 0.94 \end{aligned}$$

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 8.7 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.

Actually it can be shown that (8.1) is the same as the formula for Pearson's correlation coefficient on the ranks, see Chapter 4. The only difference between Pearson's correlation and Spearman's correlation, is that the former is computed using the original values, and the latter is actually the Pearson's correlation using the ranks as values.

Table 8.7: Ordinal and numeric variables and their ranked transformations.

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

8.4 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 same function `cor()` that we use for Pearson's correlation, but indicate explicitly that we want Spearman's rho:

```
cor(mpg$cyl, mpg$year, method = "spearman")
```

```
## [1] 0.1192822
```

Spearman's rho is here equal 0.12.

> "We computed a Spearman correlation that quantifies the extent to which the ranking of the cars in terms of cylinders is related to the ranking of the cars in terms of year. Results showed a weak positive association, $\rho(n = 234) = .12$."

Spearman equals Pearson applied to ranks

To show to you that Spearman's rho is the same as a Pearson correlation based on ranks, we can use the Pearson correlation applied to the ranks:

```
# compute the ranks
cyl_ranked <- mpg$cyl %>% rank()
year_ranked <- mpg$year %>% rank()
# compute Pearson's correlation
cor(cyl_ranked, year_ranked) # standard Pearson correlation
```

```
## [1] 0.1192822
```

Table 8.8: 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

8.5 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 8.8.

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 8.8. 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

Table 8.9: 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

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 8.8 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 8.9. 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 8.9. 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 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 8.9, 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:

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

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.

8.6 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 same `cor()` function we use for Pearson's correlation and Spearman's correlation. By default, Pearson is computed, but you can also indicate you want something else, in this case Kendall.

```
cor(mpg$cyl, mpg$year, method = "kendall")
```

```
## [1] 0.1119214
```

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).

```
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         year          cty
## cyl  0.000000e+00 0.0001896548 7.706570e-67
## year 1.896548e-04 0.0000000000 5.923697e-03
## cty  7.706570e-67 0.0059236967 0.000000e+00
##
## $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 0.1642, with an associated p-value of 0.000189.

We can report this in the following way:

> "The null-hypothesis was tested that there is no correlation between cyl and year, when one controls for cty. This was tested with a Kendall correlation coefficient. We found that the hypothesis of a Kendall correlation of 0 in the population of cars (controlling for cty) could be rejected, $\tau(n = 234) = .16, p < .001$."

8.7 Kruskal-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

Table 8.10: Field trip data.

student	group	size	rank
1	math	extra small	1.0
2	math	extra large	6.0
3	psych	medium	4.0
4	psych	small	2.5
5	engineer	large	5.0
6	math	small	2.5

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 Kruskal-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 8.10, 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.

H_0: All groups have the same average (median) rank.

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 for this null-hypothesis, a rather complex formula is used to compute the so-called KW-statistic, see Castellan & Siegel (1988), that you don't need to know. The distribution of this KW-statistic under the null-hypothesis can be approximated by a chi-square distribution. In this way we know what extreme values are, and consequently can compute p-values. This

computation can be done in R. Alternatively, instead of using the chi-square approximation of the KW statistic, we can also apply an ANOVA on the ranks, and use an F-test as an approximation. Both methods converge to the same conclusion for moderate to large sample sizes.

8.8 Kruskal-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 differences in the distribution of city miles per gallon. Figure 8.5 shows a box plot. The variances are quite different for each type of car (no homogeneity of variance) and there are quite a few extreme values, making it unlikely that residuals will have a normal distribution in the population. We therefore consider doing a non-parametric test. Based on Figure 8.5, it seems that indeed the medians 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 hypothesis, we can make it more specific and we can use a Kruskal-Wallis test to test the null-hypothesis that the average (median) rank of **cty** is the same in all groups.

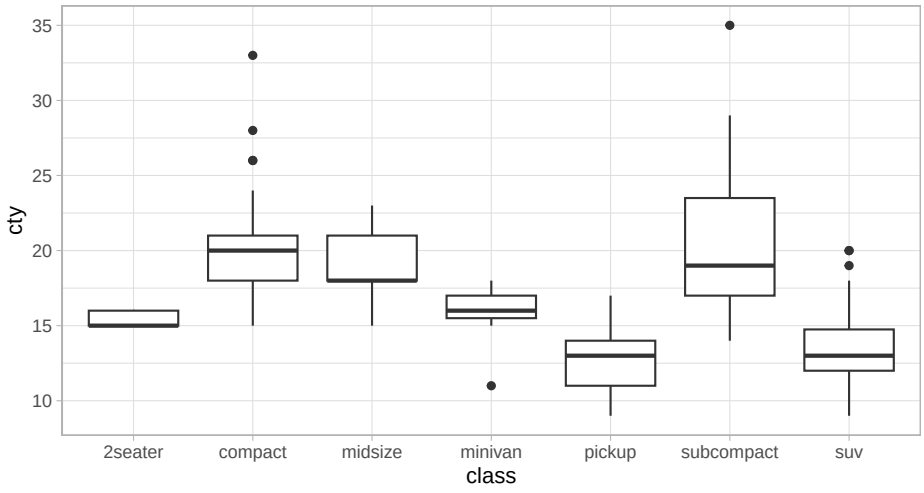


Figure 8.5: Distributions of city mileage (cty) as a function of car type (class).

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 < 2.2e-16
```

```
mpg$cty %>% length()
```

```
## [1] 234
```

```
mpg$class %>% length()
```

```
## [1] 234
```

The output allows us to report the following:

> "The null-hypothesis that city miles per gallon is distributed equally for all types of cars was tested using a Kruskal-Wallis test with an α of 0.05. Results showed that the null-hypothesis could be rejected, $X^2(6, N = 234) = 149.53$, $p < .001$."

Kruskal-Wallis equals ANOVA applied to ranks

As the Kruskal-Wallis is actually an ANOVA on ranks, we can also first rank the dependent variable, and then apply the regular analysis of variance:

```
library(car)
mpg %>%
  mutate(cty_ranked = cty %>% rank()) %>% # compute ranks
  lm(cty_ranked ~ class, data = .,
     contrasts = list(class = contr.sum)) %>%
  Anova(type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: cty_ranked
##              Sum Sq  Df F value    Pr(>F)
## (Intercept) 1584608   1 946.952 < 2.2e-16 ***
## class        680479   6  67.775 < 2.2e-16 ***
## Residuals    379857 227
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You then report instead:

> "The null-hypothesis that city miles per gallon is distributed equally for all types of cars was tested using a Kruskal-Wallis test (ANOVA on ranks) with an α of 0.05. Results showed that the null-hypothesis could be rejected, $F(6, 227) = 67.78$, $p < .001$."

8.9 Take-away points

- When a distribution of residuals looks very far removed from a normal distribution, or looks vary heteroskedastic, consider using a non-parametric method of analysis.
- When analysing variables that are ordinal or when linear model assumptions cannot be met otherwise, consider non-parametric methods.
- Many of the non-parametric methods are based on ranks, rather than the original values.
- Spearman's correlation is actually a Pearson correlation based on ranked data.
- A Kruskal-Wallis test is actually an analysis of variance based on a ranked dependent variable.

Key concepts

- Ranks
- Spearman's rank-order correlation coefficient ρ or r_s
- Kendall's τ
- Kruskal-Wallis test

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:

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

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:

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

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

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

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.

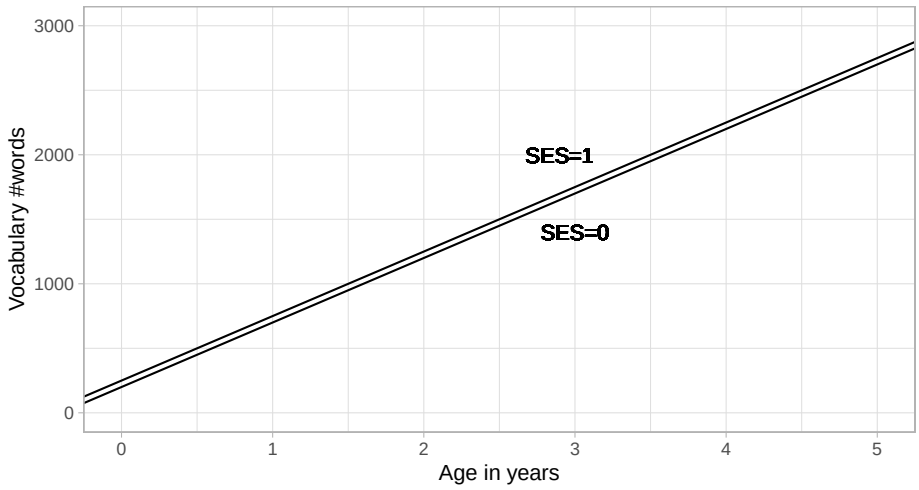


Figure 9.1: Two regression lines: one for low SES children and one for high SES children.

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**:

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

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**:

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

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

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

Multiplying this out gets us:

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

If we rearrange the terms a bit, we get:

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

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:

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

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:

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

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):

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

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** × **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* or the *moderator*: it modifies/moderates the relationship between **age** and vocabulary. Note however that you could also interpret **age** as the modifier variable (moderator): 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.

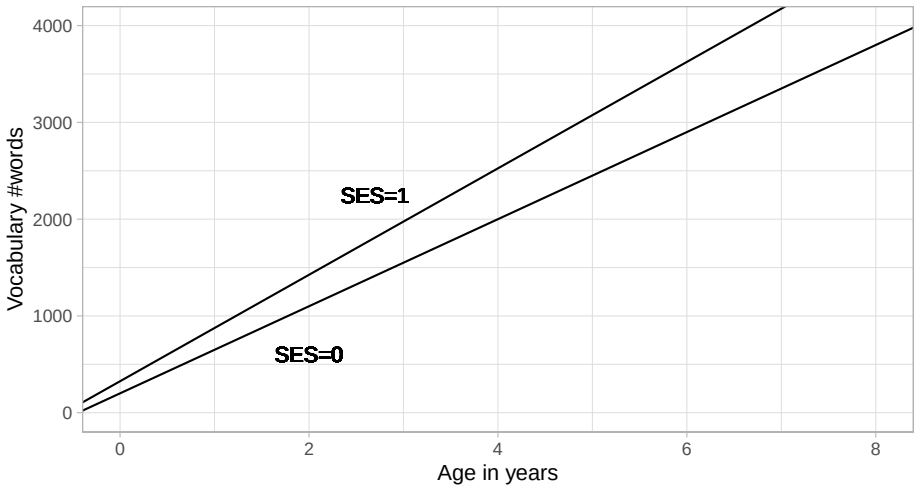


Figure 9.2: Two regression lines for the relationship between age and vocab, one for low SES children (SES = 0) and one for high SES children (SES = 1).

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 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, 5~
## $ Time   <dbl> 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 21, 0, 2, 4, 6, 8, 10, 1~
## $ Chick  <ord> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 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, 1, ~
```

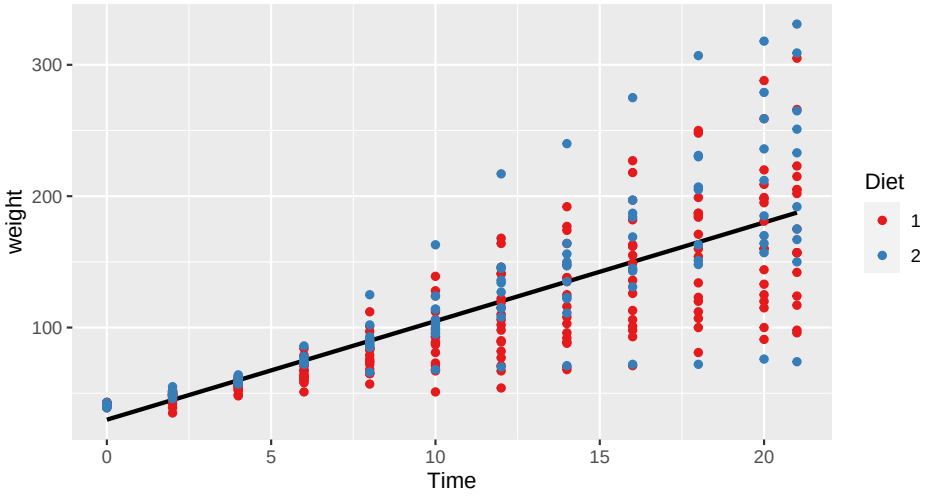


Figure 9.3: The relationship between Time and weight in all chicks with either Diet 1 or Diet 2.

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**:

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

```
## # 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 **Diet1**. 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** $\times$ **Diet2**, by multiplying the variables **Time** and **Diet1**. Results show that this interaction effect is 1.77.

These results can be plugged into the following regression equation:

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

If we fill in 1s for the **Diet1** dummy variable, we get the equation for chicks with Diet 2:

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

If we fill in 0s for the **Diet1** dummy variable, we get the equation for chicks with Diet 1:

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

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*

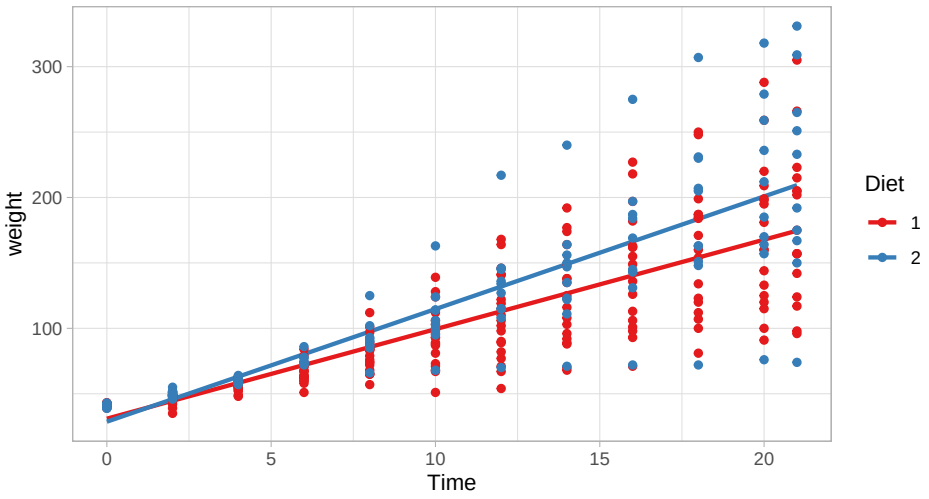


Figure 9.4: The relationship between Time and weight in chicks, separately for Diet 1 and Diet 2.

of chicks equals 0? In other words, is the value of 1.77 significantly different from 0?

The null-hypothesis is

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

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, $t(336) = 2.92, p = .004$."

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 = TRUE)
```

```
## # 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

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

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 **Diet1**, **Diet3** and **Diet4**. If we fill in these 0s, we obtain

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

For the chicks on Diet 2, we have 1s for the dummy variable **Diet1** and 0s for the other dummy variables. Hence we have

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

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 10.

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

$$\begin{aligned}\widehat{\texttt{weight}} &= 30.93 + 6.84 \times \texttt{Time} - 12.68 \times 1 + 4.58 \times \texttt{Time} \times 1 \\ &= (30.93 - 12.68) + (6.84 + 4.58) \times \texttt{Time} \\ &= 18.25 + 11.42 \times \texttt{Time}\end{aligned}$$

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:

$$\widehat{\texttt{weight}} = 30.79 + 9.71 \times \texttt{Time}$$

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.

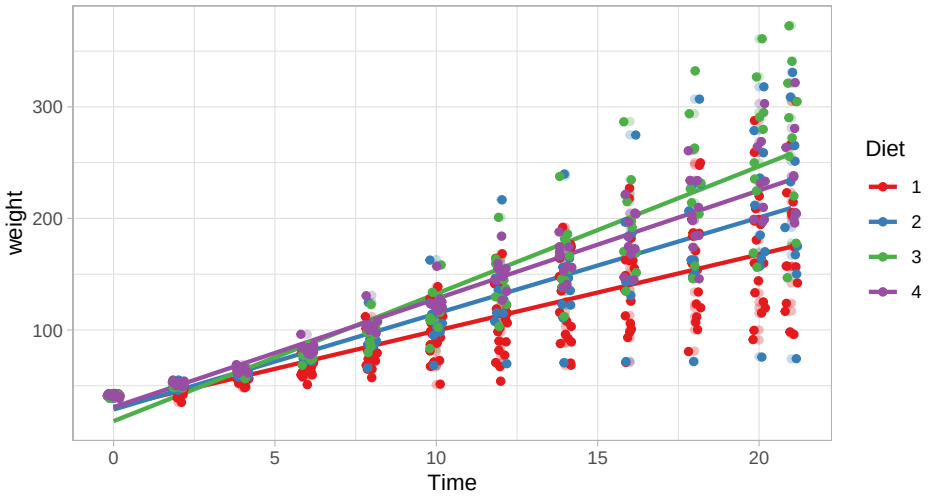


Figure 9.5: Four different regression lines for the four different diet groups.

9.4 Linear model versus ANOVA

In this book we approach all of the data analysis problems starting from the linear model, where we only use numerical variables. In the event we have

categorical variables, we transform them to dummy variables or sets of dummy variables that are in turn treated as separate numerical variables. We've also seen Analysis of Variance, a concept closely linked to the linear model. ANOVA is used a lot in practice, but it is important to realise that ANOVA and the linear model can both be used in many different ways.

For example, R by default transforms categorical variables into (sets of) dummy variables (dummy coding). An alternative way of recoding categorical variables is using sum-to-zero coding. The reference group is then coded as -1, the alternative group is coded as 1, and the other groups are coded as 0. We will discuss such alternative ways of coding categorical variables more extensively in Chapter 10. For now it suffices to realise that sum-to-zero coding leads to a different regression table than with dummy coding.

If we go back to the example of the Chickweight data, we can run a linear model with a main effect of the categorical variable Diet.

```
ChickWeight %>%
  lm(weight ~ Diet, data = .) %>%
  tidy()
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    103.       4.67     22.0  4.71e-78
## 2 Diet2           20.0      7.87      2.54 1.14e- 2
## 3 Diet3           40.3      7.87      5.12 4.11e- 7
## 4 Diet4           32.6      7.91      4.12 4.29e- 5
```

The Diet2, Diet3 and Diet4 variables are dummy variables. We see that the difference between Diet 2 and Diet 1 (reference) equals 20. If, instead of the default dummy coding, we apply sum-to-zero coding,

```
ChickWeight %>%
  lm(weight ~ Diet, data = .,
     contrasts = list(Diet = contr.sum)) %>%
  tidy()
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   126.        2.99    42.2   7.24e-178
## 2 Diet1         -23.2       4.45    -5.21  2.59e-  7
## 3 Diet2          -3.25      5.38    -0.604 5.46e-  1
## 4 Diet3          17.1       5.38     3.18  1.58e-  3
```

we get very different model parameters. This is because R doesn't use dummy coding, but codes three variables where for each variable, Diet 4 is coded as -1, and the respective other diets are coded as 1. For example, the new variable **Diet1** equals 1 for Diet 1, 0 for Diets 2 and 3, and -1 for the reference group Diet 4. For now, it is sufficient to know that most ANOVAs in social sciences are done using this particular way of coding (also known as effects coding).

Apart from the type of coding categorical variables, the way that sums of squares are calculated for an ANOVA can also differ. If you have only one independent variable, the computation of the sums of squares is exactly like explained in Chapter 6. But if you have two or more independent variables, the method of calculation has an effect on the numbers in your ANOVA table, including the statistical inference.

For example, if we apply sum-to-zero coding and apply the default way of calculating sums of squares (called Type II sums of squares) we get the following ANOVA table.

```
library(car)
ChickWeight %>%
  lm(weight ~ Time + Diet + Time:Diet, data = .,
     contrasts = list(Diet = contr.sum)) %>%
  Anova()
```

```
## Anova Table (Type II tests)
##
## Response: weight
##             Sum Sq  Df  F value    Pr(>F)
## Time       2016357   1 1737.367 < 2.2e-16 ***
## Diet        129876   3   37.302 < 2.2e-16 ***
## Time:Diet    80804   3   23.208 3.474e-14 ***
## Residuals   661532 570
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we apply Type III sums of squares, which is what is most commonly done in the social sciences, we obtain

```
ChickWeight %>%
  lm(weight ~ Time + Diet + Time:Diet, data = .,
     contrasts = list(Diet = contr.sum)) %>%
  Anova(type = 3)
```

```
## Anova Table (Type III tests)
##
```

```
## Response: weight
##               Sum Sq  Df   F value    Pr(>F)
## (Intercept)  111618   1   96.1745 < 2.2e-16 ***
## Time        2056347   1 1771.8235 < 2.2e-16 ***
## Diet           3993   3    1.1469    0.3295
## Time:Diet     80804   3   23.2079 3.474e-14 ***
## Residuals    661532 570
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the sums of squares, the F-values and the p-values are different for the main effects of Time and Diet, but that the numbers for the Time by Diet interaction effect are exactly the same.

Summarising, it is perfectly fine to run a linear model using dummy coding and interpreting the regression coefficients as we have done throughout this book. However, when reporting an ANOVA, it is best to follow convention, and apply sum-to-zero coding and Type III sums of squares. It is beyond the goals of this book to delve into the how and why.

As an aside, note that for inference it is not all that important. In a linear model with an interaction effect, one is almost always only interested in the interaction effect. As the interaction effect is not affected by the choice of type II or type III sums of squares, and also not by the type of coding, the statistical conclusion will be the same. Only the main effects will be affected. As already stressed earlier, when reporting, stick to the statistical output that relates to the research questions.

Let's look at an example of applying an ANOVA in R, with a particular null-hypothesis that can only be tested using ANOVA.

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 can test this null hypothesis

$$H_0 : \beta_{\texttt{Time}\times\texttt{Diet2}} = \beta_{\texttt{Time}\times\texttt{Diet2}} = \beta_{\texttt{Time}\times\texttt{Diet4}} = 0$$

by running an ANOVA. That is, we apply the `Anova()` function of the `car` package to the results of an `lm()` analysis. Note however that we slightly alter the `lm()` analysis, similar to what we did in Chapter 6, changing dummy coding into sum-to-zero coding by adding `contrasts = list(Diet = contr.sum)`.

```
library(car)
out <- ChickWeight %>%
  lm(weight ~ Time + Diet + Time:Diet, data = .,
     contrasts = list(Diet = contr.sum))
out_anova <- out %>%
```

```
  Anova(type = 3)
out_anova %>%
  tidy()
```

```
## # A tibble: 5 x 5
##   term           sumsq    df statistic    p.value
##   <chr>          <dbl> <dbl>     <dbl>      <dbl>
## 1 (Intercept)  111618.     1     96.2   4.38e- 21
## 2 Time        2056347.     1   1772.    4.85e-177
## 3 Diet           3993.     3      1.15  3.30e-  1
## 4 Time:Diet     80804.     3     23.2   3.47e- 14
## 5 Residuals    661532.   570     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, $F(3, 570) = 23.2, MSE = 1161, p < .001$."

The MSE is computed as the sum of squared residuals, 661532, divided by the residual degrees of freedom, 570, equals $\frac{661532}{570} = 1161$ (Ch. 6).

9.5 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
##  $ 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 = TRUE)
```

```
## # 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 the effect of one independent variable (**species**) as a function of another independent 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 = TRUE)
```

```
## # 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`).

$$\widehat{\texttt{flipperlength}} = 188 + 3.94 \times \texttt{speciesChinstrap} + 4.62 \times \texttt{sexmale} + 3.56 \times \texttt{speciesChinstrap} \times \texttt{sexmale}$$

From this we can make the following predictions. The predicted flipper length for female Adelie penguins is

$$188 + 3.94 \times 0 + 4.62 \times 0 + 3.56 \times 0 \times 0 = 188$$

The predicted flipper length for male Adelie penguins is

$$\begin{aligned} 188 + 3.94 \times 0 + 4.62 \times 1 + 3.56 \times 0 \times 1 \\ = 188 + 4.62 = 192.62 \end{aligned}$$

The predicted flipper length for female Chinstrap penguins is

$$\begin{aligned} 188 + 3.94 \times 1 + 4.62 \times 0 + 3.56 \times 1 \times 0 \\ = 188 + 3.94 = 191.94 \end{aligned}$$

and the predicted flipper length for male Chinstrap penguins is

$$\begin{aligned} 188 + 3.94 \times 1 + 4.62 \times 1 + 3.56 \times 1 \times 1 \\ = 188 + 3.94 + 4.62 + 3.56 \\ = 200.12 \end{aligned}$$

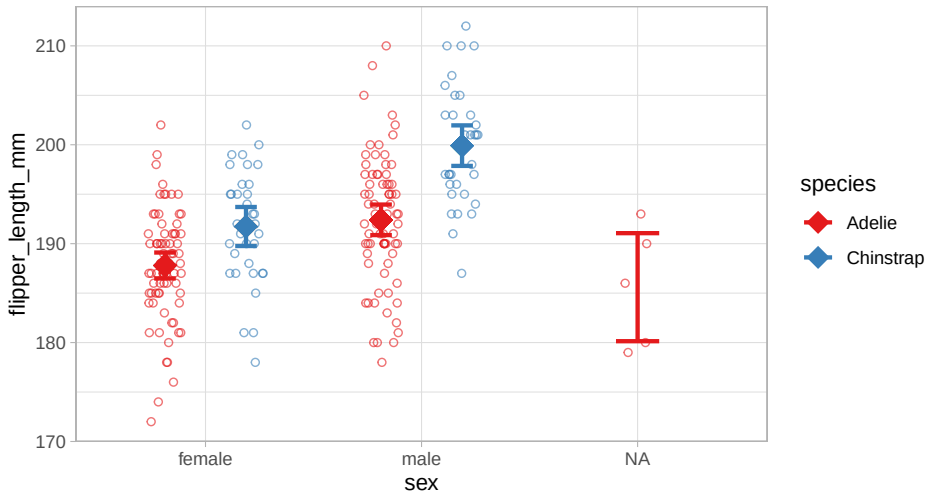
These predicted flipper lengths for each male/species combination are actually the group means. Group means and standard deviations can be computed by

```
penguin_data %>%
  group_by(sex, species) %>%
  summarise(mean = mean(flipper_length_mm, na.rm = T),
            sd = sd(flipper_length_mm, na.rm = T))
```

```
## # A tibble: 5 x 4
## # Groups:   sex [3]
##   sex    species    mean    sd
##   <fct>  <fct>     <dbl> <dbl>
## 1 female Adelie     188.  5.60
## 2 female Chinstrap  192.  5.75
## 3 male   Adelie     192.  6.60
## 4 male   Chinstrap  200.  5.98
## 5 <NA>   Adelie     186.  6.11
```

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 diamond-shaped means (`shape = 18`) are plotted with intervals that are twice (`mult = 2`) the standard error of those means (`geom = "errorbar"`).

```
penguin_data %>%
  left_join(penguin_data %>%                # adding group means to the data set
              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_colour_brewer(palette = "Set1") + # use nice colours
  theme_light() # use nice theme
```



This plot shows also the data on penguins with unknown sex (`sex = NA`). If we leave these out, we get

```
penguin_data %>%
  left_join(penguin_data %>%    # adding group means to the data set
              group_by(species, sex) %>%
              summarise(mean = mean(flipper_length_mm))
  ) %>%
  filter(!is.na(sex)) %>%   # filter out cases with missing sex data
  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_colour_brewer(palette = "Set1") + # use nice colours
  theme_light() # use nice theme
```

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 males are farther apart than the means in the females. Thus, in males the effect of species is larger than in females. 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 the moderating effect of sex on the effect of species is probably somewhere between 0.06 and 7.06 mm in the population of all penguins.

Instead of presenting the linear model results in the form of a regression table, we can also perform an ANOVA. We do that as follows, where we have to indicate we want sum-to-zero coding for both variables species and sex:

```
out <- penguin_data %>%
  lm(flipper_length_mm ~ species + sex + species:sex,
     data = .,
     contrasts = list(species = contr.sum, sex = contr.sum))
out %>% Anova(type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: flipper_length_mm
##              Sum Sq  Df    F value    Pr(>F)
## (Intercept) 6909660   1 1.8941e+05 < 2.2e-16 ***
## species        1518   1 4.1620e+01 7.518e-10 ***
## sex            1898   1 5.2031e+01 9.759e-12 ***
## species:sex     147   1 4.0293e+00     0.046 *
## Residuals      7661 210
```

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

Note that the statistical results for the interaction effect are the same as for the linear model with dummy coding. We can report either based on the linear model, reporting the t-value:

> "We found an interaction effect where the effect of sex was larger in Chinstrap than in Adelie penguins (3.56 mm, 95% CI 0.06 – 7.06, $t(210) = 2.01$, $p = 0.046$)."

or based on the ANOVA, reporting the F-value:

> "The interaction effect was significant, $F(1, 210) = 4.03$, $p = 0.046$."

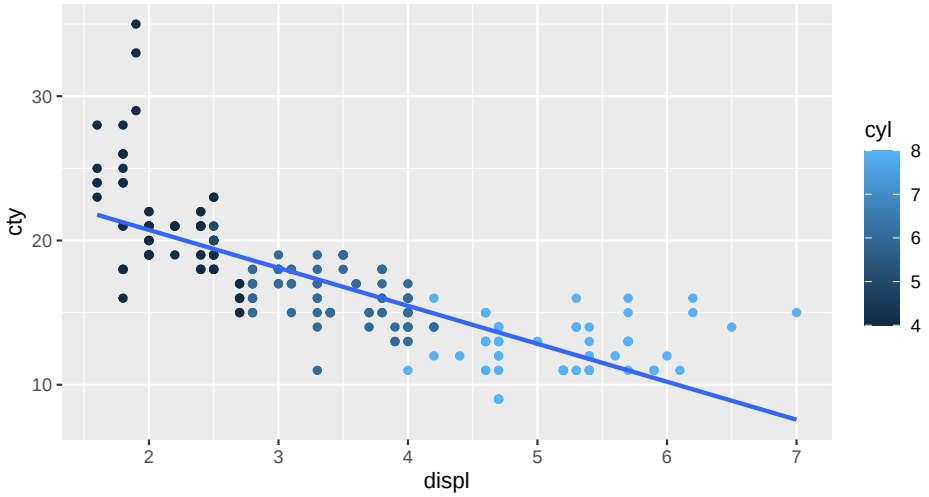
Note that the F-value is simply the square of the t-value, when we take into account rounding errors ($4.04 = 2.01^2$), and that the p-values are exactly the same. Note that the advantage of the linear model is that we have an estimate of the size of the effect, together with a confidence interval.

9.6 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.

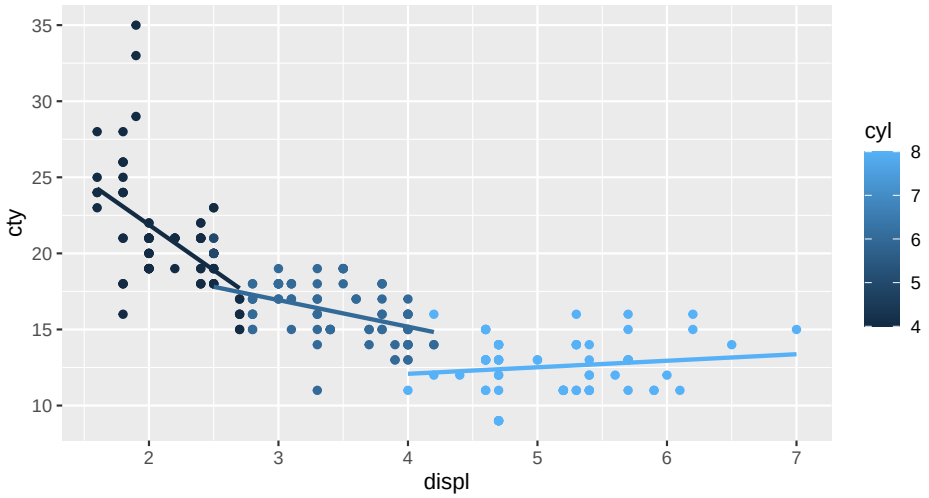
As an example data set, we look at the `mpg` data frame, available in the `ggplot2` package. It contains data on 234 cars. Let's analyse the dependent variable **cty** (city miles per gallon) as a function of the numeric variables **cyl** (number of cylinders) and **displ** (engine displacement). First 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:

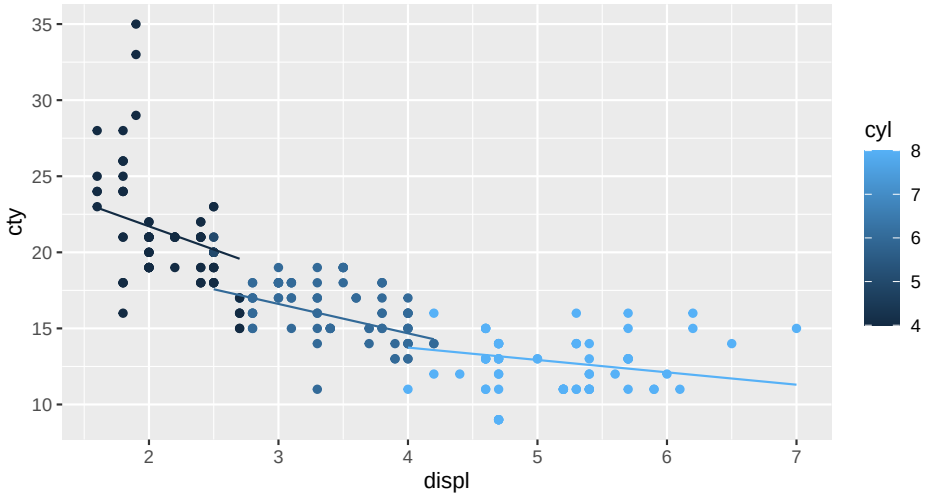
$$\begin{aligned}
\widehat{\texttt{cty}} &= 0.6 - 5.285 \times \texttt{displ} - 2.704\texttt{cyl} + 0.559 \times \texttt{displ} \times \texttt{cyl} \\
\widehat{\texttt{cty}} &= 0.6 - 5.285 \times \texttt{displ} - 2.704 \times 2 + 0.559 \times \texttt{displ} \times 2 \\
\widehat{\texttt{cty}} &= 0.6 - 5.285 \times \texttt{displ} - 5.408 + 1.118 \times \texttt{displ} \\
\widehat{\texttt{cty}} &= (0.6 - 5.408) + (1.118 - 5.285) \times \texttt{displ} \\
\widehat{\texttt{cty}} &= -4.808 - 4.167 \times \texttt{displ}
\end{aligned}$$

If we increase the number of cylinders from 2 to 3, we obtain the equation:

$$\begin{aligned}
\widehat{\texttt{cty}} &= 0.6 - 5.285 \times \texttt{displ} - 2.704 \times 3 + 0.559 \times \texttt{displ} \times 3 \\
\widehat{\texttt{cty}} &= -7.512 - 3.608 \times \texttt{displ}
\end{aligned}$$

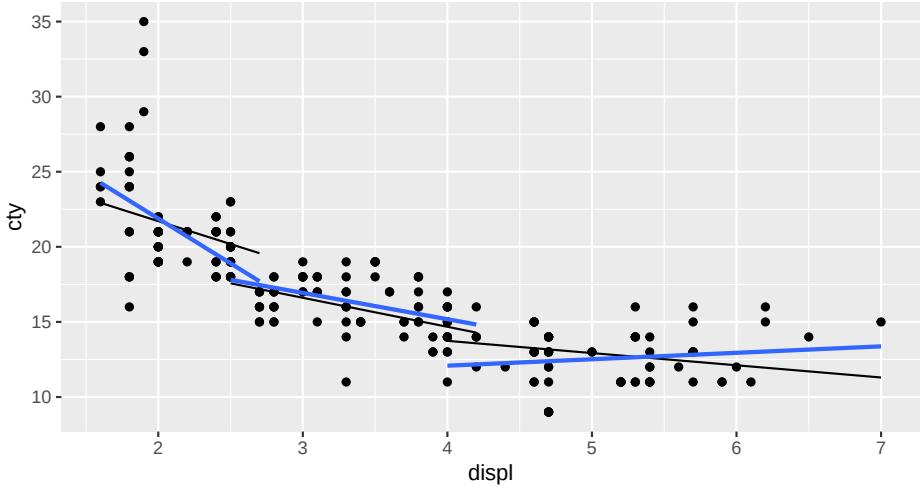
We see a different intercept and a different slope. The difference in the slope between 3 and 2 cylinders equals 0.559, which is exactly the interaction effect. If you do the same exercise with 4 and 5 cylinders, or 6 and 7 cylinders, you will always see this difference again. This parameter for the 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 simpler 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.

9.7 Take-away points

- When having two independent variables, it is possible to also quantify the extent to which one variable *moderates* (modifies) the effect of the other variable on the dependent variable.
- This quantity, the extent to which one variable *moderates* (modifies) the effect of the other variable on the dependent variable, is termed *interaction effect*.

Key concepts

- Main effect
- Interaction
- Moderation
- Means and errors plot

Chapter 10

Contrasts

10.1 Introduction

In Chapter 6 where we discussed ANOVA, we saw that we can make comparisons between means of different groups. Suppose we want to compare the mean height in three countries A, B and C. When we run a linear model, we see that usually the first group (country A) becomes the reference group. In the output then, the intercept is equal to the mean in this reference group, and the two slope parameters are the differences between countries B and A, and the difference between countries C and A, respectively. This choice of what comparisons are made, is the default choice. You may well have the desire to make other comparisons. Suppose you want to regard country C as the reference category, or perhaps you would like to make a comparison between the means of countries B and C? This chapter focuses on how to make choices regarding what comparisons you would like to make. We also explain how to do to it in R and at the same time tell you a bit more about how linear models work in the first place.

10.2 The idea of a contrast

This chapter deals with *contrasts*. We start out by explaining what we mean with contrasts and what role they play in linear models with categorical variables. In later sections, we discuss how specifying contrasts can help us to make the standard `lm()` output more relevant and easier to interpret.

A contrast is a *linear combination* of parameters or statistics. Another word for a linear combination is a *weighted sum*. A regression equation like $b_1X_1 + b_2X_2$ is also a linear combination: a linear combination of independent variables X_1 and X_2, with *weights* b_1 and b_2 (that's why we are talking about *linear* models).

Instead of variables, let's focus now on two sample statistics, the mean bloodpressure M_{Dutch} in a random sample of two Dutch persons and the mean bloodpressure M_{German} in a random sample of two German persons. We can take the sum of these two means and call it $L1$.

$$L1 = M_{German} + M_{Dutch}$$

Note that this is equivalent to defining $L1$ as

$$L1 = 1 \times M_{German} + 1 \times M_{Dutch}$$

This is a weighted sum, where the weights for the two statistics are 1 and 1 respectively. Alternatively, you could use other weights, for example 1 and -1. Let's call such a contrast $L2$:

$$L2 = 1 \times M_{German} - 1 \times M_{Dutch}$$

This $L2$ could be simplified to

$$L2 = M_{German} - M_{Dutch}$$

which shows that $L2$ is then equal to the difference between the German mean and the Dutch mean. If $L2$ is positive, it means that the German mean is higher than the Dutch mean. When $L2$ is negative, it means that the Dutch mean is higher. We can also say that $L2$ *contrasts* the German mean with the Dutch mean.

Example

If we define $L2 = M_{German} - M_{Dutch}$ and we find that $L2$ equals +2.13, it means that the mean of the Germans is 2.13 units higher than the mean of the Dutch. If we find a 95% confidence interval for $L2$ of <1.99, 2.27>, this means that the best guess is that the German population mean is between 1.99 and 2.27 units higher than the Dutch population mean.

If we would fix the order of the two means M_{German} and M_{Dutch}, we can make a nice short summary for a contrast. Suppose we order the groups alphabetically: first the Dutch, then the Germans. That means that we also fix the order of the weights: the first weight belongs to the group that comes first alphabetically, and the second weight belongs to the group that comes second alphabetically. Then we could summarise $L1$ by stating only the weights: 1 and 1. A more common way to display these values is using a *row vector*:

$$\mathbf{L1} = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

representing $1 \times M_{Dutch} + 1 \times M_{German}$. Contrast $L2$ could be summarised as -1 and 1, to represent $-1 \times M_{Dutch} + 1 \times M_{German}$. Written as a row vector we get

$$\mathbf{L2} = \begin{bmatrix} -1 & 1 \end{bmatrix}$$

We can combine these two row vectors by pasting them on top of each other and display them as a *contrast matrix* $\mathbf{L}$ with two rows and two columns:

$$\mathbf{L} = \begin{bmatrix} \mathbf{L1} \\ \mathbf{L2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

In summary, a contrast is a weighted sum of group means, and can be summarised by the weights for the respective group means. Several contrasts, represented as row vectors, can be combined into a contrast matrix. Below we discuss why this simple idea of contrasts is of interest when discussing linear models. But first, let's have a quick overview of what we learned in previous chapters.

10.3 A quick recap

Let's briefly review what we learned earlier about linear models when dealing with categorical variables. In Chapter 4 we were introduced to the simplest linear model: a numeric dependent variable Y and one numeric independent variable X. In Chapter 6 we saw that we can also use this framework when we want to use a categorical variable X. Suppose that X has two categories, say "Dutch" and "German", to indicate the nationality of participants in a study on bloodpressure. We saw that we could create a numeric variable with only 1s and 0s that conveys the same information about nationality. Such a numeric variable with only 1s and 0s is usually called a dummy variable. Let's use a dummy variable to code for nationality and call that variable **German**. All participants that are German are coded as 1 for this variable and all other participants are coded as 0. Table 10.1 shows a small imaginary data example on diastolic bloodpressure.

We learned that we can plug such a dummy variable as a numeric predictor variable into a linear regression model, here with **bp_diastolic** (diastolic bloodpressure) as the dependent variable. The data and the OLS regression line are visualised in Figure 10.1.

We learned that the slope in the model with such a dummy variable can be interpreted as the difference in the means for the two groups. In this case, the slope is equal to the difference in the means of the two groups in the sample data. The mean bloodpressure in the Dutch is $\frac{92+92}{2} = 92$ and in the Germans

Table 10.1: Small imaginary data example on bloodpressure with one dummy variable German to code for a categorical variable nationality.

participant	nationality	German	bp_diastolic
1	Dutch	0	92
2	German	1	97
3	German	1	89
4	Dutch	0	92

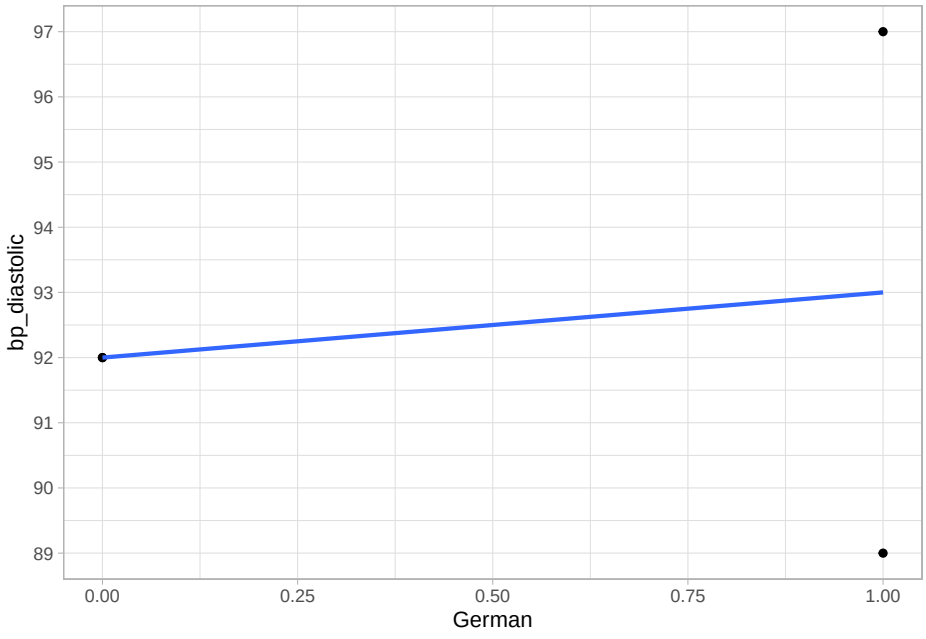


Figure 10.1: Linear regression of diastolic bloodpressure on the dummy variable German.

Table 10.2: Regression table for diastolic bloodpressure on the dummy variable German.

term	estimate	std.error	statistic	p.value
(Intercept)	92	2.828	32.527	0.0009
German	1	4.000	0.250	0.8259

Table 10.3: Small data example on bloodpressure with categorical variable nationality with three levels.

participant	nationality	bp_diastolic
1	Dutch	92
2	German	97
3	German	89
4	Dutch	92
5	Italian	91
6	Italian	96

it is $\frac{89+97}{2} = 93$. The *increase* in mean bloodpressure if we move from German = 0 (Dutch) to German = 1 (German) is therefore equal to +1.

This is also what we see if R computes the regression for us, see Table 10.2.

We see that being German has an effect of +1 units on the bloodpressure scale, with the individuals that are not German (Dutch) as the reference group. The intercept has the value 92 and that is the average bloodpressure in the non-German group. We can therefore view this +1 as the difference between mean German bloodpressure and mean Dutch bloodpressure.

Now suppose we do not have two groups, but three. We learned in Chapter 6 that we can have a hypothesis about the means of several groups. Suppose we have data from three groups, say Dutch, German and Italian. The data are in Table 10.3.

When we run an ANOVA, the null hypothesis states that the means of the three groups are equal in the population. For example, the mean diastolic bloodpressure might be 92 for the Dutch, 93 for the Germans and 93.5 for the Italians, but this could result when the actual means in the population are the same: $\mu_{Dutch} = \mu_{German} = \mu_{Italian}$.

If we would run an ANOVA we could get the following results

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
out <- bloodpressure %>%
  lm(bp_diastolic ~ nationality, data = ., contrasts = list(nationality = contr.sum))
out %>%
  Anova(type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: bp_diastolic
##              Sum Sq Df   F value   Pr(>F)
## (Intercept)  51708  1 3485.9438 1.07e-05 ***
## nationality      2  2    0.0787   0.9262
## Residuals       44  3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see an F-test with 2 and 3 degrees of freedom. The 2 results from the fact that we have three groups. This test is about the equality of the three group means. Compare this with the effects of the automatically created dummy variables in the `lm()` analysis below:

```
out <- bloodpressure %>%
  lm(bp_diastolic ~ nationality, data = .)
out %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term               estimate std.error statistic   p.value
##   <chr>                 <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)           92.0       2.72    33.8   0.0000570
## 2 nationalityGerman      1.00      3.85     0.260 0.812
## 3 nationalityItalian     1.50      3.85     0.389 0.723
```

In the output, the first t-test (33.8) is about the equality of the mean bloodpressure in the first group and 0. The second t-test (0.26) is about the equality of the means in groups 1 and 2, and the third t-test (0.389) is about the equality of the means in groups 1 and 3. The F-test from the ANOVA can be seen as a combination of two separate t-tests: the second and the third in this example, which are the ones about the equality of the last two group means and the first one. These two effects are from the two dummy variables created to code for the categorical variable **nationality**.

In the research you carry out, it is important to be clear on what you actually want to learn from the data. If you want to know if the data could have resulted from the situation where all population group means are equal, then the F-test is most appropriate. If you are more interested in whether the differences between certain countries and the first country (the reference country) are 0 in the population, the regular `lm()` table with the standard t-tests are more appropriate. If you are interested in other types of questions, then keep reading.

10.4 Contrasts and dummy coding

In the previous section we saw that when we run a linear model with a categorical predictor with two categories, we actually run a linear regression with a dummy variable. When the dummy variable codes for the category German, we see that the slope is the same as the mean of the German category minus the mean of the reference group. And we see that the intercept is the same as the mean of the reference group (the non-German group). We see that the dummy variable is 1 for Germans and 0 for Dutch, and that the use of this dummy variable in the analysis results in two parameters: (1) the intercept and (2) the slope. Actually these two parameters represent two *contrasts*. The first parameter contrasts the difference between the Dutch mean and 0 (again, the Dutch come first alphabetically and then the weight for the Germans):

$$L1 = M_{Dutch} = M_{Dutch} - 0 = 1 \times M_{Dutch} - 0 \times M_{German}$$

In the R output, this parameter (or contrast) is called '(Intercept)'. The second parameter contrasts the German mean with the Dutch mean:

$$L2 = M_{German} - M_{Dutch} = -1 \times M_{Dutch} + 1 \times M_{German}$$

This parameter is often denoted as the slope. In Table 10.2 above it is denoted as 'German'.

These two contrasts can be represented in a contrast matrix containing two rows

$$\mathbf{L} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

Suppose we want to have some different output. Suppose instead we want to see in the output the mean of the Germans first and then the extra bloodpressure in the Dutch. Earlier we learned that we can get that by using a dummy variable for being Dutch, so that the German group becomes the reference group. Here, we focus on the contrasts. Suppose we want the Germans to form the reference group, then we have to estimate the contrast:

$$L3 : M_{German} = 0 \times M_{Dutch} + 1 \times M_{German}$$

and then we can contrast the Dutch with this reference group:

$$L4 : M_{Dutch} - M_{German} = 1 \times M_{Dutch} - 1 \times M_{German}$$

We then have the following contrast matrix:

$$\mathbf{L} = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}$$

You see that when you make a choice regarding the dummy variable (either coding for being German or coding for being Dutch), this choice directly affects the contrasts that you make in the regression analysis (i.e., the output that you get). When using dummy variable **German**, the output gives contrasts $L1$ and $L2$, whereas using dummy variable **Dutch** leads to an output for contrasts $L3$ and $L4$.

To make this a bit more interesting, let's look at a data example with three nationalities: Dutch, German and Italian. Suppose we want to use the German group as the reference group. We can then use the following contrast (the intercept):

$$L1 : M_{German}$$

Next, we are interested in the difference between the Dutch and this reference group:

$$L2 : M_{Dutch} - M_{German}$$

and the difference between the Italian and this reference group:

$$L3 : M_{Italian} - M_{German}$$

If we again order the three categories alphabetically, we can summarise these three contrasts with the contrast matrix

Table 10.4: Small data example with categorical variable nationality with three levels, and two dummy variables.

participant	nationality	Dutch	Italian	bp_diastolic
1	Dutch	1	0	92
2	German	0	0	97
3	German	0	0	89
4	Dutch	1	0	92
5	Italian	0	1	91
6	Italian	0	1	96

$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

(please check this for yourself).

Based on what we learned from previous chapters, we know we can obtain these comparisons when we compute a dummy variable for being Dutch and a dummy variable for being Italian and use these in a linear regression. We see these dummy variables **Dutch** and **Italian** in Table 10.4.

This section showed that if you change the dummy coding, you change the contrasts that you are computing in the analysis. There exists a close connection between the contrasts and the actual dummy variables that are used in an analysis. We will discuss that in the next section.

10.5 Connection between contrast and coding schemes

We saw that what dummy coding you use determines the contrasts you get in the output of a linear model. In this section we discuss this intricate connection. As we saw earlier, the contrasts can be represented as a matrix $\mathbf{L}$ with each row representing a contrast. Now let's focus on the dummy coding. The way we code dummy variables, can be represented in a matrix $\mathbf{S}$ where each column gives information about how to code a new numeric variable.

For a simple example, see the following matrix

$$\mathbf{S} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Each row in this matrix represents a category: the first row is for the first category and the second row is for the second category. The columns represent the coding scheme for new variables. The first column specifies the coding scheme for the first new variable. The values are the values that should be used in constructing the new variables. The first column of **S** specifies a dummy variable where the first category is coded as 1 and the second category as 0. The second column specifies the opposite: a dummy variable that codes the first category as 0 and the second category as 1.

If you want the short story for how (dummy) coding and contrasts are related: matrix **S** (the coding scheme) is the *inverse* of the contrast matrix **L**. That is, if you have matrix **L** and you want to know the coding scheme **S**, you can simply ask R to compute the inverse of **L**. This also goes the other way around: if you know what coding is used, you can take the inverse of the coding scheme to determine what the output represents. If you want to know what is meant with the inverse, see the advanced clickable section below, although it is not strictly necessary to know it. In order to work with linear models in R, it is sufficient to know how to compute the inverse in R.

A closer look at the L and S matrices and their connection

By default, R orders the categories of a factor alphabetically. If we have two groups, R uses by default the following *coding scheme*:

$$\mathbf{S} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

This matrix has two rows and two columns. Each *row* represents a category. In our earlier example, the first category is "Dutch" and the second category is "German" (categories sorted alphabetically). Each *column* of the matrix **S** represents a new numeric variable that R will compute in order to code for the categorical factor variable. The first column (representing a new numeric variable) says that for both categories (rows), the new variable gets the value 1. The second column holds a 0 and a 1. It says that the first category is coded as 0, and the second category is coded as 1.

Thus, this matrix says that we should create two new variables. They are displayed in Table 10.5, where we call them **Intercept** and **German**, respectively. Why we name the second new variable **German** is obvious: it's simply a dummy variable for coding "German" as 1. Why we call the first variable **Intercept** is less obvious, but we will come back to that later.

Thus, with matrix **S**, you have all the information you need to perform a linear regression analysis with only numerical variables.

With matrix **S** as the default, what kind of contrasts will we get, and how can we find out? It turns out that once you know **S**, you can immediately calculate

(or have R calculate) what kind of contrast matrix $\mathbf{L}$ you get. The connection between matrices $\mathbf{L}$ and $\mathbf{S}$ is the same as the connection between 4 and $\frac{1}{4}$.

You might know that $\frac{1}{4}$ is called the *reciprocal* of 4. The reciprocal of 4 can also be denoted by 4^{-1}. Thus we can write that the reciprocal of 4 is as follows:

$$4^{-1} = \frac{1}{4}$$

We can say that if we have a quantity x, the reciprocal of x is defined as that number when multiplied with x, results in 1.

$$xx^{-1} = x^{-1}x = 1$$

Examples

- The reciprocal of 3 is equal to $\frac{1}{3}$, because $3 \times \frac{1}{3} = 1$.
- The reciprocal of $\frac{1}{100}$ is equal to 100, because $\frac{1}{100} \times 100 = 1$.
- The reciprocal of $\frac{3}{4}$ is equal to $\frac{4}{3}$, because $\frac{3}{4} \times \frac{4}{3} = 1$.

Just like a number has a reciprocal, a matrix has an *inverse*. Similar to reciprocals, we use the "-1" to indicate the inverse of a matrix. We have that by definition, $\mathbf{L}\mathbf{L}^{-1} = \mathbf{I}$ and $\mathbf{S}\mathbf{S}^{-1} = \mathbf{I}$, where matrix $\mathbf{I}$ is the matrix analogue of a 1, and its called an *identity matrix*. It turns out that the inverse of matrix $\mathbf{L}$ is $\mathbf{S}$: $\mathbf{L}^{-1} = \mathbf{S}$, and the inverse of $\mathbf{S}$ equals $\mathbf{L}$: $\mathbf{S}^{-1} = \mathbf{L}$. That implies we have that $\mathbf{L}\mathbf{S} = \mathbf{I}$. The only difference with reciprocals is that $\mathbf{L}$ and $\mathbf{S}$ are matrices, so that we are in fact performing *matrix algebra*, which is a little bit different from algebra with single numbers. Nevertheless, it is true that if we matrix multiply $\mathbf{L}$ and $\mathbf{S}$, we get a matrix $\mathbf{I}$. This matrix is as we said an *identity matrix*, which means it has only 1s on the diagonal (from top-left to bottom-right) and 0s elsewhere:

$$\mathbf{L}\mathbf{S} = \mathbf{S}\mathbf{L} = \mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

It turns out that if we have $\mathbf{S}$ as defined above (the default coding scheme), the contrast matrix $\mathbf{L}$ can only be

$$\mathbf{L} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

or written in full:

$$\mathbf{S}\mathbf{L} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Table 10.5: Small data example with a variable Intercept consisting of only 1s and a dummy variable German. This is the default way to transform a categorical variable nationality into a numeric one.

participant	nationality	Intercept	German	bp_diastolic
1	Dutch	1	0	92
2	German	1	1	97
3	German	1	1	89
4	Dutch	1	0	92

Please note that you don't have to know matrix algebra when studying this book, but it helps to understand what is going on when performing linear regression analysis with categorical variables. By having R do the matrix algebra for you, you can fully control what kind of output you want from an analysis.

10.6 Working with matrices S and L in R

In the last section we saw that coding scheme matrix **S** is the inverse of contrast matrix **L**, and contrast matrix **L** is the inverse of coding scheme matrix **S**. In this section we see how to compute the inverse of a matrix R.

Let's take the bloodpressure example again, where we had data on three nationalities: in alphabetical order the Dutch data, then the German data, and lastly the Italian data.

Let's assume we want to use the German group as the reference group. We then need a matrix **L** similar to the one from a previous section with the nationalities Dutch, German and Italian (in alphabetical order), with German as the reference category (the second group). We enter that contrast matrix **L** in R as follows:

```
L <- matrix(c(0,  1, 0,
              1, -1, 0,
              0, -1, 1), byrow = TRUE, nrow = 3)
L
```

```
##      [,1] [,2] [,3]
## [1,]    0    1    0
## [2,]    1   -1    0
## [3,]    0   -1    1
```

The first row is the contrast for the mean of the second group (German). The second row contrasts the first group (Dutch) with the second group (German). The third row contrasts the third group (Italian) with the second group (German).

If we are interested in the coding scheme matrix, we take the inverse of **L** to get matrix **S** by using the `ginv()` function, that is available in the package `MASS`.

```
library(MASS)
S <- ginv(L)  # S is calculated as the inverse of L
S %>% fractions() # to get more readable output
```

```
##      [,1] [,2] [,3]
## [1,] 1    1    0
## [2,] 1    0    0
## [3,] 1    0    1
```

The output is a coding scheme matrix with 3 *columns*:

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

This coding scheme matrix **S** tells us what kind of variables to compute to obtain these contrasts in **L** in the output of the linear model. Each column of **S** contains the information for each new variable. Since we have three columns, we know that we need to compute three variables. The first column tells us that the first variable consists of the value 1 for all groups. We will come back to this variable of 1s later (it represents the intercept). The second column indicates we need a dummy variable that codes the first group (Dutch) as 1 and the other two groups as 0. The third column tells us we need a third variable that is also a dummy variable, coding 1 for group 3 (Italian) and 0 for the other two groups.

If we apply this coding scheme to our data set, then we get the following data matrix:

Now let's do the exercise the other way around. If you know the coding scheme (how your dummy variables are coded), you can easily determine what the output will represent. For instance, let **S** be the coding scheme for a dummy variable for group 1 (Dutch) and a dummy variable for group 2 (German) (with reference group "Italian"). We enter that in R:

```
S <- matrix(c(1, 0,
              0, 1,
              0, 0), byrow = TRUE, nrow = 3)
```

Table 10.6: Data frame with the three new variables, as specified in matrix S.

participant	nationality	bp_diastolic	v1	v2	v3
1	Dutch	92	1	1	0
2	German	97	1	0	0
3	German	89	1	0	0
4	Dutch	92	1	1	0
5	Italian	91	1	0	1
6	Italian	96	1	0	1

By default, any `lm()` analysis includes an intercept, and an intercept is always coded as 1. We therefore add a variable of only 1s to represent the intercept (why these 1s are included will become clear).

```
S <- cbind(1, S)
S
```

```
##      [,1] [,2] [,3]
## [1,]    1    1    0
## [2,]    1    0    1
## [3,]    1    0    0
```

We can then determine what contrasts these three variables will lead to by taking the inverse to calculate contrast matrix **L**:

```
L <- ginv(S)
L %>% fractions() # for readability
```

```
##      [,1] [,2] [,3]
## [1,]  0    0    1
## [2,]  1    0   -1
## [3,]  0    1   -1
```

This matrix **L** contains three rows, one for each contrast. The first row gives a contrast $L1$ which is $0 \times M_1 + 0 \times M_2 + 1 \times M_3$ which is equal to M_3 (the Italian mean). The second row gives contrast $1 \times M_1 + 0 \times M_2 - 1 \times M_3$ which amounts to the first group mean minus the last one, $M_1 - M_3$ (difference between Dutch and Italian). The third row gives the contrast $0 \times M_1 + 1 \times M_2 - 1 \times M_3$ which is $M_2 - M_3$ (difference between German and Italian).

Remember, contrast matrix **L** is organised in *rows*, whereas coding scheme matrix **S** is organised in *columns*. One way to memorise this is to realise that

coding scheme matrix **S** defines new variables, and variables are organised as columns in a data matrix, as you may recall from Chapter 1.

You may wonder, what is this variable with only 1s doing in this story? Didn't we learn that we only need to compute 2 dummy variables for a categorical variable with 3 categories? The short answer is: the variable with only 1s (for all categories) represents the intercept, which is always included in a linear model by default. For more explanation, click on the link below.

A closer look at the intercept represented as a variable with 1s

In most regression analyses, we want to include an intercept in the linear model. This happens so often that we forget that we have a choice here. For instance, in R, an intercept is included by default. For example, if you use the code with the formula `bp_diastolic ~ nationality`, you get the exact same result as with `bp_diastolic ~ 1 + nationality`. In other words, by default, R computes a variable that consists of only 1s. This can be seen if we use the `model.matrix()` function. This function shows the actual variables that are computed by R and used in the analysis.

```
lm(bp_diastolic ~ nationality, data = bloodpressure) %>%
  model.matrix()
```

```
##   (Intercept) nationalityGerman nationalityItalian
## 1           1                 0                  0
## 2           1                 1                  0
## 3           1                 1                  0
## 4           1                 0                  0
## 5           1                 0                  1
## 6           1                 0                  1
## attr(,"assign")
## [1] 0 1 1
## attr(,"contrasts")
## attr(,"contrasts")$nationality
## [1] "contr.treatment"
```

In the output, we see the actual numeric variables that are used in the analysis for the six observations in the data set (six rows). Taken together, these variables, displayed as columns in a matrix, are called the *design matrix*. We see a variable called **(Intercept)** with only 1s and two dummy variables, one coding for Germans, with the name **nationalityGerman**, and one coding for Italians with the name **nationalityItalian**. These variable names we see again when we look at the results of the `lm()` analysis:

```
lm(bp_diastolic ~ nationality, data = bloodpressure) %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term              estimate std.error statistic   p.value
##   <chr>                <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)          92.0       2.72    33.8   0.0000570
## 2 nationalityGerman     1.00      3.85     0.260 0.812
## 3 nationalityItalian    1.50      3.85     0.389 0.723
```

Now the names of the newly computed *variables* have become the names of *parameters* (the intercept and slope parameters). These parameter values actually represent the (default) *contrasts* $L1$, $L2$ and $L3$, here quantified as 92.0, 1.0 and 1.5, respectively. The 'intercept' of 92.0 is simply the quantity that we find for the first contrast (the first row in the contrast matrix $\mathbf{L}$).

In summary, when you have three groups, you can invent a quantitative variable called (intercept) that is equal to 1 for all observations in your data matrix. If you then also invent a dummy variable for being German and a dummy variable for being Italian, and submit these three variables to a linear model analysis, the output will yield the mean of the Dutch, the difference between the German mean and the Dutch mean, and the difference between the Italian mean and the Dutch mean. In the code below, you see what is actually going on: the computation of the three variables and submitting these three variables to an `lm()` analysis. Note that R by default includes an intercept. To suppress this behaviour, we include `-1` in the formula:

```
dummy_German   <- c(0, 1, 1, 0, 0, 0)
dummy_Italian  <- c(0, 0, 0, 0, 1, 1)
iNtErCePt      <- c(1, 1, 1, 1, 1, 1)
bp_diastolic   <- c(92, 97, 89, 92, 91, 96)

lm(bp_diastolic ~ iNtErCePt + dummy_German + dummy_Italian - 1) %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term          estimate std.error statistic   p.value
##   <chr>            <dbl>     <dbl>     <dbl>     <dbl>
## 1 iNtErCePt        92.0       2.72    33.8   0.0000570
## 2 dummy_German      1.00      3.85     0.260 0.812
## 3 dummy_Italian     1.50      3.85     0.389 0.723
```

You see that you get the exact same results as in the default way, by specifying a categorical variable **nationality**:

```
nationality <- factor(c(1, 2, 2, 1, 3,3 ),
                      labels = c("Dutch", "German", "Italian"))

lm(bp_diastolic ~ nationality) %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term               estimate std.error statistic   p.value
##   <chr>                 <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)           92.0       2.72    33.8   0.0000570
## 2 nationalityGerman      1.00      3.85     0.260 0.812
## 3 nationalityItalian     1.50      3.85     0.389 0.723
```

Or when doing the dummy coding yourself:

```
lm(bp_diastolic ~ dummy_German + dummy_Italian) %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term          estimate std.error statistic   p.value
##   <chr>            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)      92.0       2.72    33.8   0.0000570
## 2 dummy_German      1.00      3.85     0.260 0.812
## 3 dummy_Italian     1.50      3.85     0.389 0.723
```

and in a way where we explicitly include an intercept of 1s:

```
lm(bp_diastolic ~ 1 + nationality) %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term               estimate std.error statistic   p.value
##   <chr>                 <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)           92.0       2.72    33.8   0.0000570
## 2 nationalityGerman      1.00      3.85     0.260 0.812
## 3 nationalityItalian     1.50      3.85     0.389 0.723
```

```
lm(bp_diastolic ~ 1 + dummy_German + dummy_Italian) %>%
  tidy()
```

```
## # A tibble: 3 x 5
```

```
##   term          estimate std.error statistic   p.value
##   <chr>            <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)      92.0       2.72    33.8   0.0000570
## 2 dummy_German      1.00      3.85     0.260 0.812
## 3 dummy_Italian     1.50      3.85     0.389 0.723
```

with the only difference that the first parameter is now called '(Intercept)'.

In conclusion: linear models by default include an 'intercept' that is actually coded as an independent variable consisting of all 1s.

10.7 Choosing the reference group in R for dummy coding

Let's see how to run an analysis in R, when we want to control what contrasts are actually computed. This time, we want to let the Germans form the reference group, and then compare the Dutch and Italian means with the German mean.

Let's first put the data from Table 10.3 in R, so that you can follow along using your own computer (copy the code below into an R-script).

```
# create a dataframe (a tibble is the tidyverse version of a dataframe)
bloodpressure <- tibble(
  participant = c(1, 2, 3, 4, 5, 6),
  nationality = c("Dutch", "German", "German", "Dutch", "Italian", "Italian"),
  bp_diastolic = c(92, 97, 89, 92, 91, 96))
bloodpressure
```

```
## # A tibble: 6 x 3
##   participant nationality bp_diastolic
##         <dbl> <chr>              <dbl>
## 1           1 Dutch                 92
## 2           2 German                97
## 3           3 German                89
## 4           4 Dutch                 92
## 5           5 Italian               91
## 6           6 Italian               96
```

Note that we omit dummy variables for now. Next we turn the **nationality** variable in the tibble `bloodpressure` into a factor variable. This ensures that R knows that it is a categorical variable.

```
bloodpressure <-
  bloodpressure %>%
  mutate(nationality = as.factor(nationality))
```

If we then look at this factor,

```
bloodpressure$nationality
```

```
## [1] Dutch   German  German  Dutch   Italian Italian
## Levels: Dutch German Italian
```

we see that it has three levels. With the `levels()` function, we see that the first level is Dutch, the second level is German, and the third level is Italian. This is default behaviour: R chooses the order alphabetically.

```
bloodpressure$nationality %>% levels()
```

```
## [1] "Dutch"   "German"  "Italian"
```

This means that if we run a standard `lm()` analysis, the first level (the Dutch) will form the reference group, and that the two slope parameters stand for the contrasts of the bloodpressure in Germans and Italians versus the Dutch, respectively:

```
bloodpressure %>%
  lm(bp_diastolic ~ nationality, data = .)
```

```
##
## Call:
## lm(formula = bp_diastolic ~ nationality, data = .)
##
## Coefficients:
##        (Intercept)   nationalityGerman  nationalityItalian
##               92.0                 1.0                 1.5
```

In the case that we want to use the Germans as the reference group instead of the Dutch, we have to specify a set of contrasts that is different from this default set of contrasts.

One way to easily do this is to permanently set the reference group to another level of the **nationality** variable. We work with the variable **nationality** in the dataframe (or *tibble*) called `bloodpressure`. We use the `relevel()` function to change the reference group to "German".

```
bloodpressure <- bloodpressure %>%
  mutate(nationality = relevel(nationality, ref = "German"))
```

If we check this by using the function `levels()` again, we see that the first group is now German, and no longer Dutch:

```
bloodpressure$nationality %>% levels()
```

```
## [1] "German"  "Dutch"   "Italian"
```

If we run an `lm()` analysis, we see from the output that the reference group is now indeed the German group (i.e., with contrasts for being Dutch and Italian).

```
bloodpressure %>% lm(bp_diastolic ~ nationality, data = .)
```

```
##
## Call:
## lm(formula = bp_diastolic ~ nationality, data = .)
##
## Coefficients:
##        (Intercept)    nationalityDutch  nationalityItalian
##               93.0                -1.0                 0.5
```

To check this fully, we can ask R to show us the dummy variables that it created for the analysis, using the function `model.matrix()`.

```
bloodpressure %>%
  lm(bp_diastolic ~ nationality, data = .) %>%
  model.matrix()
```

```
##   (Intercept) nationalityDutch nationalityItalian
## 1           1                1                  0
## 2           1                0                  0
## 3           1                0                  0
## 4           1                1                  0
## 5           1                0                  1
## 6           1                0                  1
## attr(,"assign")
## [1] 0 1 1
## attr(,"contrasts")
## attr(,"contrasts")$nationality
## [1] "contr.treatment"
```

The output gives the values for the new variables for all individuals in the original dataset. We see that R created a dummy variable for being Dutch (named **nationalityDutch**), and another dummy variable for being Italian (named **nationalityItalian**). Therefore, the reference group is now formed by the Germans. Note that R also created a variable called **(Intercept)**, that consists of only 1s for all six individuals.

The variable **nationality** in the dataframe `bloodpressure` is now permanently altered. From now on, every time we use this variable in an analysis, the reference group will be formed by the Germans. By using `mutate()` and `relevel()` again, we can change everything back or pick another reference group.

This fixing of the reference category is very helpful in data analysis for experimental data. Suppose a researcher wants to quantify the effects of vaccine A and vaccine B on hospitalisation for COVID-19. She randomly assigns individuals to one of three groups: A, B and C, where group C is the control condition in which individuals receive a saline injection (placebo). Individuals in group A and group B get vaccines A and B, respectively. In order to quantify the effect of vaccine A, you would like to contrast group A with group C, and in order to quantify the effect of vaccine B, you would like to contrast group B with group C. In that case, it is helpful if the reference group is group C. By default, R uses the first category as the reference category, when ordered alphabetically. By default therefore, R would choose group A as the reference group. R would then by default compute the contrast between B and A, and between C and A, apart from the mean of group A (the intercept). By changing the reference category to group C, the output would be more helpful to answer the research question.

Summary

General steps to take for using dummy coding and choosing the reference category:

1. Use `levels()` to check what group is named first. That group will generally be the reference group. For example, do `levels(data$variable)`.
2. If the reference group is not the right one, use `mutate()` and `relevel()` to change the reference group, for example do something like: `data <- data %>% mutate(variable = relevel(variable, ref = "German"))`.
3. Verify with `levels()` whether the right reference category is now mentioned first. Any new analysis using `lm()` will now yield an analysis with dummy variables for all categories except the first one.

10.8 Alternative coding schemes

By default, R uses dummy coding. Contrasts that are computed by using dummy variables are called *treatment contrasts*: these contrasts are the difference between the individual groups and the reference group. Similar to defining the reference group for a particular variable (so that the first parameter value in the output is the mean of this reference group, usually termed **(Intercept)**), you can set all contrasts that are used in the linear model analysis. For instance by typing

```
bloodpressure$nationality %>% contrasts()
```

```
##         Dutch Italian
## German      0       0
## Dutch       1       0
## Italian     0       1
```

you see that the first variable that is created (the first column) is a dummy variable for being Dutch, and the second variable is a dummy for being Italian. What you see in the `contrasts()` matrix is not the contrasts *per se*, but the coding scheme for the new variables that are created by R. In other words, with `contrasts()` you get to see the coding scheme matrix **S**. Note however that the default variable with 1s (the intercept) is not displayed.

This default treatment coding (dummy coding) is fine for linear regression, but not so great for ANOVAs. We will come back to this issue in a later section when discussing moderation. For now it suffices to know that if you are interested in an ANOVA rather than a linear regression analysis, it is best to use *effects coding* (sometimes called *sum-to-zero contrasts*). Above we saw that the first column after the `contrasts()` function (called 'Dutch') sums to 1, and the second column (called 'Italian') also sums to 1. In effects coding on the other hand, the weights for each contrast sum to 0 (hence the name *sum-to-zero contrasts*). Effects coding is also useful if we have specific research questions where we are interested in differences between *combinations* of groups. We will see in a moment what that means. Below we discuss effects coding in more detail, and also discuss other standard ways of coding (*Helmert contrasts* and *successive differences contrast coding*).

10.8.1 Effects coding

Below we will illustrate how to use and interpret a linear model when we apply effects coding. You can type along in R.

First, let's check that we have everything in the original alphabetical order. This makes it easier to not get distracted by details.

```
bloodpressure$nationality %>% levels() # German is now the first group
```

```
## [1] "German"  "Dutch"   "Italian"
```

```
# change the order such that "Dutch" becomes the first group
bloodpressure <- bloodpressure %>%
  mutate(nationality = relevel(nationality, ref = "Dutch"))
bloodpressure$nationality %>%
  levels() # Check that Dutch is now the first group
```

```
## [1] "Dutch"   "German"  "Italian"
```

Next, we need to change the default dummy coding into effects coding for the **nationality** variable in the `bloodpressure` dataframe. We do that by again fixing something permanently for this variable. We use the function `code_deviation()` from the `codingMatrices` package.

```
library(codingMatrices) # install if not already done
contrasts(bloodpressure$nationality) <- code_deviation
```

When we run `contrasts()` again, we see a different set of columns, where each column now sums to 0. We no longer see dummy coding.

```
contrasts(bloodpressure$nationality)
```

```
##         MD1 MD2
## Dutch     1   0
## German    0   1
## Italian  -1  -1
```

The values in this coding scheme matrix are used to code for the new numeric variables. The first column says that a variable should be created with 1s for Dutch, 0s for Germans and -1s for Italians. The second column indicates that another variable should be created with 0s for Dutch, 1s for Germans and -1s for Italians. Any new analysis with the variable **nationality** in the dataframe `bloodpressure` will now by default use this sum-to-zero coding scheme. If we run a linear model analysis with this factor variable, we see that we get very different values for the parameters.

```
bloodpressure %>%
  lm(bp_diastolic ~ nationality, data = .) %>%
  tidy()
```

```
## # A tibble: 3 x 5
##   term           estimate std.error statistic   p.value
##   <chr>             <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)      92.8        1.57   59.0    0.0000107
## 2 nationalityMD1   -0.833      2.22   -0.375  0.733
## 3 nationalityMD2    0.167      2.22    0.0750 0.945
```

Let's check that we understand how the new variables are computed.

```
bloodpressure %>%
  lm(bp_diastolic ~ nationality, data = .) %>%
  model.matrix()
```

```
##   (Intercept) nationalityMD1 nationalityMD2
## 1           1              1              0
## 2           1              0              1
## 3           1              0              1
## 4           1              1              0
## 5           1             -1             -1
## 6           1             -1             -1
## attr(,"assign")
## [1] 0 1 1
## attr(,"contrasts")
## attr(,"contrasts")$nationality
##         MD1 MD2
## Dutch     1   0
## German    0   1
## Italian  -1  -1
```

We still have a variable called **(Intercept)** with only 1s, which is used by default, as we saw earlier. There are also two other new variables, one called **nationalityMD1** and the other called **nationalityMD2**, where there are 1s, 0s and -1s.

Based on this coding scheme it is hard to say what the new values in the output represent. But we saw earlier that the output values are simply the values for the contrasts, and the contrasts can be determined by taking the inverse of the coding scheme matrix **S**.

Matrix **S** consists now of the effects coding scheme. **S** is therefore the same as the output of `contrasts(bloodpressure$nationality)`.

```
S <- contrasts(bloodpressure$nationality)
S
```

```
##         MD1 MD2
## Dutch     1   0
## German    0   1
## Italian  -1  -1
```

We see that it has the two new numeric variables. We miss however the variable with 1s that is used by default in the analysis. If we add that we get the full matrix **S** with the three new variables.

```
S <- cbind(1, S)
S
```

```
##           MD1 MD2
## Dutch   1   1   0
## German  1   0   1
## Italian 1  -1  -1
```

Next we take the inverse to compute contrast matrix **L**.

```
L <- ginv(S)
L
```

```
##             [,1]       [,2]       [,3]
## [1,]  0.3333333  0.3333333  0.3333333
## [2,]  0.6666667 -0.3333333 -0.3333333
## [3,] -0.3333333  0.6666667 -0.3333333
```

```
L %>% fractions()
```

```
##      [,1] [,2] [,3]
## [1,]  1/3  1/3  1/3
## [2,]  2/3 -1/3 -1/3
## [3,] -1/3  2/3 -1/3
```

This gives us what we need to know about how to interpret the output. Briefly, the first parameter in the output now represents what is called the *grand mean*. This is the mean of all group means. This is so, because if we put the group means in the right order, and we use the weights from the first row in **L**, we get

$$L1 : \frac{1}{3}M_{Dutch} + \frac{1}{3}M_{German} + \frac{1}{3}M_{Italian}$$

which can be simplified to

$$L1 : \frac{M_{Dutch} + M_{German} + M_{Italian}}{3}$$

which is the mean of the three group means. We generally call the mean of group means the *grand mean*.

For the meaning of the second parameter we look at the second row of the contrast matrix and fill in the numbers for a second contrast:

$$L2 : \frac{2}{3}M_{Dutch} - \frac{1}{3}M_{German} - \frac{1}{3}M_{Italian}$$

$L2$ can be rewritten as

$$L2 : \frac{2}{3} \times (M_{Dutch} - \frac{M_{German} + M_{Italian}}{2})$$

$L2$ is therefore the difference between the Dutch mean and the mean of the other two means, multiplied by $\frac{2}{3}$.

Although it makes sense to interpret this contrast as the difference between the Dutch mean and the other two means, the fraction $\frac{2}{3}$ looks weird here. Let's therefore rewrite contrast $L2$ in a different way:

$$\begin{aligned} L2 : & \frac{2}{3}M_{Dutch} - \frac{1}{3}M_{German} - \frac{1}{3}M_{Italian} \\ & = \frac{2}{3}M_{Dutch} + \frac{1}{3}M_{Dutch} - \frac{1}{3}M_{Dutch} - \frac{1}{3}M_{German} - \frac{1}{3}M_{Italian} \\ & = M_{Dutch} - \frac{M_{Dutch} + M_{German} + M_{Italian}}{3} \end{aligned}$$

In other words, this contrast is about the difference between the Dutch mean and the grand mean.

For the meaning of the third parameter we look at the third row of the contrast matrix and fill in the numbers for a second contrast:

$$L3 : -\frac{1}{3}M_{Dutch} + \frac{2}{3}M_{German} - \frac{1}{3}M_{Italian}$$

which can be written as

$$L3 : \frac{2}{3} \times (M_{German} - \frac{M_{Dutch} + M_{Italian}}{2})$$

and interpreted as the difference between the German mean and the mean of the other two means, multiplied by $\frac{2}{3}$, or, when writing it out like

$$L3 : M_{German} - \frac{M_{Dutch} + M_{German} + M_{Italian}}{3}$$

as the difference between the German mean and the grand mean.

Let's check whether this makes sense. Let's get an overview of the group means in the data, and the coefficients again:

```
bloodpressure %>%
  group_by(nationality) %>%
  summarise(mean = mean(bp_diastolic))
```

```
## # A tibble: 3 x 2
##   nationality  mean
##   <fct>       <dbl>
## 1 Dutch        92
## 2 German       93
## 3 Italian      93.5
```

```
bloodpressure %>%
  lm(bp_diastolic ~ nationality, data = .) %>% coef()
```

```
##    (Intercept) nationalityMD1 nationalityMD2
##     92.8333333     -0.8333333      0.1666667
```

The first parameter '(Intercept)' equals 92.833 and this is indeed equal to the grand mean $\frac{(92+93+93.5)}{3} = 92.833$.

The second parameter 'nationalityMD1' equals -0.833 and this is equal to the difference between the Dutch mean 92 and the German and Italian means combined, $\frac{(93+93.5)}{2} = 93.25$, so $92 - 93.25 = -1.25$. When we multiply this by $\frac{2}{3}$, we get $-1.25 \times \frac{2}{3} = -0.833$.

The third parameter 'nationalityMD2' equals 0.167 and this is indeed equal to the difference between the German mean 93 and the Dutch and Italian means combined, $\frac{(92+93.5)}{2} = 92.75$, so $93 - 92.75 = 0.25$. When multiplied by $\frac{2}{3}$, we get $0.25 \times \frac{2}{3} = 0.167$. This multiplication is explained in more detail below.

It also computes when we define the contrasts in terms of the grand mean. Let's first compute the grand mean (the mean of all three means):

```
bloodpressure %>%
  group_by(nationality) %>%
  summarise(mean = mean(bp_diastolic)) %>% # the group means
  summarise(grandmean = mean(mean)) # compute mean of the group means
```

```
## # A tibble: 1 x 1
##   grandmean
##       <dbl>
## 1      92.8
```

and compare it to the group means:

```
bloodpressure %>%
  group_by(nationality) %>%
  summarise(mean = mean(bp_diastolic))
```

```
## # A tibble: 3 x 2
##   nationality  mean
##   <fct>       <dbl>
## 1 Dutch        92
## 2 German       93
## 3 Italian      93.5
```

We see that $L2$ should be $92-92.8 = -0.8$, and that $L3$ should be $93-92.8 = 0.2$, which is exactly what we find as coefficients in the output (safe rounding).

```
bloodpressure %>%
  lm(bp_diastolic ~ nationality, data = .) %>% coef()
```

```
##    (Intercept) nationalityMD1 nationalityMD2
##     92.8333333     -0.8333333      0.1666667
```

Effects coding is interesting in all cases where you want to test a null-hypothesis about one group versus the mean of the other groups (or versus the grand mean). As we saw above, the second parameter in the output is about the difference between the first group mean and the average of all group means (the grand mean). The t-test that goes along with that parameter informs us about the null-hypothesis that the first group has the same population mean as the average of all the other group means (in other words, that the group mean is the same as the grand mean). This is useful in a situation where you for example want to compare the control/placebo condition with several experimental conditions combined: do the various treatments in general result in a different mean than when doing nothing?

Suppose for example that we do a study on various types of therapy to quit smoking. In one group, people are put on a waiting list, in a second group people receive treatment using nicotine patches and in a third group, people receive treatment using cognitive behavioural therapy.

If your main questions are (1) what is the effect of nicotine patches on the number of cigarettes smoked, and (2) what is the effect of cognitive behavioural therapy on number of cigarettes smoked, you may want to go for the default dummy coding strategy and make sure that the control condition is the first level of your factor. The parameters and the t-tests that are relevant for your research questions are then in the second and third row of your regression table.

However, if your main question is: "Do treatments like nicotine patches and cognitive behavioural therapy help reduce the number of cigarettes smoked, compared to doing nothing?" you might want to opt for a sum-to-zero (effects coding) strategy, where you make sure that the control condition (waiting list) is the first level of your factor. The relevant null-hypothesis test for your research question is then given in the second line of your regression table (contrasting the control group with the two treatments groups, i.e., contrasting the control group with the grand mean).

Summarising, there are multiple ways of carrying out a linear model analysis with categorical variables. The R default way is to use dummy coding, and the second way is effects coding. Unless you are particularly interested in an ANOVA with multiple independent categorical variables, or about combinations of groups, the default way with dummy coding is just fine. Regarding ANOVA, most software packages and functions that perform ANOVA actually use effects coding by default, so it is important to be familiar with it. It must be said however, that ANOVA results for analyses with one independent variable will not be affected by using a different coding scheme. The coding scheme underlying an ANOVA only matters as soon as you have multiple independent variables.

In previous chapters we learned how to read the output from an `lm()` analysis when it is based on dummy coding. In the current subsection we saw how to figure out what the numbers in the output represent in case of effects coding. Now we will look at a number of other alternatives of coding categorical variables, after which we will look into how to make user-specified contrasts, that help you answer your own particular research question.

10.8.2 Helmert contrasts

One other well-known coding scheme is a Helmert coding scheme. We can specify we want Helmert contrasts for the **nationality** variable if we code the following

```
contrasts(bloodpressure$nationality) <- code_helmert
contrasts(bloodpressure$nationality) %>% fractions()
```

```
##         H2   H3
## Dutch   -1/2 -1/3
```

```
## German   1/2 -1/3
## Italian    0  2/3
```

The first column indicates how the first new variable is coded (Dutch as $-\frac{1}{2}$, Germans as $\frac{1}{2}$ and Italians as 0), and the second column indicates a new variable with $-\frac{1}{3}$ for Germans and Dutch, and $\frac{2}{3}$ for Italians. By default another variable is computed as 1s for all groups, representing some kind of intercept.

If we do the same trick to this matrix as before, by adding a column of 1s and then using `ginv()`, we get the following weights for the contrasts that we are quantifying this way

```
S <- cbind(1, contrasts(bloodpressure$nationality))
L <- ginv(S)
L %>% fractions()
```

```
##      [,1] [,2] [,3]
## [1,]  1/3  1/3  1/3
## [2,]   -1    1    0
## [3,] -1/2 -1/2    1
```

Again each row gives the weights for the contrasts that are computed. Similar as with effects coding, the first contrast represents the grand mean (the mean of the group means):

$$L1: \frac{M_{Dutch} + M_{German} + M_{Italian}}{3}$$

The second row represents the difference between German and Dutch means:

$$L2: -1 \times M_{Dutch} + 1 \times M_{German} + 0 \times M_{Italian} = M_{German} - M_{Dutch}$$

and the third row represents the difference between the Italian mean on the one hand, and the Dutch and German means combined.

$$-\frac{1}{2} \times M_{Dutch} - \frac{1}{2} \times M_{German} + 1 \times M_{Italian} = M_{Italian} - \frac{M_{Dutch} + M_{German}}{2}$$

In general, with a Helmert coding scheme (1) the first group is compared to the second group, (2) the first and second group combined is compared to the third group, (3) the first, second and third group combined is compared to the fourth group, and so on. Let's imagine we have a country factor variable with five levels. Let's then indicate that we want to have Helmert contrasts.

```
country <- c("A", "B", "C", "D", "E") %>% as.factor()
contrasts(country) <- code_helmert
contrasts(country)
```

```
##     H2         H3    H4   H5
## A -0.5 -0.3333333 -0.25 -0.2
## B  0.5 -0.3333333 -0.25 -0.2
## C  0.0  0.6666667 -0.25 -0.2
## D  0.0  0.0000000  0.75 -0.2
## E  0.0  0.0000000  0.00  0.8
```

then we do the trick to get the rows with the contrasts weights

```
S <- cbind(1, contrasts(country))
L <- ginv(S)
L %>% fractions() # to make it easier to read
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  1/5  1/5  1/5  1/5  1/5
## [2,]   -1    1    0    0    0
## [3,] -1/2 -1/2    1    0    0
## [4,] -1/3 -1/3 -1/3    1    0
## [5,] -1/4 -1/4 -1/4 -1/4    1
```

Then we see that the first contrast is the general mean. The second row is a comparison between A and B, the third one a comparison between (A+B)/2 and C, the fourth one a comparison between (A+B+C)/3 and D, and the fifth one a comparison between (A+B+C+D)/4 with E. To put it differently, the Helmert contrasts compare each level with the average of the 'preceding' levels. Thus, we can compare categories of a variable with the mean of the preceding categories of the variable.

You can also choose to go the other way around. If we use the function `code_helmert_forward` instead of `code_helmert`

```
contrasts(country) <- code_helmert_forward
S <- cbind(1, contrasts(country))
L <- ginv(S)
L %>% fractions() # to make it easier to read
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  1/5  1/5  1/5  1/5  1/5
## [2,]    1 -1/4 -1/4 -1/4 -1/4
## [3,]    0    1 -1/3 -1/3 -1/3
## [4,]    0    0    1 -1/2 -1/2
## [5,]    0    0    0    1   -1
```

you see that we go 'forward': after the general mean, we see that country A is compared to (B+C+D+E)/4, then country B is compared to (C+D+E)/3; next, country C is compared to (D + E)/2 and lastly D is compared to E. This forward type of Helmert coding is usually the default Helmert coding in other statistical packages.

Helmert contrasts are useful when you have a particular logic for ordering the levels of a categorical variable (i.e. if you have an ordinal variable). An example where a Helmert contrast might be useful is when you want to know the minimal effective dose in a dose-response study. Suppose we have a new medicine and we want to know how much of it is needed to end a headache. We split individuals with headache into four groups. In group A, people take a placebo pill, in group B people take 1 tablet of the new medicine, in group C people take 3 tablets, and in group D people take 4 tablets. In the data analysis, we first want to establish whether 1 tablet leads to a significant reduction compared to placebo. If that is not the case, we can test whether 3 tablets (group C) work better than either 1 tablet or no tablet (groups A and B combined). If that doesn't show a significant improvement either, we may want to compare 6 tablets (group D) versus fewer tablets (groups A, B and C combined).

There are a couple of other standard options in R for coding your variables. *Successive differences contrast coding* is also an option for meaningfully ordered categories. It compares the second level with the first level, the third level with the second level, the fourth level with the third level, and so on. The function to be used is `code_diff` (available in the `codingMatrices` package). It could be a useful alternative to Helmert coding for a dose-response study. Other alternatives are given in Table 10.7.

Table 10.7: Contrast options available in the ‘codingMatrices‘ package.

option	description
code_control	For contrasts that compare the group means (the 'treatments') with the first class mean (the reference group).
code_control_last	Similar to 'code_control', but using the final class mean as the reference group
code_diff	The contrasts are the successive differences of the treatment means, e.g. group 2 - group 1, group 3 - group 2, etc.
code_diff_forward	Very similar to 'code_diff', but using forward differences: group 1 vs group 2, group 2 - group 3, etc.
code_helmert	The contrasts now compare each group mean, starting from the second, with the average of all group previous group means.
code_helmert_forward	Similar to code_helmert, but comparing each class mean, up to the second last, with the average of all class means coming after it.
code_deviation	Effect coding or sum-to-zero coding. A more precise description might be to say that the contrasts are the deviations of each group mean from the average of the other ones. To avoid redundancy, the last deviation is omitted.
code_deviation_first	Very similar to code_deviation, but omitting the first deviation to avoid redundancy rather than the last.
code_poly	For polynomial contrasts. When you are interested in linear, quadratic, and cubic effects of different treatment levels. (advanced)
contr.diff	Very similar in effect to 'code_diff', yielding the same differences as the contrasts, but using the first group mean as the intercept rather than the simple average of the remaining class means, as with 'code_diff'. Some would regard this as making it unsuitable for use in some non-standard ANOVA tables, so use with care.

Overview

Pre-set coding schemes in R using the `codingMatrices` package

1) Default is dummy coding (referring to the coding scheme **S**), also called treatments contrasts (referring to matrix **L**). R can compute this coding scheme matrix by using `code_control`. With this coding, the first level of the factor variable is used as the reference group. Alternatively, with `code_control_last`, dummy coding is used with the last level of the factor variable as the reference group.
2) For ANOVAs, generally sum-to-zero coding is used, also called effects coding. R can compute the coding scheme matrix using `code_deviation`. It is the same as `contr.sum` in the `lm()` function.
3) For categories that are ordered meaningfully, Helmert contrasts are often used. R can compute the coding scheme matrix using `code_helmert`. Another option is successive differences contrast coding, using `code_diff` or `code_diff_forward`.

10.9 Custom-made contrasts

Sometimes, the standard contrasts available in the `codingMatrices` package are not what you are looking for. In this section we will look at a situation where we have more elaborate research questions that we can only answer with special, tailor-made contrasts. We will give an example of a data set with ten groups.

Thus far, we've seen different ways of specifying contrasts and in each case, *the number of contrasts was equal to the number of categories*. In the two-category dummy coding examples, you saw two categories and the output shows an intercept (contrast 1) and a slope (contrast 2). In the three-category dummy example, we saw an intercept for one category (the reference group), one slope for the second category vs. the first category and one slope for the third category vs. the first category. In the Helmert example, we saw the same thing: for three categories and Helmert coding, we saw one general mean (the intercept), then the difference between group 2 and 1, and then the difference between group 3 on the one hand and group 1 and 2 combined on the other hand. Thus in general we state that if we run a model for a variable with J categories, we see J parameters in the output (including the intercept). We will not go into the details here but it is paramount that even when you have fewer than J questions, you should still have J contrasts in your analysis. We will illustrate

Table 10.8: Overview of the grapes used in a wine tasting study.

Wine_ID	grape	origin	colour
1	Cabernet Sauvignon	French	Red
2	Merlot	French	Red
3	Airén	Spanish	White
4	Tempranillo	Spanish	Red
5	Chardonnay	French	White
6	Syrah (Shiraz)	French	Red
7	Garnacha (Garnache)	Spanish	Red
8	Sauvignon Blanc	French	White
9	Trebbiano Toscana (Ugni Blanc)	Italian	White
10	Pinot Noir	French	Red

Table 10.9: Part of the wine tasting data.

winetasterID	Wine_ID	grape	origin	rating
1	9	Trebbiano Toscana (Ugni Blanc)	Italian	20
2	8	Sauvignon Blanc	French	8
3	8	Sauvignon Blanc	French	27
4	1	Cabernet Sauvignon	French	70
5	9	Trebbiano Toscana (Ugni Blanc)	Italian	32

this principle here. In this section we show you an example where it can be very helpful to take full control over the contrasts that R computes for you, but where we must make sure that the number of parameters in the output matches the total number of categories.

Suppose you do a wine tasting study, where you have ten different wines, each made from one of ten different grapes, and each evaluated by one of ten different groups of people. Each person tastes only one wine and gives a score between 1 (worst quality imaginable) and 100 (best quality imaginable). Table 10.8 gives an overview of the ten wines made from ten different grapes.

The units of this research are the people that taste the wine. The kind of wine is a grouping variable: some people taste the Sauvignon Blanc, whereas other people taste the Pinot Noir. Table 10.9 shows a small part of the data.

Suppose we would like to answer the following research questions:

1. How large is the difference in quality experienced between the red and the white wines?
2. How large is the difference in quality experienced between French origin and Spanish origin grapes?

3. How large is the difference in quality experienced between the Italian origin grapes and the other two origins?

Note that these three questions can be conceived as three contrasts for the wine tasting data, whereas the grouping variable has ten categories. We will come back to this issue later.

We address these three research questions by translating them into three contrasts for the factor variable **Wine_ID**. Suppose that the ordering of the grapes is similar to above, where Cabernet Sauvignon is the first level, Merlot is the second level, etc.

Contrast $L1$ is about the difference between the mean of 6 red wine grapes and the mean of 4 white wine grapes:

$$L1 : \frac{M_1 + M_2 + M_4 + M_6 + M_7 + M_{10}}{6} - \frac{M_3 + M_5 + M_8 + M_9}{4}$$

This is equivalent to:

$$L1 : \frac{1}{6}M_1 + \frac{1}{6}M_2 - \frac{1}{4}M_3 + \frac{1}{6}M_4 - \frac{1}{4}M_5 + \frac{1}{6}M_6 + \frac{1}{6}M_7 - \frac{1}{4}M_8 - \frac{1}{4}M_9 + \frac{1}{6}M_{10}$$

We can store this contrast for the variable grape as row vector

$$\begin{bmatrix} \frac{1}{6} & \frac{1}{6} & -\frac{1}{4} & \frac{1}{6} & -\frac{1}{4} & \frac{1}{6} & \frac{1}{6} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{6} \end{bmatrix}$$

The second question can be answered by the following contrast where we have the mean of the 6 French mean ratings minus the mean of the 3 Spanish mean ratings:

$$L2 : \frac{M_1 + M_2 + M_5 + M_6 + M_8 + M_{10}}{6} - \frac{M_3 + M_4 + M_7}{3}$$

This is equivalent to

$$L2 : \frac{1}{6}M_1 + \frac{1}{6}M_2 - \frac{1}{3}M_3 - \frac{1}{3}M_4 + \frac{1}{6}M_5 + \frac{1}{6}M_6 - \frac{1}{3}M_7 + \frac{1}{6}M_8 + 0 \times M_9 + \frac{1}{6}M_{10}$$

We can store this contrast for the variable grape as row vector

$$\begin{bmatrix} \frac{1}{6} & \frac{1}{6} & -\frac{1}{3} & -\frac{1}{3} & \frac{1}{6} & \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} & 0 & \frac{1}{6} \end{bmatrix}$$

The third question can be stated as the contrast

$$L3: M_9 - \frac{\left(\frac{M_1+M_2+M_5+M_6+M_8+M_{10}}{6} + \frac{M_3+M_4+M_7}{3}\right)}{2}$$

In other words, we take the grand mean for the 6 French wines, the grand mean of the 3 Spanish wines, and take the average of these two grand means. We then contrast this grand mean with the mean rating of the single Italian wine. This contrast is equivalent to

$$L3: -\frac{1}{12}M_1 - \frac{1}{12}M_2 - \frac{1}{6}M_3 - \frac{1}{6}M_4 - \frac{1}{12}M_5 - \frac{1}{12}M_6 - \frac{1}{6}M_7 - \frac{1}{12}M_8 + 1 \times M_9 - \frac{1}{12}M_{10}$$

and can be stored in row vector

$$\begin{bmatrix} -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{6} & -\frac{1}{12} & 1 & -\frac{1}{12} \end{bmatrix}$$

Combining the three row vectors, you get the contrast matrix **L**

$$\mathbf{L} = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & -\frac{1}{4} & \frac{1}{6} & -\frac{1}{4} & \frac{1}{6} & \frac{1}{6} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & -\frac{1}{3} & -\frac{1}{3} & \frac{1}{6} & \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} & 0 & \frac{1}{6} \\ -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{12} & -\frac{1}{12} & -\frac{1}{6} & -\frac{1}{12} & 1 & -\frac{1}{12} \end{bmatrix}$$

We can enter this matrix **L** in R, and then compute the inverse:

```
l1 <- c(1/6, 1/6, -1/4, 1/6, -1/4, 1/6, 1/6, -1/4, -1/4, 1/6)
l2 <- c(1/6, 1/6, -1/3, -1/3, 1/6, 1/6, -1/3, 1/6, 0, 1/6)
l3 <- c(-1/12, -1/12, -1/6, -1/6, -1/12, -1/12, -1/6, -1/12, 1, -1/12)

L <- rbind(l1, l2, l3) # bind the rows together
L %>% fractions()
```

```
##     [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]  [,10]
## l1   1/6   1/6  -1/4   1/6  -1/4   1/6   1/6  -1/4  -1/4   1/6
## l2   1/6   1/6  -1/3  -1/3   1/6   1/6  -1/3   1/6     0   1/6
## l3 -1/12 -1/12  -1/6  -1/6 -1/12 -1/12  -1/6 -1/12     1 -1/12
```

```
S <- ginv(L) # compute the inverse
S
```

```
##                [,1]       [,2]          [,3]
##  [1,]  4.000000e-01  0.3333333 -2.193395e-17
##  [2,]  4.000000e-01  0.3333333 -8.707943e-20
##  [3,] -8.000000e-01 -0.6166667 -3.000000e-01
```

```
##  [4,]  4.000000e-01 -0.6666667 -2.841861e-17
##  [5,] -8.000000e-01  0.3833333 -3.000000e-01
##  [6,]  4.000000e-01  0.3333333 -4.437680e-17
##  [7,]  4.000000e-01 -0.6666667 -2.841861e-17
##  [8,] -8.000000e-01  0.3833333 -3.000000e-01
##  [9,]  1.164701e-17 -0.1500000  9.000000e-01
## [10,]  4.000000e-01  0.3333333 -4.437680e-17
```

Note that we have three new variables to be computed. An intercept is missing, but R will include that by default when running a linear model.

Let's invent some data and submit that to R.

```
# inventing two ratings for each of the 10 wines
winedata <- tibble(rating = c(70,88,81,25,59,2,93,27,32,23,13,21,8,7,76,5,57,8,20,96),
                   Wine_ID = rep(1:10, 2) %>% as.factor(),
                   grape = c("Cabernet Sauvignon",
                             "Merlot",
                             "Airén",
                             "Tempranillo",
                             "Chardonnay",
                             "Syrah (Shiraz)",
                             "Garnacha (Garnache)",
                             "Sauvignon Blanc",
                             "Trebbiano Toscana (Ugni Blanc)",
                             "Pinot Noir") %>% rep(2),
                   colour = factor(c(1, 1, 2, 1, 2, 1, 1, 2, 2, 1),
                                   labels = c("Red", "White")) %>% rep(2),
                   origin = factor(c(1, 1, 3, 3, 1, 1, 3, 1, 2, 1),
                                   labels = c("French", "Italian", "Spanish")) %>% rep
```

Let's look at the group means per grape:

```
winedata %>%
  group_by(grape) %>%
  summarise(mean = mean(rating))
```

```
## # A tibble: 10 x 2
##    grape                          mean
##    <chr>                         <dbl>
##  1 Airén                          44.5
##  2 Cabernet Sauvignon             41.5
##  3 Chardonnay                     67.5
##  4 Garnacha (Garnache)            75
##  5 Merlot                         54.5
```

```
##  6 Pinot Noir                       59.5
##  7 Sauvignon Blanc                  17.5
##  8 Syrah (Shiraz)                    3.5
##  9 Tempranillo                      16
## 10 Trebbiano Toscana (Ugni Blanc)   26
```

For the first research question, we want to estimate the difference in mean rating between the red wines and the white wines.

If we look at the means and compute this difference by hand, by calculating the grand means for red and white wines separately and take the difference, we get $41.7 - 38.9 = 2.8$.

It is a bit tiresome and error prone to compute things by hand, while we have such a great computing device as R available to us. Alternatively therefore we can compute this by simply taking the weighted sum of all the means, with the weights from the first contrast $L1$. In R this is easy: we take the vector of means for the 10 different groups and piecewise multiply them by the $L1$ contrast weights that we have in `l1`, and then sum them (in algebra this is called the *inner product*, if that rings any bell). It's exactly the same as computing the weighted sum of the means.

```
# compute the sample means for each level of the Wine_ID variable
means <- winedata %>%
  group_by(Wine_ID) %>% # for each level of the Wine_ID factor
  summarise(mean = mean(rating)) %>%  # compute the mean per level
  dplyr::select(mean) %>% # select only the column of the means
  as_vector() # turn the column into a vector
means
```

```
##  mean1  mean2  mean3  mean4  mean5  mean6  mean7  mean8  mean9 mean10
##   41.5   54.5   44.5   16.0   67.5    3.5   75.0   17.5   26.0   59.5
```

```
l1 %*% means  # take a weighted sum of the group means
```

```
##          [,1]
## [1,] 2.791667
```

The difference between the grand means of the French and Spanish wine grapes (research question 2) can be computed by using the weights for contrast $L2$:

```
l2 %*% means
```

```
##      [,1]
## [1,] -4.5
```

We see that the difference between the French and Spanish grand means equals -4.5.

The third question was about the Italian mean versus the mean of the grand means of the French and Spanish wines. if we take the weights from contrast $L3$ we get

```
l3 %*% means
```

```
##           [,1]
## [1,] -16.91667
```

The output tells us that the difference in rating of the Italian wines and the other wines combined equals -16.9.

If we want to have information about these differences in the *population*, such as standard errors, confidence intervals and null-hypothesis testing, R needs the coding scheme matrix and run the linear model. We can do that by providing the **S** matrix for the relevant factor in the analysis.

First we need to assign matrix **S** to the variable **Wine_ID**. Now pay close attention. Remember that **S** contained 3 variables, one for each of the research questions.

```
S %>% fractions()
```

```
##       [,1]   [,2]   [,3]
##  [1,]  2/5    1/3      0
##  [2,]  2/5    1/3      0
##  [3,] -4/5 -37/60  -3/10
##  [4,]  2/5   -2/3      0
##  [5,] -4/5  23/60  -3/10
##  [6,]  2/5    1/3      0
##  [7,]  2/5   -2/3      0
##  [8,] -4/5  23/60  -3/10
##  [9,]    0  -3/20   9/10
## [10,]  2/5    1/3      0
```

```
contrasts(winedata$Wine_ID) <- S
```

But now look at the variable `winedata$Wine_ID`:

```
contrasts(winedata$Wine_ID)
```

```
##             [,1]       [,2]          [,3]          [,4]          [,5]
## 1   4.000000e-01  0.3333333 -2.193395e-17 -4.655128e-01 -3.703222e-01
## 2   4.000000e-01  0.3333333 -8.707943e-20  5.080988e-01 -2.142072e-01
## 3  -8.000000e-01 -0.6166667 -3.000000e-01 -7.710846e-02  1.500931e-01
## 4   4.000000e-01 -0.6666667 -2.841861e-17 -1.657550e-02 -1.073408e-01
## 5  -8.000000e-01  0.3833333 -3.000000e-01  5.385542e-01 -7.504653e-02
## 6   4.000000e-01  0.3333333 -4.437680e-17 -5.984724e-02  8.673112e-01
## 7   4.000000e-01 -0.6666667 -2.841861e-17  9.368396e-02 -4.275228e-02
## 8  -8.000000e-01  0.3833333 -3.000000e-01 -4.614458e-01 -7.504653e-02
## 9   1.164701e-17 -0.1500000  9.000000e-01  3.459485e-17  3.538807e-18
## 10  4.000000e-01  0.3333333 -4.437680e-17 -5.984724e-02 -1.326888e-01
##             [,6]          [,7]          [,8]          [,9]
## 1  -7.677688e-02 -4.655128e-01 -2.533806e-01 -3.703222e-01
## 2  -3.116414e-01  5.080988e-01 -2.693679e-01 -2.142072e-01
## 3  -4.077412e-01 -7.710846e-02 -4.703661e-01  1.500931e-01
## 4  -3.273038e-01 -1.657550e-02  6.931016e-01 -1.073408e-01
## 5   2.038706e-01 -4.614458e-01  2.351831e-01 -7.504653e-02
## 6  -9.661462e-03 -5.984724e-02  2.619120e-02 -1.326888e-01
## 7   7.350450e-01  9.368396e-02 -2.227354e-01 -4.275228e-02
## 8   2.038706e-01  5.385542e-01  2.351831e-01 -7.504653e-02
## 9   1.590885e-17 -1.439127e-17  9.098508e-17  1.741659e-17
## 10 -9.661462e-03 -5.984724e-02  2.619120e-02  8.673112e-01
```

It seems that R automatically added six other columns. The total number of columns is now nine, and if we run a linear model together with an intercept we will see ten parameters in our linear model output.

We tell R to run a linear model, with **rating** as the dependent variable, **Wine_ID** as the independent variable, where the coding scheme matrix is the new coding scheme matrix.

```
winedata %>%
  lm(rating ~ Wine_ID, data = .) %>%
  tidy(conf.int = TRUE)
```

```
## # A tibble: 10 x 7
##    term        estimate std.error statistic  p.value conf.low conf.high
##    <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
##  1 (Intercept)    40.5       7.21     5.62  0.000221     24.5      56.6
##  2 Wine_ID1        2.79     14.7      0.190 0.853       -30.0      35.6
##  3 Wine_ID2       -4.50     16.1     -0.279 0.786       -40.4      31.4
##  4 Wine_ID3      -16.9      24.2     -0.699 0.500       -70.8      37.0
##  5 Wine_ID4       36.2      22.8      1.59  0.144       -14.6      87.0
##  6 Wine_ID5      -36.5      22.8     -1.60  0.140       -87.4      14.3
##  7 Wine_ID6       28.3      22.8      1.24  0.243       -22.5      79.1
```

```
##  8 Wine_ID7      -13.8      22.8     -0.604 0.559       -64.6      37.0
##  9 Wine_ID8      -30.1      22.8     -1.32  0.216       -80.9      20.7
## 10 Wine_ID9       19.5      22.8      0.854 0.413       -31.4      70.3
```

In the output we see ten different parameters. The first (intercept) we can ignore as it has no research question associated with it. The next three parameters do have research questions associated with them.

We see the values of 2.79 for research question 1, -4.50 for research question 2 and -16.9 for research question 3, similar to what we found by looking at the inner products (verify this for yourself). This check tells us that we have indeed computed the contrasts that we wanted to compute. In addition to these estimates for the contrasts, we see standard errors, null-hypothesis tests and 95% confidence intervals for these contrasts. We can ignore all other parameters (`Wine_ID4` thru `Wine_ID_ID9`): these are only there to make sure that we have as many parameters as we have categories, so that the residual error variance is computed correctly.

Note that also the intercept was included automatically. In order to see what it represents, we can add this intercept of 1s to $\mathbf{S}$ and take the inverse to obtain the contrast matrix

```
cbind(1, contrasts(winedata$Wine_ID)) %>% ginv()
```

```
##              [,1]        [,2]        [,3]       [,4]        [,5]          [,6]
##  [1,]  0.10000000  0.10000000  0.10000000  0.1000000  0.10000000  0.100000000
##  [2,]  0.16666667  0.16666667 -0.25000000  0.1666667 -0.25000000  0.166666667
##  [3,]  0.16666667  0.16666667 -0.33333333 -0.3333333  0.16666667  0.166666667
##  [4,] -0.08333333 -0.08333333 -0.16666667 -0.1666667 -0.08333333 -0.083333333
##  [5,] -0.46551281  0.50809884 -0.07710846 -0.0165755  0.53855423 -0.059847243
##  [6,] -0.37032217 -0.21420722  0.15009306 -0.1073408 -0.07504653  0.867311224
##  [7,] -0.07677688 -0.31164140 -0.40774121 -0.3273038  0.20387060 -0.009661462
##  [8,] -0.46551281  0.50809884 -0.07710846 -0.0165755 -0.46144577 -0.059847243
##  [9,] -0.25338058 -0.26936792 -0.47036610  0.6931016  0.23518305  0.026191197
## [10,] -0.37032217 -0.21420722  0.15009306 -0.1073408 -0.07504653 -0.132688776
##              [,7]        [,8]          [,9]        [,10]
##  [1,]  0.10000000  0.10000000  1.000000e-01  0.100000000
##  [2,]  0.16666667 -0.25000000 -2.500000e-01  0.166666667
##  [3,] -0.33333333  0.16666667 -5.620379e-18  0.166666667
##  [4,] -0.16666667 -0.08333333  1.000000e+00 -0.083333333
##  [5,]  0.09368396 -0.46144577 -1.629503e-17 -0.059847243
##  [6,] -0.04275228 -0.07504653  1.186856e-17 -0.132688776
##  [7,]  0.73504500  0.20387060 -7.619694e-17 -0.009661462
##  [8,]  0.09368396  0.53855423  7.997995e-18 -0.059847243
##  [9,] -0.22273545  0.23518305  1.209841e-16  0.026191197
## [10,] -0.04275228 -0.07504653  3.908467e-17  0.867311224
```

We see that the first row yields the grand mean over all 10 wines:

$$\frac{1}{10}M_1 + \frac{1}{10}M_2 + \frac{1}{10}M_3 + \frac{1}{10}M_4 + \frac{1}{10}M_5 + \frac{1}{10}M_6 + \frac{1}{10}M_7 + \frac{1}{10}M_8 + \frac{1}{10}M_9 + \frac{1}{10}M_{10} = \frac{M_1 + M_2 + M_3 + M_4 + M_5 + M_6 + M_7 + M_8 + M_9 + M_{10}}{10}$$

You may ignore this value in the output, if it is not of interest.

Overview

User-defined contrasts in an `lm()` analysis

1) Check the order of your levels using `levels()`.
2) Specify your contrast row vectors and combine them into a matrix **L**.
3) Have R compute matrix **S**, the inverse of matrix **L**, using `ginv()` from the `MASS` package.
4) Assign the matrix **S** to the grouping variable, so something like `contrasts(data$variable) <- S`.
5) Run the linear model.
6) R will automatically generate an intercept. Its value in the output will often be the grand mean. The remaining parameters are estimates for your contrasts in **L**. If the number of rows in **L** is smaller than the number of categories, R will automatically add extra columns in the actual coding scheme matrix. The parameters for these columns in the output can be ignored.

10.10 Contrasts in the case of two categorical variables

Up till now we only discussed situations with one categorical independent variable in your data analysis. Here we discuss the situation where we have two independent variables. We start with two categorical independent variables. The next section will discuss the situation with one numerical and one categorical independent variable.

Specifying contrasts for two categorical variables is a straightforward extension of what we saw earlier. For each categorical variable, you can specify what contrasts you would like to have in the output.

Let's have a look at a data set called `jobsatisfaction` to see how that is done practically. The dataframe is part of the package `datarium` (don't forget to install that package, if you haven't already).

```
data("jobsatisfaction", package = "datarium")

jobsatisfaction %>%
  glimpse()
```

```
## Rows: 58
## Columns: 4
## $ id              <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ gender          <fct> male, male, male, male, male, male, male, male, male, ~
## $ education_level <fct> school, school, school, school, school, school, school~
## $ score           <dbl> 5.51, 5.65, 5.07, 5.51, 5.94, 5.80, 5.22, 5.36, 4.78, ~
```

We see one numeric variable `score` and three factor variables. Factors are always stored with the categories in a certain order. Let's have a look at the variable gender.

```
jobsatisfaction$gender %>% levels()
```

```
## [1] "male"   "female"
```

Variable gender has two levels, with "male" being the first. Note that the order is not alphabetic, so it was clearly intended by the person who prepared this dataframe for the males to be the first category.

The pre-specified coding matrix for this variable can also be inspected:

```
contrasts(jobsatisfaction$gender)
```

```
##        female
## male        0
## female      1
```

We see that it is standard dummy coding (treatment coding) with a dummy variable for being female, so that "male" will be the reference category if we don't change anything.

Let's now look at the factor variable `education_level`.

```
jobsatisfaction$education_level %>% levels()
```

```
## [1] "school"     "college"    "university"
```

Again, the ordering is not alphabetical, but in an ordered fashion from 'lower' to 'higher' education.

The coding scheme matrix shows standard dummy coding, with "school" being the reference category.

```
contrasts(jobsatisfaction$education_level)
```

```
##            college university
## school           0          0
## college          1          0
## university       0          1
```

Suppose we want to run a linear model where job satisfaction is predicted or explained by gender and education level. For the gender effect, we are content with the males forming the reference group, so we keep treatment coding. We also are content with dummy coding for **education_level** with "school" being the reference category.

```
# run linear model with gender and education_level
out1 <- jobsatisfaction %>%
  lm(score ~ gender + education_level,
     data = .)
out1 %>% tidy()
```

```
## # A tibble: 4 x 5
##   term                       estimate std.error statistic  p.value
##   <chr>                         <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                   5.66      0.164    34.6   1.66e-38
## 2 genderfemale                 -0.125     0.161    -0.777 4.41e- 1
## 3 education_levelcollege        0.757     0.198     3.82  3.47e- 4
## 4 education_leveluniversity     3.25      0.196    16.6   7.01e-23
```

In the output we see that job satisfaction scores are on average 0.125 points *lower* in females than in males. Furthermore, we see that the contrast between college and school (`education_levelcollege`) equals 0.757. That means that job satisfaction scores in male college graduates are 0.757 higher than those males with only high school. The contrast between university and school (`education_leveluniversity`) shows that university graduates score 3.25 points higher on job satisfaction than school graduates. The intercept equals 5.66, meaning that the people that are in the reference group of male students, graduated from school, score on average 5.66.

To check whether we are right in our interpretations, we can check the group means.

```
jobsatisfaction %>%
  group_by(education_level, gender) %>%
  summarise(mean = mean(score))
```

```
## `summarise()` has grouped output by 'education_level'. You can override using
## the `.groups` argument.
```

```
## # A tibble: 6 x 3
## # Groups:   education_level [3]
##   education_level gender  mean
##   <fct>           <fct>  <dbl>
## 1 school          male    5.43
## 2 school          female  5.74
## 3 college         male    6.22
## 4 college         female  6.46
## 5 university      male    9.29
## 6 university      female  8.41
```

To check interpretation of the intercept, we do

```
jobsatisfaction %>%
  filter((gender == "male") & (education_level == "school")) %>% # only males with sch
  summarise(mean = mean(score))  # compute mean
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  5.43
```

We see a slight discrepancy: the observed mean is not what the linear model predicts!

The interpretation of the output that we just discussed and the predictions that are made are only valid if the model is a good representation of the actual situation in the population. In reality, there might be *moderation*: the effect of gender might be moderated by education level. That is, the difference between females and males may be different, depending on the education level. The model with only two main effects assumes that the differences between females and males is -0.125, regardless of education. If we want to know whether this difference in gender is moderated by education, we need to look for interaction effects. For that we need to include an interaction term in the model. We keep the default treatment coding (dummy coding).

```
out2 <- jobsatisfaction %>%
  lm(score ~ gender + education_level + gender:education_level, data = .)
out2 %>% tidy(conf.int = TRUE)
```

```
## # A tibble: 6 x 7
##   term                   estimate std.error statistic  p.value conf.low conf.high
##   <chr>                     <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)              5.43       0.183    29.6   3.26e-34    5.06      5.79
## 2 genderfemale             0.314      0.253     1.24  2.19e- 1   -0.193     0.821
## 3 education_levelcolle~    0.797      0.259     3.07  3.37e- 3    0.276     1.32
## 4 education_levelunive~    3.87       0.253    15.3   6.87e-21    3.36      4.37
## 5 genderfemale:educati~   -0.0747     0.357    -0.209 8.35e- 1   -0.792     0.643
## 6 genderfemale:educati~   -1.20       0.353    -3.40  1.29e- 3   -1.91     -0.493
```

Now we are only interested in whether the interaction term is different from 0. If we look at the `out2` output, we see two interaction terms: one for the extra gender effect in college students (compared to high school graduates), and one for the extra gender effect in university graduates. In Chapter 9 we learned that if we want to know whether the overall moderation is significant, we should perform an analysis of variance.

```
library(car)
jobsatisfaction %>%
  lm(score ~ gender + education_level + gender:education_level, data = .,
     contrasts = list(gender = contr.sum, education_level = contr.sum)) %>%
  Anova(type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: score
##                          Sum Sq Df   F value    Pr(>F)
## (Intercept)            2774.84  1 9172.2752 < 2.2e-16 ***
## gender                    0.18  1    0.5856  0.447601
## education_level         114.46  2  189.1724 < 2.2e-16 ***
## gender:education_level    4.44  2    7.3379  0.001559 **
## Residuals                15.73 52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We compare the gender effect across three different education levels, so we should see two degrees of freedom for the interaction term. The F-value is more than 1, and shows to be significant at the 0.05 level, $F(2, 52) = 7.34, p = .002$. We can therefore conclude that the gender effect is different, depending on education level.

Now this `Anova()` output and the `lm()` output only tell us that the gender effect is different for different education levels: that the gender effect is -0.07 larger in college graduates than in school graduates, and -1.20 larger in university graduates, compared to school graduates. However, what if one research question was to estimate the gender effect for various education levels? That is, how large is the effect of gender in school graduates, how large is the effect of gender in college graduates, and how large is the gender effect in university graduates?

We should be able to get estimates for these quantities by working with the various parameters in the output of the `lm()` analysis. For instance, the gender effect in school graduates must be equal to parameter 2 (`genderfemale`). Since the reference group is male school graduates, the `genderfemale` parameter gives the difference with female school graduates. This parameter has a standard error and we can ask for the confidence interval. But it is trickier when we want to estimate the gender effect in college graduates and university graduates. On the basis of the parameters in the model output, we could figure out the expected gender effects in these groups, but we would not have information about the standard errors and consequently the confidence intervals.

Mathematically, the standard error of a linear combination of parameters depends on the standard errors of the original parameters and the extent that the estimates are related to each other (correlated). But since most of us are not mathematicians, we do not want to worry about that. We simply ask R to do the necessary computations for us. We show you now how to obtain standard errors and confidence intervals for the effect of one categorical variable, given the level of an other categorical level.

We use the `reghelper` package (install it first). We first run the model with the interaction effects included, and then use the function `simple_slopes()` on the output of the model.

```
# using reghelper package
library(reghelper)

jobsatisfaction %>%
  lm(score ~ gender + education_level + gender:education_level, data = .) %>%
  simple_slopes(confint = TRUE,
                ci.width = 0.95)
```

```
##   gender     education_level Test Estimate Std. Error    2.5%   97.5% t value
## 1   male    sstest (college)        0.7967     0.2593  0.2764  1.3170  3.0726
## 2   male sstest (university)        3.8653     0.2527  3.3582  4.3724 15.2950
## 3 female    sstest (college)        0.7220     0.2460  0.2284  1.2156  2.9352
## 4 female sstest (university)        2.6650     0.2460  2.1714  3.1586 10.8343
## 5 sstest              school        0.3143     0.2527 -0.1928  0.8214  1.2438
## 6 sstest             college        0.2397     0.2527 -0.2674  0.7468  0.9484
```

```
## 7 sstest          university       -0.8860     0.2460 -1.3796 -0.3924 -3.6020
##   df  Pr(>|t|) Sig.
## 1 52 0.0033741   **
## 2 52 < 2.2e-16  ***
## 3 52 0.0049525   **
## 4 52 6.074e-15  ***
## 5 52 0.2191473
## 6 52 0.3473349
## 7 52 0.0007058  ***
```

In the output you see seven rows. In the first row, you see the effect of being a college graduate (reference group is "school"), given that one is male. In other words, in males we see that the difference between college graduates and school graduates equals 0.7967, with standard error of 0.2593 and the 95% confidence interval, and a null-hypothesis test.

We ignore the first four rows, because they are about the effects of **education_level** given a particular value for gender (male or female). Because we are interested in the effect of gender, given a certain level of education level, we only inspect the last three lines. In lines 5, 6, and 7, we see that the effect of gender equals 0.31 for school graduates, 0.24 for college graduates and -0.89 for university graduates. The respective standard errors, confidence intervals and null-hypothesis tests are also given. We can therefore report:

> "In a data set on job satisfaction, we estimated the gender effect (females relative to males) at three levels of education: school, college and university graduates. Results showed that in school graduates, the gender difference was 0.314 (SE = 0.253, 95% CI: -0.193, 0.821). In college graduates the difference was 0.240 (SE = 0.253 , 95% CI: -0.267, 0.747) and in university graduates the difference was -0.886 (SE = 0.246, 95% CI: -1.380, -0.392)."

Note that the general term "simple slopes" refers to the effects of one independent variable on a dependent variable, given a certain value for an other independent variable. Another term for this is "simple effects". Such simple effects can be studied when you have two categorical independent variables, but also in situations with numeric independent variables, as we will see in the next section.

10.11 Contrasts in the case of one categorical variable and one numeric variable

Let's return to the example in Chapter 9 where we analysed vocabulary as a function of age and socio-economic status (SES). The dependent variable was

Table 10.10: Example data set on vocabulary in children from country X.

childID	SES	age	words
1	average	4	111
2	low	1	94
3	average	4	116
4	average	1	82
5	high	1	103
6	low	4	124

the number of words one knows. We looked for moderation of the effect of **age** (numerical), by **SES** (categorical). In other words, we modelled a different linear relationship between **age** and **vocabulary**, depending on **SES**. For this chapter, we extend that example a bit, by introducing three levels of SES: low, average and high.

Let's suppose we want to know what the slopes are for each of the SES levels. The data have the structure as in Table 10.10. In total the data came from 100 children, with one measurement per child.

Similar to the previous section, we ask what the effect is of one variable, given a certain level of the second variable. The only difference now is that the first variable is numeric, instead of categorical. Similar to the previous section therefore, we can solve the problem in the following way.

We run a standard analysis, using dummy coding and look at the output, and make sure that we understand the values in the output correctly. For that we need to know the ordering of the levels and the coding scheme.

```
levels(data_words$SES) # verify level ordering
```

```
## NULL
```

```
data_words <- data_words %>%
  mutate(SES = factor(SES))  # turn SES into a factor variable
levels(data_words$SES)
```

```
## [1] "average" "high"    "low"
```

```
contrasts(data_words$SES) # verify dummy coding
```

```
##         high low
## average    0   0
```

```
## high       1   0
## low        0   1
```

```
out_words <- data_words %>%
  lm(words ~ age + SES + age:SES, data = .)  # run model with dummy coding
out_words %>% tidy()
```

```
## # A tibble: 6 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    87.0       4.95    17.6   1.90e-31
## 2 age             3.70      1.53     2.42  1.76e- 2
## 3 SEShigh        12.5       6.66     1.87  6.43e- 2
## 4 SESlow          8.81      7.12     1.24  2.19e- 1
## 5 age:SEShigh    -4.04      2.04    -1.98  5.06e- 2
## 6 age:SESlow     -1.85      2.19    -0.844 4.01e- 1
```

In Chapter 9 we learned how to read this output. We see that the "average" SES group is the reference group. The effect of age for this group is equal to 3.70 (the slope).

Now let's use the `simple_slopes()` function again to also see the effect of age for the other SES groups.

```
library(reghelper)
data_words %>%
  lm(words ~ age + SES + age:SES, data = .) %>%
  simple_slopes(confint = TRUE,
                ci.width = 0.95)
```

```
##        age           SES Test Estimate Std. Error     2.5%   97.5% t value df
## 1 1.803566 sstest (high)        5.1778     3.5549  -1.8805 12.2360  1.4565 94
## 2 1.803566  sstest (low)        5.4767     3.7220  -1.9134 12.8669  1.4714 94
## 3     3.03 sstest (high)        0.2204     2.5593  -4.8612  5.3020  0.0861 94
## 4     3.03  sstest (low)        3.2097     2.5773  -1.9076  8.3271  1.2454 94
## 5 4.256434 sstest (high)       -4.7369     3.6048 -11.8942  2.4204 -1.3141 94
## 6 4.256434  sstest (low)        0.9427     3.7217  -6.4468  8.3322  0.2533 94
## 7   sstest       average        3.6957     1.5297   0.6583  6.7330  2.4159 94
## 8   sstest          high       -0.3464     1.3511  -3.0290  2.3362 -0.2564 94
## 9   sstest           low        1.8472     1.5662  -1.2626  4.9570  1.1794 94
##   Pr(>|t|) Sig.
## 1  0.14858
## 2  0.14451
## 3  0.93155
## 4  0.21610
```

```
## 5  0.19202
## 6  0.80059
## 7  0.01763    *
## 8  0.79820
## 9  0.24122
```

In the last three rows, we see the effects (slopes) of age given average, high or low SES. Hence we can report:

> "In a data set on the vocabulary size of 100 children of various ages, we estimated the slope coefficient for the regression of number of words on age in years at three levels of SES: low, average and high. Results showed that in low SES children, the slope was 1.85 (SE = 1.57, 95% CI: -1.26, 4.96). In average SES children the slope was 3.70 (SE = 1.53, 95% CI: 0.66, 6.73) and in high SES children the slope was -0.35 (SE = 1.35, 95% CI: -3.03, 2.34)."

Now that our research questions are answered, you may wonder how to interpret the other six lines in the simple slopes output. For instance, in the first line, we see the effect of having a high SES (relative to "average"), given a certain value for age, namely 1.803566 years. In other words, for children of age 1.80 (almost two years old), we see that high SES children score 5.1778 points higher on vocabulary than average SES children.

In the second line we see that in children of age 1.80, low SES children score 5.4767 points higher than average SES children.

In the next two lines (lines 3 and 4) we see other estimates for children of age 3.03, and in lines 5 and 6, we see the effects of SES in children of age 4.26.

These three values for age (1.80, 3.03 and 4.26) are selected by the function `simple_slopes()` by computing the mean (3.03), subtracting the standard deviation of the age variable from the mean ($3.03 - 1.23 = 1.80$), and by adding the standard deviation to the mean ($3.03 + 1.23 = 4.26$, when rounding to two decimals). This can be verified by getting some descriptive statistics for the **age** variable.

```
data_words %>% summarise(mean = mean(age),
                         sd = sd(age))
```

```
## # A tibble: 1 x 2
##    mean    sd
##   <dbl> <dbl>
## 1  3.03  1.23
```

Because with a numeric variable, you do not have categories, but rather an infinite number of possible values, we simply have to choose a couple of values for which we study the effect of SES. You could say that we look at the effect of SES, given three values for age: "relatively young" (1.80 years), "average" (3.03 years), and "relatively old" (4.26).

10.12 Why not simply partition the data in subsets?

You may wonder, wouldn't it be easier to divide the data into three parts? For instance, in the previous section, we could divide the data into low, average and high SES children, and then model the effect of age on words for each separate group? Or in the section before that: divide the data into three data sets, one for each level of education and then look for the difference between females and males for every data set separately?

Well, you may be right that it would perhaps be simpler, but it would lead to slightly different results. If we would cut up the data into three pieces, and do three separate analyses, we would have fewer data points to base our conclusions on. In each separate analysis, we would have three model parameters: one for the intercept, one for the slope, but also one for the variance of the residuals. This residual variance is used to compute the standard error of the intercept and slope parameters. Therefore, if you cut up the data and do the analysis on each part separately, you would end up with three different values for this residual variance. If on the other hand, you carry out the analysis once using all the data, and computing the contrasts, there is only one estimate of the residual variance. And not only that: also the degrees of freedom is larger. It's the total sample size minus the number of parameters in the regression table. Together with the calculated single residual variance, this affects the size of the standard errors, and also the critical t-value for the confidence intervals. In general you can expect more statistical power when using all the data at once, then when doing separate analyses.

In general it is better to perform one analysis on your data, and to answer you research questions using the results of that single analysis, rather than cutting up the data and do separate analyses. Unless you have doubts that the residual variance is the same for different groups (homoscedasticity, see Chapter 7), you should use only one linear model.

10.13 Take-away points

- When analysing categorical variables in a linear model, the categorical variable is represented by a new set of numeric variables.

- These new variables can be dummy variables (default) but can also be other types of numeric variables.
- How these numeric variables relate to the original categorical variable is summarised in a coding matrix **S**.
- The coding matrix **S** determines what values are printed in the regression table. These values are actually contrasts.
- Contrasts are weighted sums of group means. The contrasts are represented in a contrast matrix **L**.
- Contrasts are meant to address specific research questions.
- Matrix **L** is the inverse of **S**, and **S** is the inverse of **L**.
- If your model involves moderation, you can calculate "simple effects" (or "simple slopes"): the effects of one independent variable given particular values of a second independent variable.

Key concepts

- Contrasts
- Weighted sum
- Linear combination
- Contrast matrix **L**
- Coding scheme matrix **S**
- Dummy coding/treatment coding
- Effect coding
- Helmert contrasts
- Successive differences contrast coding
- Simple effects (simple slopes)

Chapter 11

Post-hoc comparisons

11.1 Introduction

Analysis of variance, as we have seen, can be used to test null-hypotheses about *overall* effects of certain factors (categorical variable) or combinations of factors (moderation). This is done with F-test statistics, with degrees of freedom that depend on the number of (combinations of) categories. The regression table, with t-tests in the output, can be used to compute specific contrasts, either the standard contrasts based on dummy coding, or contrasts based on alternative coding schemes.

In the previous chapter, all alternatives for specifying contrasts have been discussed in the context of specific research questions. It is important to make a distinction here between research questions that are posed *before* the data gathering and analysis, and research questions that pop up *during* the data analysis. Research questions of the first kind we call *a priori* ("at the outset") questions, and questions of the second kind we call *post hoc* ("after the fact") questions.

We've seen that the whole data analysis approach for inference is based on sampling distributions, for instance the sampling distribution of the t-statistic given that a population value equals 0. We then look at what t-value can be deemed large enough to reject the null-hypothesis (or to construct a confidence interval). Such a critical t-value is chosen in a way that if the null-hypothesis is true, it only happens in a low percentage of cases (α) that we find a t-value more extreme than this critical value. This helps us to reject the null-hypothesis: we see something extreme that can't be explained very well by the null-hypothesis.

However, if we look at a linear regression output table, we often see many t-values: one for the intercept, several for slope coefficients, and if the model includes moderation, we also may see several t-values for interaction effects.

Every single t-value is based on a hypothesis that the null-hypothesis is true, that is, that the actual parameter value (or contrast) is 0 in the population. For every single t-test, we therefore know that if we would draw many many samples, in only α% of the samples, we would find a t-value more extreme than the critical value (given that the null-hypothesis is true). But if we have for instance 6 different t-values in our output, how large is the probability that any of these 6 different t-values is more extreme than than the critical value?

Let's use a very simple example. Let's assume we have a relatively large data set, so that the t-distibution is very close to the normal distribution. When we assume we use two-sided testing with an α of 5%, we know that the critical values for the null-hypothesis are -1.96 and 1.96. Imagine we have a dependent variable Y and a categorical independent variable X that consists of two levels, A and B. If we run a standard linear model on those variables Y and X, using dummy coding, we will see two parameters in the regression table: one for reference level A (labelled "(Intercept)"), and one coefficient for the difference between level B and A (labelled "XB"). Suppose that in reality, the population values for these parameters are both 0. That would mean that the two group means are equal to 0. When we do the research many times, drawing a large sample and repeat taking new samples 100 times, how many times will the intercept have a t-value more extreme than ± 1.96? Well, by definition, that frequency would be about 5, because we know that for the t-distribution, 5% of this distribution has values more extreme than ± 1.96. Thus, if the intercept is truly 0 in the population, we will see a significant t-value in 5% of the samples.

The same is true for the second parameter "XB": if this value is 0 in the population, then we will see a significant t-value for this parameter in the output in 5% of the samples.

Both of these events would be Type I errors: the kind of error that you make when you reject the null-hypothesis while it is actually true (see Chapter 2).

For any given sample, there can be either no Type I error, or there is one Type I error, of there are two Type I errors. Now, if the probability for a Type I error is 5% for a significant value for "(Intercept)", and the probability is 5% for a type I error for "XB", what then is the probability for *at least one Type I error*?

This is a question for probability theory. If we assume that the Type I errors for the intercept and the slope are independent, we can use the binomial distribution (Ch. 3) and know that the probability of finding no Type I errors equals

$$P(errors = 0|\alpha = 0.05) = \binom{2}{0} \times 0.05^0 \times (1 - 0.05)^2 = 0.95^2 = 0.9025$$

[Note: In case you skipped Chapter 3, $\binom{4}{2}$ is pronounced as "4 choose 2" and it stands for the number of combinations of two elements that you can have when you have four elements. For instance, if you have 4 letters A, B, C and D, then there are 6 possible pairs: AB, AC, AD, BC, BD, and CD. The general case of

$\binom{a}{b}$ can be computed by R using `choose (a, b)`. $\binom{2}{0}$ is defined as 1 (there is only one way in which neither of the 2 tests results in a Type I error.)]

Therefore, since probabilities sum to 1, we know that the probability of *at least* one Type I error equals $1 - 0.9025 = 0.0975$.

We see that when we look at two t-tests in one analysis, the probability of a Type I error is no longer 5%, but almost twice that: 9.75%. This is under the assumption that the t-tests are independent of each other, which is often not the case. We will discuss what we mean with independent later. For now it suffices to know that the more null-hypotheses you test, the higher the risk of a Type I error.

For instance, if you have a categorical variable X with not two, but ten different groups, your regression output table will contain ten null-hypothesis tests: one for the intercept (reference category) and nine tests for the difference between the remaining groups and the reference group. In that case, the probability of at least one Type I error, if you perform each test with an α of 5%, will be

$$1 - P(errors = 0 | \alpha = 0.05) = 1 - \binom{10}{0} \times 0.05^0 \times 0.95^{10} = 1 - 0.95^{10} = 0.4013$$

And all this is only in the situation that you stick to the default output of a regression. Imagine that you not only test the difference between each group and the reference group, but that you also make many other contrasts: difference between group 2 and group 9, etcetera. If we would look at each possible pair among these ten groups, there would be $\binom{10}{2} = 144$ contrasts and consequently 144 t-tests. The probability of a Type I error will then be very close to 1, almost certainty!

The problem is further complicated by the fact that tests for all these possible pairs cannot be *independent* of each other. This is easy to see: If Jim is 5 centimetres taller than Sophie, and Sophie is 4 centimetres taller than Wendy, we already know the difference in height between Jim and Sophie: $5 + 4 = 9$ centimetres. In other words, if you want to know the contrast between Jim and Wendy, you don't need a new analysis: the answer is already there in the other two contrasts. Such a dependence in the contrasts that you estimate can lead to an even higher probability of a Type I error.

In order to get some grip on this dependence problem, we can use the so-called *Bonferroni inequality*. In the context of null-hypothesis testing, this states that the probability of at least one Type I error is less than or equal to α times the number of contrasts J. This is called the upper bound.

$$P(errors > 0) = 1 - P(errors = 0) \leq J\alpha$$

This inequality is true whether two contrast are heavily dependent (as in the height example above) or only slightly, or not at all. For instance, if you have

two contrasts in your output (the intercept and the slope), the probability of at least one Type I error equals 0.0975, but only if we assume these two contrasts are independent. In contrasts, if the two contrasts are *dependent*, we can use the Bonferroni inequality to know that the probability of at least one Type I error is less than or equal to $0.05 \times 2 = 0.10$. Thus, if there is dependency we know that the probability of at least one Type I error is *at the most* 0.10 (it could be less bad).

Note that if $J\alpha > 1$, then the upper bound is set equal to 1.

This Bonferroni upperbound can help us to take control over the overall probability of making Type I errors. Here we make a distinction between the *test-wise* Type I error rate, α_{TW}, and the *family-wise* Type I error rate, α_{FW}. Here, α_{TW} is the probability of a Type I error used for one individual hypothesis test, and α_{FW} is the probability of at least one Type I error among all tests performed. If we have a series of null-hypothesis tests, and if we want to have an overall probability of at most 5% (i.e., $\alpha_{FW} = 0.05$), then we should set the level for any individual test α_{TW} at α_{FW}/J. Then we know that the probability of at least one error is 5% or less.

Note that what is true here for null-hypothesis testing is also true for the calculation of confidence intervals. Also note that we should only look at output for which we have research questions. Below we see an example of how to apply these principles.

Example

We use the `ChickWeight` data available in R. It is a data set on the weight of chicks over time, where the chicks are categorised into four different groups, each on a different feeding regime. Suppose we do a study on diet in chicks with one control group and three experimental groups. For each of these three experimental groups, we want to estimate the difference with the control condition (hence there are three research questions). We perform a regression analysis with dummy coding with the control condition (Diet 3) as the reference group to obtain these three contrasts. For the calculation of the confidence intervals, we want to have a family-wise Type I error rate of 0.05. That means that we need to have a test-wise Type I error rate of $0.05/3 = 0.0167$. We therefore need to compute $100 - 1.67 = 98.33\%$ confidence intervals and we do null-hypothesis testing where we reject the null-hypothesis if $p < 0.0167$. The R code would look something like the following:

```
ChickWeight <- ChickWeight %>%
  mutate(Diet = relevel(Diet, ref = "3"))
model <- ChickWeight %>%
  lm(weight ~ Diet, data =.)
model %>%
  tidy(conf.int = TRUE, conf.level = 0.9833)
```

```
## # 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)   143.        6.33    22.6   2.59e-81    128.     158.
## 2 Diet1         -40.3       7.87    -5.12  4.11e- 7    -59.2    -21.4
## 3 Diet2         -20.3       8.95    -2.27  2.35e- 2    -41.8      1.15
## 4 Diet4          -7.69      8.99    -0.855 3.93e- 1    -29.3     13.9
```

```
model$df
```

```
## [1] 574
```

In the output we see the three contrasts that we need to answer our research questions. We can then report:

> "We estimated the difference between each of the three experimental diet with the control diet. In order to control the family-wise Type I error rate and keep it below 5%, we used Bonferroni correction and chose a test-wise significance level of 0.0167 and computed 98.3% confidence intervals. The chicks on Diet 1 had a significantly lower weight than the chicks in the control conditions (Diet 3, $b = -40.3$, $SE = 7.87$, $t(574) = -5.12$, $p < .001$, 98.33% CI: -59.2, -21.4). The chicks on Diet 2 also had a lower weight than chicks on Diet 3, although the null-hypothesis could not be rejected ($b = -20.3$, $SE = 8.95$, $t(574) = -2.27$, $p = .024$, 98.33% CI: -41.8, 1.15). The chicks on Diet 4 also had a weight not significantly different from chicks on Diet 3, ($b = -7.69$, $SE = 8.99$, $t(574) = -0.855$, $p = .039$, 98.33% CI: -29.3, 13.9). "

Note that we do not report on the Intercept. Since we had no research question about the average weight of chicks on Diet 3, we ignore those results in the regression table, and divide the desired family-wise error rate by 3 (and not 4).

As we said, there are two kinds of research questions: a priori questions and post hoc questions. A priori questions are questions posed before the data collection. Often they are the whole reason why data were collected in the first place. Post hoc questions are questions posed during data analysis. When analysing data, some findings may strike you and they inspire you to do some more analyses. In order to explain the difference, let's think of two different scenarios for analysing the `ChickWeight` data.

In Scenario 1, researchers are interested in the relationship between the diet and the weight of chicks. They see that in different farms, chicks show different mean sizes, and they are also on different diets. The researchers suspect that the weight differences are induced by the different diets, but they are not sure,

because there are also many other differences between the farms (differences in climate, chicken breed and daily regime). In order to control for these other differences, the researchers pick one specific farm, they use one particular breed of chicken, and assign the chicks randomly to one of four diets. They reason that if they find differences between the four groups regarding weight, then diet is the factor responsible for those differences. Thus, their research question is: "Are there any differences in mean weight as a function of diet?" They run an ANOVA and find the following results:

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
ChickWeight %>%
  lm(weight ~ Diet, data =.,
     contrasts = list(Diet = contr.sum)) %>%
  Anova(type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: weight
##              Sum Sq  Df F value    Pr(>F)
## (Intercept) 8538737   1 1776.65 < 2.2e-16 ***
## Diet         155863   3   10.81 6.433e-07 ***
## Residuals   2758693 574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

They answer their (a priori) research question in the following way.

> "We tested the null-hypothesis of equal mean weights in the four diet groups at an alpha of 5%, and found that it could be rejected, $F(3, 574), p < .001$. We conclude that diet does have an influence on the mean weight in chicks."

When analysing the data more closely, they also look at the mean weights per group.

```
ChickWeight %>%
  group_by(Diet) %>%
  summarise(mean = mean(weight))
```

```
## # A tibble: 4 x 2
##   Diet   mean
##   <fct> <dbl>
## 1 3      143.
## 2 1      103.
## 3 2      123.
## 4 4      135.
```

They are struck by the relatively small difference in mean weight between Diet 4 and Diet 3. They are surprised because they know that one of them contains a lot more protein than the other. They are therefore curious to see whether the difference between Diet 4 and Diet 3 is actually significant. Moreover, they are very keen on finding out which Diets are different from each other, and which Diets are not different from each other. They decide to perform 6 additional t-tests: one for every possible pair of diets.

In this Scenario I, we see two kinds of research questions: the initial a priori question was whether Diet affects weight, and they answer this question with one F-test. The second question only arose during the analysis of the data. They look at the means, see some kind of pattern and want to know whether all means are different from each other. This follow-up question that arises during data analysis is a post hoc question.

Let's look at Scenario II. A group of researchers wonder whether they can get chicks to grow more quickly using alternative diets. There is one diet, Diet 3, that is used in most farms across the world. They browse through the scientific literature and find three alternative diets. These alternative diets each have a special ingredient that makes the researchers suspect that they might lead to weight gain in chicks. The objective of their research is to estimate the differences between these three alternative diets and the standard diet, Diet 3. They use one farm and one breed of chicken and assign chicks randomly to one of the four diets. They perform the regression analysis with the dummy coding and reference group Diet 3 as shown above, and find that the differences between the three experimental diets and the standard diet are all negative: they show *slower* growth rather than faster growth. They report the estimates, the standard errors and the confidence intervals.

In general we can say that a priori questions can be answered with regular alphas and confidence intervals. For instance, if you state that your Type I error rate

α is set at 0.05, then you can use this $\alpha = 0.05$ also for all the individual tests that you perform confidence intervals that you calculate. However, for post hoc questions, where your questions are guided by the results that you see, you should correct the test-wise error rate α_{TW} in such a way that you control the family-wise error rate α_{FW}.

Returning to the two scenarios, let's look at the question whether Diet 4 differs from Diet 3. In Scenario I, this is a post hoc question, where in total you have 6 post hoc questions. You should therefore do the hypothesis test with an alpha of $0.05/6 = 0.0083$, and/or compute a 99.17% confidence interval. In contrast, in Scenario II, the same question about Diets 4 and 3 is an a priori question, and can therefore be answered with an $\alpha = 5\%$ and/or a 95% confidence interval.

Summarising, for post hoc questions you adjust your test-wise type I error rate, whereas you do not for a priori questions. The reason for this different treatment has to do with the dependency in contrasts that we talked about earlier. It also has to do with the fact that you only have a limited number of model degrees of freedom. In the example of the `ChickWeight` data, we have four groups, hence we can estimate only four contrasts. In the regression analysis with dummy coding, we see one contrast for the intercept and then three contrasts between the experimental groups and the reference group. Also if we use Helmert contrasts, we will only obtain four estimates in the output. This has to do with the dependency between contrasts: if you know that group A has a mean of 5, group B differs from group A by +2 and group C differs from group A by +3, you don't need to estimate the difference between B and C any more, because you know that based on these numbers, the difference can only be +1. In other words, the contrast C to B totally depends on the contrast A versus B and A versus C. The next section discusses the dependency problem in more detail.

11.2 Independent (orthogonal) contrasts

Whether two contrasts are dependent is easily determined. Suppose we have J independent samples (groups), each containing values from a population of normally distributed values (assumption of normality). Each group is assumed to come from a population with the same variance σ_e^2 (assumption of equal variance). For the moment also assume that the J groups have an equal sample size n. Any group j will have a mean $\bar{Y}_j$. Now imagine two contrasts among these means. The first contrast, $L1$, has the weights c_{1j}, and the second contrasts, $L2$, has the weights c_{2j}. Then we know that contrasts $L1$ and $L2$ are independent if

$$\sum_{j=1}^{J} c_{1j}c_{2j} = 0$$

Thus, if you have J independent samples (groups), each of size n, one can decide if two contrasts are dependent by checking if the products of the weights sum to zero:

$$c_{11}c_{21} + c_{12}c_{22} + \cdots + c_{1J}c_{2J} = 0$$

Another word for independent is *orthogonal*. Two contrasts are said to be orthogonal if the two contrasts are independent. Let's look at some examples for a situation of four groups: one set of dependent contrasts and a set of orthogonal contrasts. For the first example, we look at default dummy coding. For contrast $L1$, we estimate the mean of group 1. Hence

$$L1 = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

Let contrast $L2$ be the contrast between group 2 and group 1:

$$L2 = \begin{bmatrix} -1 & 1 & 0 & 0 \end{bmatrix}$$

If we calculate the products of the weights, we get:

$$\sum_j c_{1j}c_{2j} = 1 \times -1 + 0 \times 1 + 0 \times 0 + 0 \times 0 = -1$$

So we see that when we use dummy coding, the contrasts are not independent (not orthogonal).

For the second example, we look at Helmert contrasts. Helmert contrasts are independent (orthogonal). The Helmert contrast matrix for four groups looks like

$$L = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ -1 & 1 & 0 & 0 \\ -\frac{1}{2} & -\frac{1}{2} & 1 & 0 \\ -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & 1 \end{bmatrix}$$

For the first two contrasts, we see that the product of the weights equals zero:

$$\sum_j c_{1j}c_{2j} = \frac{1}{4} \times -1 + \frac{1}{4} \times 1 + \frac{1}{4} \times 0 + \frac{1}{4} \times 0 = 0$$

Check for yourself and find that all four Helmert contrasts are independent of each other.

11.3 The number of independent contrasts is limited

Earlier we saw that there is only so much information you can gain from a data set. Once you have certain information, asking further questions leads to answers that depend on the answers already available.

This dependency has a bearing on the number of orthogonal comparisons that can be made with J group means. Given J independent sample means, there can be, apart from the grand mean, no more than $J - 1$ comparisons, without them being dependent on each other. This means that if you have J completely independent contrasts for J group means, it is impossible to find one more comparison which is also orthogonal to the first J ones.

This implies that if you ask more questions (i.e., ask for more contrasts) you should tread carefully. If you ask more questions, the answers to your questions will not be independent of each other (you are to some extent asking the same thing twice).

As an example, earlier we saw that if you know that group B differs from group A by +2 and group C differs from group A by -3, you don't need to estimate the difference between B and C any more, because you know that based on these numbers, the difference can only be 5. In other words, the contrast C to B totally depends on the contrasts A versus B and A versus C. You can also see this in the contrast matrix for groups A, B and C below:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix}$$

The last contrast is dependent both on the third and the second contrast. Contrast $L4$ can be calculated as $L3 - L2$ by doing the calculation element-wise:

$$\begin{bmatrix} -1 & 0 & 1 \end{bmatrix} - \begin{bmatrix} -1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 & 1 \end{bmatrix}$$

In other words, $L4$ is a linear combination (weighted sum) of $L2$ and $L3$: $L4 = 1 \times L3 - 1 \times L2$. Statistically therefore, contrast $L4$ is completely redundant given the contrasts $L2$ and $L3$: it doesn't provide any extra information.

It should however be clear that if you have a research question that can be answered with contrast $L4$, it is perfectly valid to make this contrast. However, you should realise that the number of independent research questions is limited. It is a wise idea to limit the number of research questions to the number of contrasts you can make: apart from the grand mean, you should make no more

than $J - 1$ comparisons (your regression table should have no more than J parameters).

These contrasts that you specify belong to the a priori research question. Good research has a limited number of precisely worded research questions that should be answerable by a limited number of contrasts, usually 2 or 3, sometimes only 1. These can be answered by using the regular significance level. In social and behavioural sciences, oftentimes α for each individual test or confidence interval equals 5%. However, the follow-up questions that arise only after the initial data analysis (computing means, plotting the data, etc.), should however be corrected to control the overall Type I error rate.

11.4 Fishing expeditions

Research and data analysis can sometimes be viewed as a fishing expedition. Imagine you fish the high seas for herring. Given your experience and what colleagues tell you (you browse the scientific literature, so to speak), you choose a specific location where you expect a lot of herring. By choosing this location, you maximise the probability of finding herring. This is analogous to the setting up of a data collection scheme where you maximise the probability of finding a statistically significant effect, or you maximise the precision of your estimates; in other words, you maximise statistical power (see Chapter 5). However, while fishing in that particular spot for herring, irrespective of whether you actually find herring, you find a lot of other fish and seafood. This is all coincidence, as you never *planned* to find these kinds of fish and seafood in your nets. The fact that you find a crab in your nets, might seem very interesting, but it should never be reported as if you were looking for that crab. You would have equally regarded it interesting if you had found a lobster, or a seahorse, or a baseball hat. You have to realise that it is pure random sampling error: you hang out your nets, and just pick up what's there by chance. In research it works the same way: if you do a lot of statistical tests, or compute a large number of confidence intervals, you're bound to find something that *seems* interesting, but is actually merely random noise due to sampling error. If the family-wise error rate is large, say 60%, then you cannot tell your family and friends ashore that the base-ball hat you found is very fascinating. Similarly, in research you have to control the number of Type I errors by adjusting the test-wise error rate in such a way that the family-wise error rate is low.

11.5 Several ways to define your post hoc questions

One question that often arises when we find that a categorical variable has an effect in an ANOVA, is to ask where this overall significant effect is coming

from. For instance, we find that the four diets result in different mean weights in the chicks. This was demonstrated with an F-test at an α of 5%. A follow-up question might then be, what diets are different from each other. You might then set up contrasts for all $\binom{4}{2} = 6$ possible pairs of the four diets.

Alternatively, you may specify your post hoc questions as simple or more complex contrasts in the same way as for your a priori questions, but now with no limit on how many. For instance, you may ask what alternative diets are significantly different from the standard diet (Diet 3). The number of comparisons is then limited to 3. Additionally, you might ask whether the alternative diets combined (grand mean of diets 1, 2 and 4) are significantly different from Diet 3.

Be aware, however, that the more comparisons you make, the more severe the correction must be to control the family-wise Type I error rate.

The analyses for the questions that you answer by setting up the entire data collection, and that are thus planned before the data collection (a priori), can be called *confirmatory* analyses. You would like to *confirm* the workings of an intervention, or you want to precisely estimate the size of a certain effect. Preferably, the questions that you have are statistically independent of each other, that is, the contrasts that you compute should preferably be orthogonal (independent).

In contrast, the analyses that you do for questions that arise while analysing the data (post hoc) are called *exploratory* analyses. You *explore* the data for any interesting patterns. Usually, while exploring data, a couple of questions are not statistically independent. Any interesting findings in these exploratory analyses could then be followed up by confirmatory analyses using a new data collection scheme, purposely set up to confirm the earlier findings. It is important to do that with a new or different sample, since the finding could have resulted from mere sampling error (i.e., a Type I error).

Also be aware of the immoral practice of p-hacking. P-hacking, sometimes referred to as *selective reporting*, is defining your research questions and setting up your analysis (contrasts) in such a way that you have as many significant results as possible. With p-hacking one presents their research in such a way that they find all these interesting results, ignoring the fact that they made a selection of the results based on what they saw in the data (post-hoc). For instance, their research was set up to find evidence for the workings of medicine A on the alleviation of migraine. Their study included a questionnaire on all sorts of other complaints and daily activities, for the sake of completeness. When analysing the results, they might find that the contrast between medicine A with placebo is not significant for migraine. But when exploring the data further, they find that medicine A was significantly better with regards to bloodpressure and the number of walks in the park. A p-hacker would write up the research as a study of the workings of medicine A on bloodpressure and walks in the park. This form of p-hacking is called *cherry-picking*: only reporting statistically

significant findings and pretending you never set out to find the other things and not reporting them. Another p-hacking example would be to make a clever selection of the migraine data after which the effect becomes significant, for instance by filtering out the males in the study. Thus, p-hacking is the practice of trying to select the data or choose the method of analysis in such a way that the p-values in the report are as small as possible. The research questions are then changed from exploratory to confirmatory, without informing the reader.

11.6 Controlling the family-wise Type I error rate

There are several strategies that control the number of Type I errors. One is the *Bonferroni method*, where we adjust the test-wise error rate by dividing the family-wise error rate by the number of comparisons, $\alpha_{TW} = \alpha_{FW}/J$. This method is pretty conservative, in that the α_{TW} becomes low with already a couple of comparisons, so that the statistical power to spot differences that also exist in the population becomes very low. The upside is that this method is easy to understand and perform. Alternative ways of addressing the problem are *Scheffé's procedure*, and the *Tukey HSD* method. Of these two, Scheffé's procedure is also relatively conservative (i.e., little statistical power). The logic of the Tukey HSD method is based on the probability that a difference between two group means is more than a critical value, by chance alone. This critical value is called *Honestly Significant Difference (HSD)*. We fix the probability of finding such a difference (or more) between the group means under the null-hypothesis at α_{FW}. The details of the actual method will not be discussed here. Interested readers may refer to Wikipedia and references therein.

11.7 Post-hoc analysis in R

In general, post hoc contrasts can be done in the same way as in the previous chapter: specifying the contrasts in an $\mathbf{L}$ matrix, taking the inverse and assigning the matrix to the variable in your model. Here, you are therefore limited to the number of levels of a factor: you can only have $J - 1$ new variables, apart from the intercept of 1s. You can then adjust α yourself using Bonferroni. For instance if you want to have a family-wise type I error rate of 0.05, and you look at two post-hoc contrasts, you can declare a contrast significant if the corresponding p-value is less than 0.025.

There are also other options in R to get post hoc contrasts, where you can ask for as many comparisons as you want.

There are two ways in which you can control the overall Type I error rate: either by using an adjusted α yourself (as above), or adjusting the p-value. For now

we assume you generally want to test at an α of 5%. But of course this can be any value.

In the first approach, you run the model and the contrast just as you would normally do. If the output contains answers to post hoc questions, you do not use $\alpha = 0.05$, but you use 0.05 divided by the number of tests that you inspect: $\alpha_{TW} = \alpha_{FW}/k$, with k being the number of tests you do.

For instance, if the output for a linear model with a factor with four levels contains the comparison on groups 1 and 2, and it applies to an a priori question, you simply report the statistics and concludes significance if the $p < .05$.

If the contrast pertains to a post hoc question and you compare all six possible pairs, you report the usual statistics and conclude significance if the $p < \frac{0.05}{6}$.

In the second approach, you can change the p-value itself: you multiply the plotted value by the number of comparisons and declare a difference to be significant if the corresponding *adjusted* p-value is less than 0.05. As an example, suppose you make six comparisons. Then you multiply the usual p-values by a factor 6: $p_{adj} = 6p$. Thus, if you see a p-value of 0.04, you compute p_{adj} to be 0.24 and conclude that the contrast is not significant. This is often done in R: the output yields *adjusted* p-values. Care should be taken with the confidence intervals: make sure that you know whether these are adjusted 95% confidence intervals or not. If not, then you should compute your own. Note that when you use the adjusted p-values, you should no longer adjust the α. Thus, an adjusted p-value of 0.24 is not significant, because $p_{adj} > .05$.

In this section we will see how to perform post hoc comparisons in two situations: either with only one factor in your model, or when you have two factors in your model.

11.7.1 ANOVA with only one factor

We give an example of an ANOVA post hoc analysis with only one factor, using the data from the four diets. We first run an ANOVA to answer the primary research question whether diet has any effect on weight gain in chicks.

```
ChickWeight %>%
  lm(weight ~ Diet, data = .,
     contrasts = list(Diet = contr.sum)) %>%
  Anova(type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: weight
##              Sum Sq  Df F value    Pr(>F)
## (Intercept) 8538737   1 1776.65 < 2.2e-16 ***
```

```
## Diet         155863   3   10.81 6.433e-07 ***
## Residuals   2758693 574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Seeing these results, noting that there is indeed a significant effect of diet, a secondary question pops up: "Which pairs of two diets show significant differences?" We answer that by doing a post hoc analysis, where we study each pair of diets, and control Type I error rate using the Bonferroni method. We can do that in the following way:

```
# Check your sample means
ChickWeight %>%
  group_by(Diet) %>%
  summarise(mean = mean(weight))
```

```
## # A tibble: 4 x 2
##   Diet   mean
##   <fct> <dbl>
## 1 3      143.
## 2 1      103.
## 3 2      123.
## 4 4      135.
```

```
pairwise.t.test(ChickWeight$weight, # dependent variable
                ChickWeight$Diet,  # independent variable
                p.adjust.method = 'bonferroni') # adjustment method
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  ChickWeight$weight and ChickWeight$Diet
##
##   3       1       2
## 1 2.5e-06 -       -
## 2 0.14077 0.06838 -
## 4 1.00000 0.00026 0.95977
##
## P value adjustment method: bonferroni
```

In the output we see six Bonferroni adjusted p-values for all six possible pairs. The column and row numbers refer to the levels of the Diet factor: Diet 3, Diet 1 and Diet 2 in the three columns, and Diet 1, Diet 2 and Diet 4 in the three rows. We see that all p-values are non significant ($p > .05$), except for two

comparisons: the difference between Diet 3 and Diet 1 is significant, $p < .001$. as well as the difference between Diet 4 and Diet 1, $p < .001$.

> "An analysis of variance showed that the mean weight was signficantly different for the four diets, $F(3, 574) = 10.8, p < .001$. We performed post hoc pair-wise comparisons, for all six possible pairs of diets. A family-wise Type I error rate of 0.05 was used, with Bonferroni correction. The difference between Diet 1 and Diet 3 was significant, and the difference between Diets 4 and 1 was significant. All other differences were not signficantly different from 0. "

11.7.2 ANOVA with two factors and moderation

In the previous subsection we did pair-wise comparisons in a one-way ANOVA (i.e., ANOVA with only one factor). In the previous chapter we also discussed how to set up contrasts in the situation of two factors that are modelled with interaction effects (Ch. 10). Let's return to that example.

```
data("jobsatisfaction", package = "datarium")

# Run linear model with interaction effect and compute simple slopes
sslopes <- jobsatisfaction %>%
  lm(score ~ gender + education_level + gender:education_level,
     data = .) %>%
  simple_slopes(confint = TRUE, ci.width = 0.95)
sslopes
```

```
##   gender     education_level Test Estimate Std. Error    2.5%   97.5% t value
## 1   male    sstest (college)        0.7967     0.2593  0.2764  1.3170  3.0726
## 2   male sstest (university)        3.8653     0.2527  3.3582  4.3724 15.2950
## 3 female    sstest (college)        0.7220     0.2460  0.2284  1.2156  2.9352
## 4 female sstest (university)        2.6650     0.2460  2.1714  3.1586 10.8343
## 5 sstest              school        0.3143     0.2527 -0.1928  0.8214  1.2438
## 6 sstest             college        0.2397     0.2527 -0.2674  0.7468  0.9484
## 7 sstest          university       -0.8860     0.2460 -1.3796 -0.3924 -3.6020
##   df  Pr(>|t|) Sig.
## 1 52 0.0033741   **
## 2 52 < 2.2e-16  ***
## 3 52 0.0049525   **
## 4 52 6.074e-15  ***
## 5 52 0.2191473
## 6 52 0.3473349
## 7 52 0.0007058  ***
```

In the example, we were only interested in the gender effect for each of the education levels. That means only the last three lines are relevant.

In the `simple_slopes()` code we used the argument `ci.width = 0.95`. That was because we hadn't discussed post hoc analysis yet, nor adjustment of p or α. In the case that we want to control the Type I error rate, we could use Bonferroni correction. We should then make the relevant p-values three times bigger than what they are uncorrected, because we are interested in three contrasts.

```
# Obtain adjusted p-values
p_uncorrected <- sslopes[5:7, "Pr(>|t|)"] # get rows 5:7 and column with p-values
p_uncorrected # check you got the right ones
```

```
## [1] 0.2191473106 0.3473348890 0.0007057855
```

```
p_corrected <- p_uncorrected*3 # Bonferroni correction, multiplying by number of contrasts
p_corrected  # for school, college and university, respectively
```

```
## [1] 0.657441932 1.042004667 0.002117357
```

Confidence intervals should also be changed. For that we need to adjust the Type I error rate α.

```
# Correct Type I error rate
alpha <- 0.05 / 3 # adjust alpha
CI <- 1 - alpha # adjust confidence interval
CI
```

```
## [1] 0.9833333
```

```
sslopes <- jobsatisfaction %>%
  lm(score ~ gender + education_level + gender:education_level, data = .) %>%
  simple_slopes(confint = TRUE, ci.width = CI) # adjusted conf. interval
sslopes
```

```
##   gender     education_level Test Estimate Std. Error 0.833333333333336%
## 1   male    sstest (college)        0.7967     0.2593             0.1552
## 2   male sstest (university)        3.8653     0.2527             3.2401
## 3 female    sstest (college)        0.7220     0.2460             0.1135
## 4 female sstest (university)        2.6650     0.2460             2.0565
## 5 sstest              school        0.3143     0.2527            -0.3109
## 6 sstest             college        0.2397     0.2527            -0.3855
## 7 sstest          university       -0.8860     0.2460            -1.4945
```

```
##   99.1666666666667% t value df  Pr(>|t|) Sig.
## 1            1.4381  3.0726 52 0.0033741   **
## 2            4.4905 15.2950 52 < 2.2e-16  ***
## 3            1.3305  2.9352 52 0.0049525   **
## 4            3.2735 10.8343 52 6.074e-15  ***
## 5            0.9395  1.2438 52 0.2191473
## 6            0.8649  0.9484 52 0.3473349
## 7           -0.2775 -3.6020 52 0.0007058  ***
```

In the output we see adjusted confidence intervals (note that the p-values are the original ones). We conclude for our three contrasts that in the "school" group the females score 0.31 (adjusted 95% CI: -0.31, 0.94) higher than boys, in the "college" group females score 0.24 (adjusted 95% CI: -0.39, 0.86) higher than boys, and in the "university" group 0.89 (adjusted 95% CI: -1.49, -0.28) *lower* than the boys.

The same logic of adjusting p-values and adjusting confidence intervals can be applied in situations with numeric independent variables.

11.8 Take-away points

- Your main research questions are generally very limited in number. If they can be translated into contrasts, we call these a priori contrasts.
- Your a priori contrasts can be answered using a pre-set level of significance, in the social and behavioural sciences this is often 5% for p-values and using 95% for confidence intervals. No adjustment necessary.
- This pre-set level of significance, α, should be set *before* looking at the data (if possible before the collection of the data).
- If you are looking at the data and want to answer specific research questions that arise because of what you see in the data (post hoc), you should use adjusted p-values and confidence intervals.
- There are several ways of adjusting the test-wise αs to obtain a reasonable family-wise α: Bonferroni is the simplest method but rather conservative (low statistical power). Many alternative methods exist, among them are Scheffé's procedure, and Tukey HSD method.

Key concepts

- A priori
- Post hoc
- Orthogonality/independence

- p-hacking
- Family-wise Type I error rate
- Test-wise Type I error rate
- Bonferroni correction

Chapter 12

Linear mixed modelling: introduction

12.1 Fixed effects and random effects

In the simplest form of linear modelling, we have one dependent numeric variable, one intercept and one or more independent variables. Let's look at a simple regression equation where dependent variable Y is predicted by an intercept b_0 and a linear effect of independent variable X with regression slope parameter b_1, and an error term e, where we assume that the error term e comes from a normal distribution.

$$Y = b_0 + b_1 X + e$$
$$e \sim N(0, \sigma^2)$$

Using this model, we know that for a person with a value of 5 for X, we expect Y to be equal to $b_0 + b_1 \times 5$. As another example, if Y is someone's IQ score, X is someone's brain size in cubic millilitres, b_0 is equal to 70, and b_1 is equal to 0.1, we expect on the basis of this model that a person with a brain size of 1500 cubic millimetres has an IQ score of $70 + 0.01 \times 1500$, which equals 85.

Now, for any model the predicted values usually are not the same as the observed values. If the model predicts on the basis of my brain size that my IQ is 140, my true IQ might be in fact 130. This discrepancy is termed the residual: the observed Y, minus the predicted Y, or $\widehat{Y}$, so in this case the residual is $Y - \widehat{Y} = 130 - 140 = -10$.

Here we have the model for the relationship between IQ and brain size.

$$\texttt{IQ} = 70 + 0.1 \times \texttt{Brain size} + e$$
$$e \sim N(0, \sigma^2)$$

Note that in this model, the values of 70 and 0.1 are *fixed*, that is, we use the same intercept and the same slope for everyone. You use these values for any person, for Henry, Jake, Liz, and Margaret. We therefore call these effects of intercept and slope *fixed effects*, as they are all the same for all units of analysis. In contrast, we call the e term, the random error term or the residual in the regression, a *random effect*. This is because the error term is *different for every unit*. We don't know the specific values of these random errors or residuals for every person, but nevertheless, we assume that they come from a distribution, in this case a normal distribution with mean 0 and an unknown variance. This unknown variance is given the symbol σ^2.

Here are a few more examples.

1. Suppose we study a number of schools, and for every school we use a simple linear regression equation to predict the number of students (dependent variable) on the basis of the number of teachers (independent variable). For every unit of analysis (in this case: school), the intercept and the regression slope are the same (fixed effects), but the residuals are different (random effect).

2. Suppose we study reaction times, and for every measure of reaction time – a trial – we use a simple linear regression equation to predict reaction time in milliseconds on the basis of the characteristics of the stimulus. Here, the unit of analysis is trial, and for every trial, the intercept and the regression slope are the same (fixed effects), but the residuals are different (random effect).

3. Suppose we study a number of students, and for every student we use a simple linear regression equation to predict the math test score on the basis of the number of hours of study the student puts in. Here, the unit of analysis is student, and for every student, the intercept and the regression slope are the same (fixed effects), but the residuals are different (random effect).

Let's focus for now on the last example. What happens when we have a lot of data on students, but the students come from different schools? Suppose we want to predict average grade for every student, on the basis of the number of hours of study the student puts in. We again could use a simple linear regression equation.

$$Y = b_0 + b_1\texttt{hourswork} + e$$
$$e \sim N(0, \sigma^2)$$

That would be fine if all schools would be all very similar. But suppose that some schools have a lot of high scoring students, and some schools have a lot of low scoring students? Then school itself would also be a very important predictor, apart from the number of hours of study. One could say that the data are *clustered*: math test scores coming from the same school are more similar than math test scores coming from different schools. When we do not take this into account, the residuals will not show independence (see Chapter 7 on the assumptions of linear models).

One thing we could therefore do to remedy this is to include school as a categorical predictor. We would then have to code this school variable into a number of dummy variables. The first dummy variable called `school1` would indicate whether students are in the first school (`school1 = 1`) or not (`school1 = 0`). The second dummy variable `school2` would indicate whether students are in the second school (`school2 = 1`) or not (`school2 = 0`), etcetera. You can then add these dummy variables to the regression equation like this:

$$\begin{aligned} Y &= b_0 + b_1\texttt{hourswork} + b_2\texttt{school1} + b_3\texttt{school2} + b_4\texttt{school3} + \cdots + e \\ e &\sim N(0, \sigma^2) \end{aligned}$$

In the output we would find a large number of effects, one for each dummy variable. For example, if the students came from 100 different schools, you would get 99 fixed effects for the 99 dummy variables. However, one could wonder whether this is very useful. As stated earlier, fixed effects are called fixed because they are the same for every unit of research, in this case every student. But working with 99 dummy variables, where students mostly score 0, this seems very much over the top. In fact, we're not even interested in these 99 effects. We're interested in the relationship between test score and hours of work, meanwhile taking into account that there are test score differences across schools. The dummy variables are only there to account for differences across schools; the prediction for one school is a little bit higher or lower than for another school, depending on how well students generally perform in each school.

We could therefore try an alternative model, where we treat the school effect as *random*: we assume that every school has a different average test score, and that these averages are normally distributed. We call these average test score deviations *school effects*:

$$\begin{aligned} Y &= b_0 + b_1\texttt{hourswork} + \texttt{schooleffect} + e \\ \texttt{schooleffect} &\sim N(0, \sigma_s^2) \\ e &\sim N(0, \sigma_e^2) \end{aligned}$$

So in this equation, the intercept is fixed, that is, the intercept is the same for all observed test scores. The regression coefficient b_1 for the effect of hours of

work is also fixed. But the schooleffect is random, since it is different for every school. The residual e is also random, being different for every student. It could also be written like this:

$$\begin{aligned} Y &= (b_0 + \texttt{schooleffect}) + b_1\texttt{hourswork} + e \quad (12.1) \\ \texttt{schooleffect} &\sim N(0, \sigma_s^2) \\ e &\sim N(0, \sigma_e^2) \end{aligned}$$

This representation emphasises that for every school, the intercept is a little bit different: for school A the intercept might be $b_0 + 2$, and for school B the intercept might be $b_0 - 3$.

So, equation (12.2) states that every observed test score is

1. partly influenced by an intercept that is random, with a certain average b_0 and variance σ_s^2, that is dependent on which school students are in,
2. partly influenced by the number of hours of work, an effect that is the same no matter what school a student is in (fixed), and
3. partly influenced by unknown factors, indicated by a random residual e coming from a normal distribution with variance σ_e^2.

To put it more formally: test score Y_{ij}, that is, the test score from student j in school i, is the sum of an effect of the school $b_0 + \texttt{schooleffect}_i$ (the average test score in school i), plus an effect of hours of work, $b_1 \times \texttt{hourswork}$, and an unknown residual e_{ij} (a specific residual for the test score for student j in school i).

$$\begin{aligned} Y_{ij} &= b_0 + \texttt{schooleffect}_i + b_1\texttt{hourswork} + e_{ij} \\ \texttt{schooleffect}_i &\sim N(0, \sigma_s^2) \\ e_{ij} &\sim N(0, \sigma_e^2) \end{aligned}$$

So in addition to the assumption of residuals that have a normal distribution with mean 0 and variance σ_e^2, we also have an assumption that the school averages have a normal distribution, in this case with mean b_0 and variance σ_s^2.

Let's go back to the example of reaction times. Suppose in an experiment we measure reaction time in a large number of trials. We want to know whether the size of the stimulus (large or small) has an effect on reaction time. Let's also suppose that we carry out this experiment with 20 participants, where every participant is measured during 100 trials: 50 large stimuli and 50 small stimuli, in random order. Now probably, some participants show generally very fast

responses, and some participants show generally very slow responses. In other words, the average reaction time for the 100 trials may vary from participant to participant. This means that we can use participant as an important predictor of reaction times. To take this into account we can use the following linear equation:

$$\begin{aligned} Y_{ij} &= b_0 + \texttt{speed}_i + b_1\texttt{size} + e_{ij} \\ \texttt{speed}_i &\sim N(0, \sigma_s^2) \\ e_{ij} &\sim N(0, \sigma_e^2) \end{aligned}$$

where Y_{ij}, is the reaction time j from participant i, $(b_0 + \texttt{speed}_i)$ is a random intercept representing the average speed for each participant i (where b_0 is the overall average across all participants and $\texttt{speed}_i$ the random deviation for each and every participant), b_1 is the fixed effect of the size of the stimulus. Unknown residual e_{ij} is a specific residual for the reaction time for trial j of participant i.

The reason for introducing random effects is that when your observed data are clustered, for instance student scores clustered within schools, or trial response times are clustered within participants, you violate the assumption of independence: two reaction times from the same person are more similar than two reaction times from different persons. Two test scores from students from the same school may be more similar than two scores from students in different schools (see Chapter 7). When this is the case, when data are clustered, it is very important to take this into account. When the assumption of independence is violated, you are making wrong inference if you use an ordinary linear model, the so-called general linear model (GLM). With clustered data, it is therefore necessary to work with an extension of the general linear model or GLM, the *linear mixed model*. The above models for students' test scores across different schools and reaction times across different participants, are examples of *linear mixed models*. The term *mixed* comes from the fact that the models contain a mix of both fixed and random effects. GLMs only contain fixed effects, apart from the random residual.

If you have clustered data, you should take this clustering into account, either by using the grouping variable as a categorical predictor or by using a random factor in a linear mixed model. As a rule of thumb: if you have fewer than 10 groups, consider a fixed categorical factor; if you have 10 or more groups, consider a random factor. Two other rules you can follow are: Use a random factor if the assumption of normally distributed group differences is tenable. Use a fixed categorical factor if you are actually interested in the *size* of group differences.

Below, we will start with a very simple example of a linear mixed model, one that we use for a simple pre-post intervention design.

Table 12.1: Headache measurements in NY Times readers suffering from headaches.

patient	pre	post
001	55	45
002	63	50
003	66	56
004	50	37
005	63	50

12.2 Pre-post intervention designs

Imagine a study where we hope to show that aspirin helps reduce headache. We ask 100 patients to rate the severity of their headache before they use aspirin (on a scale from 1 - 100), and to rate the severity again 3 hours after taking 500 mg of aspirin. These patients are randomly selected among people who read the NY Times and suffer from regular headaches. So here we have clustered data: we have 100 patients, and for each patient we have two scores, one before (pre) and one after (post) the intervention of taking aspirin. Of course, overall headache severity levels tend to vary from person to person, so we might have to take into account that some patients have a higher average level of pain than other patients.

The data could be represented in different ways, but suppose we have the data matrix in Table 12.1 (showing only the first five patients). What we observe in that table is that the severity seems generally lower after the intervention than before the intervention. But you may also notice that the severity of the headache also varies across patients: some have generally high scores (for instance patient 003), and some have generally low scores (for example patient 001). Therefore, the headache scores seem to be clustered, violating the assumption of independence. We can quantify this clustering by computing a correlation between the pre-intervention scores and the post-intervention scores. We can also visualise this clustering by a scatter plot, see Figure 12.1. Here it appears that there is a strong positive correlation, indicating that the higher the pain score before the intervention, the higher the pain score after the intervention.

There is an alternative way of representing the same data. Let's look at the same data in a new format in Table 12.2. In Chapter 1 we saw that this representation is called long format.

By representing the data in long format we acknowledge that there is really only one dependent measure: headache severity. The other two variables indicate that this variable varies across both patients and time point (pre intervention and post intervention). There is therefore a variable `measure` that indicates

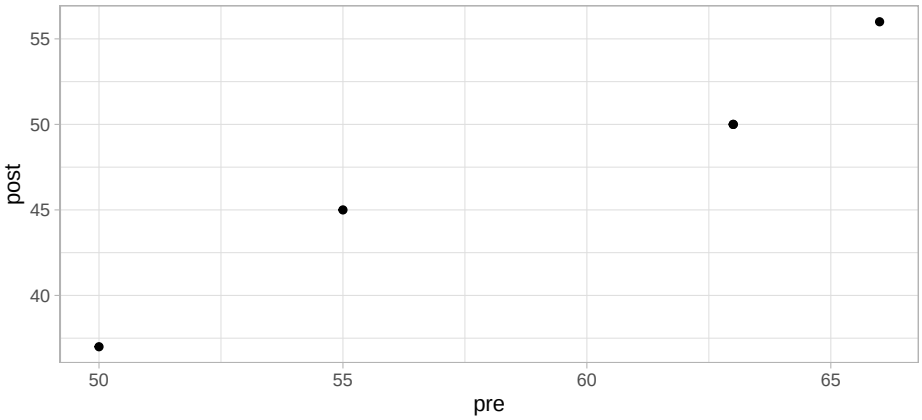


Figure 12.1: Scatterplot of pre and post headache levels.

Table 12.2: Headache severity measures in long format.

patient	measure	headache
1	1	55
1	2	45
2	1	63
2	2	50
3	1	66
3	2	56
4	1	50
4	2	37
5	1	63
5	2	50

whether the headache severity was measured pre intervention (`measure = 1`) or post intervention (`measure = 2`).

Here we might consider applying a simple linear regression model, using `headache` as the dependent variable and `measure` (1st or 2nd) as a categorical predictor. However, since we know that there is a correlation between the pre and post severity measures, we know that measures also systematically vary across patients: some score high on average and some score low on average. The assumption of independence is therefore not tenable. Thus we have to run a linear model, including not only an effect of `measure` but also an effect of `patient`. We then have to decide between fixed effects or random effects for these variables.

Let's first look at the variable `measure`. Since we are really interested in the effect of the intervention, that is, we want to know how large the effect of aspirin is, we use a fixed effect for the time effect (the variable `measure`). Moreover, the variable `measure` has only two levels, which is another reason to opt for a fixed effect.

Then for the patient effect, we see that we have 100 patients. Are we really interested by how much the average pain level in say patient 23 differs from the average pain level in say patient 45? No, not really. We only want to acknowledge that the individual differences exist, we want to take them into account in our model so that our inference regarding the confidence intervals and hypothesis testing is correct. We therefore prefer to assume random effects, assuming they are normally distributed. We therefore end up with a fixed effect for `measure` and a random effect for `patient` resulting in the following model:

$$\begin{aligned} Y_{ij} &= b_0 + \texttt{patient}_i + b_1\texttt{measure2} + e_{ij} \\ \texttt{patient}_i &\sim N(0, \sigma_p^2) \\ e_{ij} &\sim N(0, \sigma_e^2) \end{aligned}$$

where Y_{ij} is the jth headache severity score (first or second) for patient i, $(b_0 + \texttt{patient}_i)$ is the average amount of headache before aspirin that patient i deviates from the mean, `measure2` is a dummy variable, and b_1 is the effect of the intervention (by how much the severity changes from pre to post). We assume that the average pain level for each patient shows a normal distribution with average b_0 and variance σ_p^2. And of course we assume that the residuals show a normal distribution.

An analysis with this model can be done with the R package `lme4`. That package contains the function `lmer()` that works more or less the same as the `lm()` function that we saw earlier, but requires the addition of at least one random variable. Below, we run a linear mixed model, with dependent variable `headache`, a regular fixed effect for the categorical variable `measure`, and a random effect for the categorical variable `patient`.

```r
library(lme4)
out <- data %>%
  lmer(headache ~ measure + (1|patient), data = .)
summary(out)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: headache ~ measure + (1 | patient)
##    Data: .
##
## REML criterion at convergence: 1189.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.36239 -0.42430 -0.04973  0.49702  2.17096
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  patient  (Intercept) 27.144   5.210
##  Residual              8.278   2.877
## Number of obs: 200, groups:  patient, 100
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  59.6800     0.5952  100.28
## measure2    -10.3600     0.4069  -25.46
##
## Correlation of Fixed Effects:
##          (Intr)
## measure2 -0.342
```

In the output we see the results. We're mainly focused on the fixed effect of the intervention: does aspirin reduce headache? Where it says 'Fixed effects:' in the output, we see the linear model coefficients, with an intercept of around 59 and a negative effect of the intervention dummy variable `measure2`, around -10. We see that the dummy variable was coded 1 for the second measure (after taking aspirin). So, for our dependent variable `headache`, we see that the expected headache severity for the observations with a 0 for the dummy variable `measure2` (that is, measure 1, which is *before* taking aspirin), is equal to $59 - (10) \times 0 = 59$.

Similarly, we see that the expected headache severity for the observations with a 1 for the dummy variable `measure2` (that is, *after* taking aspirin), is equal to $59-(10)\times 1 = 49-10 = 49$. So, expected pain severity is 10 points lower after the intervention than before the intervention. Whether this difference is significant is indicated by a t-statistic. We see here that the average headache severity after

taking an aspirin is significantly different from the average headache severity before taking an aspirin, $t = -25.46$. However, note that we do not see a p-value. That's because the degrees of freedom are not clear, because we also have a random variable in the model. The determination of the degrees of freedom in a linear mixed model is a complicated matter, with different choices, the discussion of which is beyond the scope of this book.

However, we do know that for whatever the degrees of freedom really are, a t-statistic of -25.46 will always be in the far tail of the t-distribution, see Appendix B. So we have good reason to conclude that we have sufficient evidence to reject the null-hypothesis that headache levels before and after aspirin intake are the same.

If we do not want to test a null-hypothesis, but want to construct a confidence interval, we run into the same problem that we do not know what t-distribution to use. Therefore we do not know with what value the standard error of 0.40 should be multiplied to compute the margin of error. We could however use a coarse rule of thump and say that the critical t-value for a 95% confidence interval is more or less equal to 2 (see Appendix B). If we do that, then we get the confidence level for the effect of aspirin: between $-10.36 - 2 \times 0.4069 = -11.17$ and $-10.36 + 2 \times 0.4069 = -9.55$.

Taking into account the direction of the effect and the confidence interval for this effect, we might therefore carefully conclude that aspirin reduces headache in the population of NY Times readers with headache problems, where the reduction is around 10 points on a 1...100 scale (95% CI: 9.55 – 11.17).

Now let's look at the output regarding the random effect of `patient` more closely. The model assumed that the individual differences in headache severity in the 100 patients came from a normal distribution. How large are these individual differences actually? This can be gleaned from the 'Random effects:' part of the R output. The 'intercept' (i.e., the patient effect in our model) seems to vary with a variance of 27, which is equivalent to a standard deviation of $\sqrt{27}$ which is around 5.2. What does that mean exactly? Well let's look at the equation again and fill in the numbers:

$$\begin{aligned} Y_{ij} &= b_0 + \texttt{patient}_i + b_1\texttt{measure2} + e_{ij} \\ Y_{ij} &= 59 + \texttt{patient}_i - 10\texttt{measure2} + e_{ij} \\ patient_i &\sim N(0, 27) \\ e_{ij} &\sim N(0, 8) \end{aligned}$$

Since R used the the headache level before the intervention as the reference category, we conclude that the average pain level before taking aspirin is 59. However, not everybody's pain level before taking aspirin is 59: people show variance (variation). The pain level before aspirin varies with a variance of 27, which is equivalent to a standard deviation of around 5.2. Figure 12.2 shows

how much this variance actually is. It depicts a normal distribution with a mean of 59 and a standard deviation of 5.2.

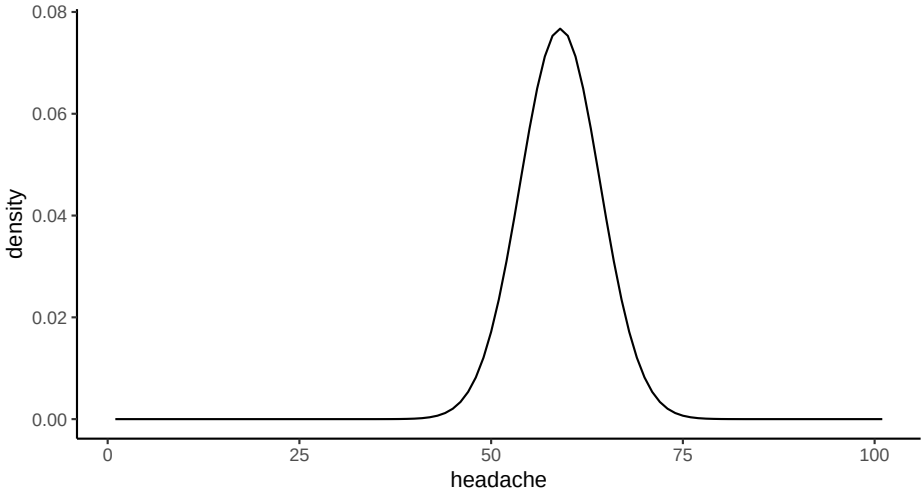


Figure 12.2: Distribution of headache scores before taking aspirin, according to the linear mixed model.

So *before* taking aspirin, most patients show headache levels roughly between 50 and 70. More specifically, if we would take the middle 95% by using plus or minus twice the standard deviation, we can estimate that 95% of the patients shows levels between $59 - 2 \times 5.2 = 48.6$ and $59 + 2 \times 5.2 = 69.4$.

Now let's look at the levels *after* taking aspirin. The average headache level is equal to $59 - 10 = 49$. So 95% of the patients shows headache levels between $49 - 2 \times 5.2 = 38.6$ and $49 + 2 \times 5.2 = 59.4$ before taking aspirin.

Together these results are visualised in Figure 12.3. In this plot you see there is variability in headache levels before taking aspirin, and there is variation in headache levels after taking aspirin. We also see that these distributions have the same spread (variance): in the model we assume that the variability in headache before aspirin is equal to the variability after aspirin (homoscedasticity). The distributions are equal, except for a horizontal shift: the distribution for headache after aspirin is the same as the distribution before aspirin, except for a shift to the left of about 10 points. This is of course the effect of aspirin in the model, the b_1 parameter in our model above.

The fact that the two distributions before and after aspirin show the same spread (variance) was an inherent assumption in our model: we only have one random effect for patient in our output with one variance (σ_p^2). If the assumption of equal variance (homoscedasticity) is not tenable, then one should consider other linear mixed models. But this is beyond the scope of this book. The assumption can

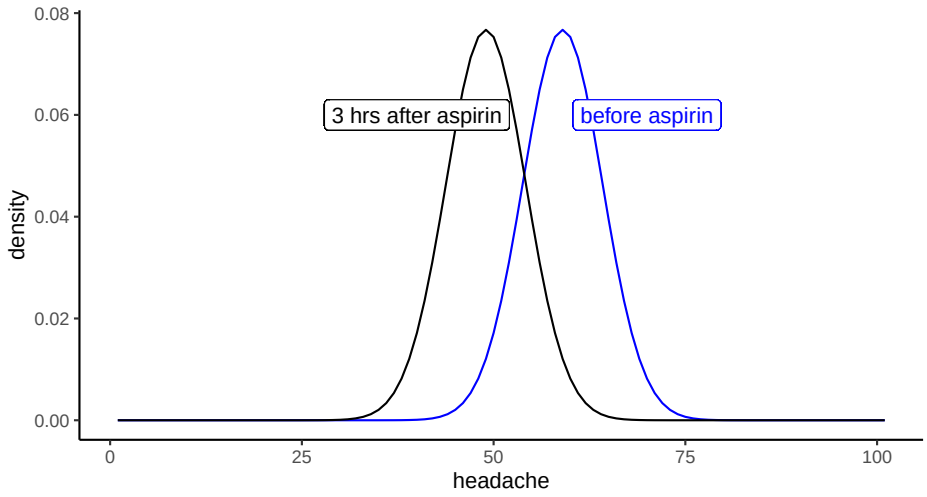
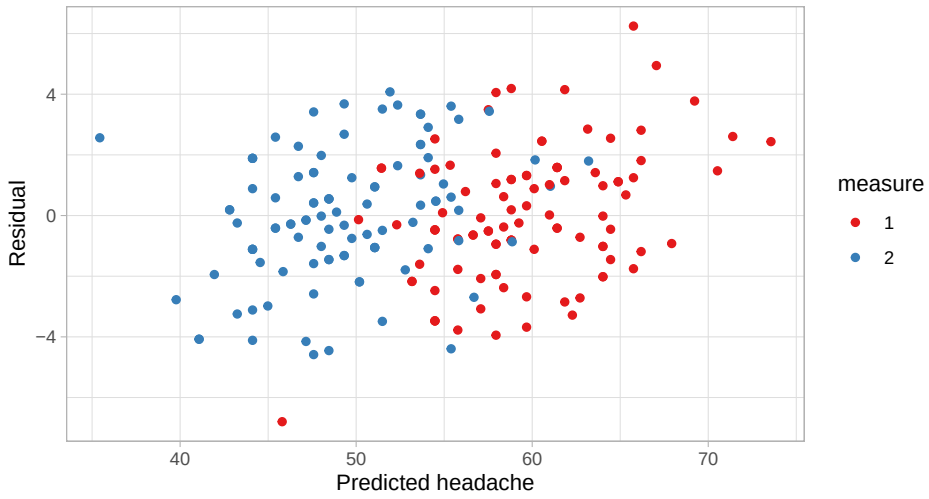


Figure 12.3: Distribution of headache scores before and after taking aspirin, according to the linear mixed model.

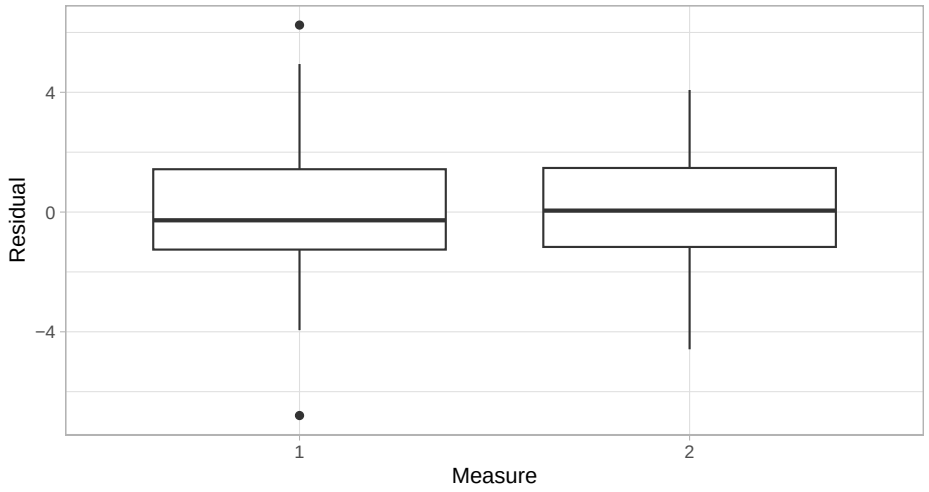
be checked by plotting the residuals, using different colours for residuals from before taking aspirin and for residuals from after taking aspirin.

```
library(modelr)
data %>%
  add_residuals(out) %>%
  add_predictions(out) %>%
  ggplot(aes(x = pred, y = resid, colour = measure)) +
  geom_point() +
  xlab("Predicted headache") +
  ylab("Residual") +
  scale_colour_brewer(palette = "Set1") + # use nice colours
  theme_light()
```



Alternatively one could create a box plot.

```
data %>%
  add_residuals(out) %>%
  ggplot(aes(x = measure, y = resid)) +
  geom_boxplot() +
  xlab("Measure") +
  ylab("Residual") +
  theme_light()
```



The plots show that the variation in the residuals is about the same for pre and post aspirin headache levels (box plot) and for all predicted headache levels (scatter plot). This satisfies the assumption of homogeneity of variance.

If all assumptions are satisfied, you are at liberty to make inferences regarding the model parameters. We saw that the effect of aspirin was estimated at about 10 points with a rough 95% confidence interval, that was based on the rule of thumb of 2 standard errors around the estimate. For people that are uncomfortable with such quick and dirty estimates, it is also possible to use Satterthwaite's approximation of the degrees of freedom. You can obtain these, once you load the package `lmerTest`. After loading, your `lmer()` analysis will yield (estimated) degrees of freedom and the p-values associated with the t-statistics and those degrees of freedom.

```
library(lmerTest)
out <- data %>%
  lmer(headache ~ measure + (1|patient), data = .)
summary(out)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: headache ~ measure + (1 | patient)
##    Data: .
##
## REML criterion at convergence: 1189.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.36239 -0.42430 -0.04973  0.49702  2.17096
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  patient  (Intercept) 27.144   5.210
##  Residual              8.278   2.877
## Number of obs: 200, groups:  patient, 100
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)  59.6800     0.5952 124.7464  100.28   <2e-16 ***
## measure2    -10.3600     0.4069  99.0000  -25.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr)
## measure2 -0.342
```

The degrees of freedom for effect of `measure2` is 99. If we look in Appendix B or check using R, we see that the critical t-value with 99 degrees of freedom

for a 95% confidence interval is equal to 1.98. That's very close to our rough estimate of 2.

```
qt(0.975, df = 99)
```

```
## [1] 1.984217
```

If we calculate the 95% confidence interval using the new critical value, we obtain $-10.36 - 1.98 \times 0.41 = -11.17$ and $-10.36 + 1.98 \times 0.41 = -9.55$. That's only a minor difference with what we saw before.

12.3 Reporting on a linear mixed model for pre-post data

When you want to write down the results from a linear mixed model, it is important to explain what the model looked like, in terms of fixed and random effects. Usually it is not necessary to mention what method you used to determine degrees of freedom (e.g., Satterthwaite's method).

If you would have to report on the aspirin study illustrated above, you could write down the results as follows:

> "In order to estimate the effect of aspirin, headache measures were collected in 100 patients: one measure before aspirin intake and one measure after aspirin intake. The data were analysed using a linear mixed model, with a fixed effect for measure (before/after) and a random effect for patient. The results showed that the average headache score was 10.36 (SE = 0.41) points lower after aspirin intake than before intake."

If you want to include a confidence interval, you can compute it as shown above using the degrees of freedom reported by the lmerTest package. We saw that the 95% confidence interval runs from $-10.36 - 1.98 \times 0.41$ to $-10.36 + 1.98 \times 0.41$, that is, from -11.17 to -9.55. The value of -10.36 is *negative*, meaning that the scores were *lower* after aspirin. You can report this negative number, but it sounds strange to state that scores were on average -10.36 *higher* after aspirin intake. Makes more sense to switch the signs and state that the scores were on average 10.36 *lower*.

We therefore can report:

> "The results showed that the average headache score was 10.36 (SE = 0.41, 95% CI: 9.55, 11.17) points lower after aspirin intake than before intake."

If the objective of the study was to test the null-hypothesis that the effect of aspirin equals 0, then you could report:

> "In order to test the null-hypothesis that the effect of aspirin on headache equals 0, headache measures were collected in 100 patients: one measure before aspirin intake and one measure after aspirin intake. The data were analysed using a linear mixed model, with a fixed effect for measure (before/after) and a random effect for patient. The results showed that the null-hypothesis could be rejected, $t(99) = -25.46, p < .001$. Headache level after aspirin intake ($M = 49.32$) was significantly lower than headache level before aspirin intake ($M = 59.68$)."

Usually no mention is made of the variance of the random effects, if it is not relevant to the research question.

Chapter 13

Linear mixed models for more than two measurements

13.1 Pre-mid-post intervention designs

In many intervention studies, one has more than two measurement moments. Let's go back to the example of the effect of aspirin on headache in Chapter 12. Suppose you'd like to know whether there is not only a short-term effect of aspirin, but also a long-term effect. Imagine that the study on headache among NY Times readers was extended by asking patients not only to rate their headache before aspirin and 3 hours after intake, but also 24 hours after intake. In this case our data could look like as presented in Table 13.1.

So for each patient we have three measures: **pre**, `post1` and `post2`. To see if there is some clustering, it is no longer possible to study this by computing a single correlation. We could however compute 3 different correlations: **pre**-`post1`, **pre**-`post2`, and `post1`-`post2`, but this is rather tedious, and moreover does not give us a single measure of the extent of clustering of the data. But there is an alternative: one could compute not a Pearson correlation, but an *intraclass correlation* (ICC). To do this, we need to bring the data again into *long format*, as opposed to *wide format*, see Chapter 1. This is done in Table 13.2.

Next, we can perform an analysis with the `lmer()` function from the `lme4` package.

Table 13.1: Headache measures in NY Times readers.

patient	pre	post1	post2
1	52	45	47
2	59	50	55
3	65	56	58
4	51	37	42
5	62	50	55
6	61	53	57
7	56	44	55
8	62	48	53
9	56	48	49
10	58	45	44

Table 13.2: Headache measures in NY Times readers in long format.

patient	measure	headache
1	pre	52
1	post1	45
1	post2	47
2	pre	59
2	post1	50
2	post2	55
3	pre	65
3	post1	56
3	post2	58
4	pre	51

```r
library(lme4)
model1 <- datalong %>%
  lmer(headache ~ measure + (1|patient), data = .)
model1
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: headache ~ measure + (1 | patient)
##    Data: .
## REML criterion at convergence: 1730.488
## Random effects:
##  Groups   Name        Std.Dev.
##  patient  (Intercept) 5.316
##  Residual             2.923
## Number of obs: 300, groups:  patient, 100
## Fixed Effects:
##  (Intercept)  measurepost2    measurepre
##        49.32          2.36          9.85
```

In the output we see the fixed effects of two automatically created dummy variables `measurepost2` and `measurepre`, and the intercept. We also see the standard deviations of the random effects: the standard deviation of the residuals and the standard deviation of the random effects for the patients.

From this output, we can plug in the values into the equation:

$$\begin{aligned} \texttt{headache}_{ij} &= 49.32 + patient_i + 2.36\ \texttt{measurepost2} + 9.85\ \texttt{measurepre} + e_{ij} \\ patient_i &\sim N(0, \sigma_p = 5.316) \\ e_{ij} &\sim N(0, \sigma_e = 2.923) \end{aligned}$$

Based on this equation, the expected headache severity score in the population three hours after aspirin intake is 49.32 (the first post measure is the reference group). Dummy variable `measurepost2` is coded 1 for the measurements 24 hours after aspirin intake. Therefore, the expected headache score 24 hours after aspirin intake is equal to $49.32 + 2.36 = 51.68$. Dummy variable `measurepre` was coded 1 for the measurements before aspirin intake. Therefore, the expected headache before aspirin intake is equal to $49.32 + 9.85 = 59.17$. In sum, in this sample we see that the average headache level decreases directly after aspirin intake from 59.17 to 49.32, but then increases again to 51.68.

There is quite some variation in individual headache levels: the variance is equal to $5.316^2 = 28.260$, since the standard deviation (its square root) is equal to 5.316. Therefore, if we look at roughly 95% of the sample, we see that prior to taking aspirin, the scores vary between $59.17 - 2 \times 5.316 = 48.538$ and

$59.17 + 2 \times 5.29 = 69.802$. For the short-term effect of aspirin after 3 hours, we see that roughly 95% of the scores lie between $49.32 - 2 \times 5.316 = 38.688$ and $49.32 + 2 \times 5.316 = 59.952$. The normal distributions, predicted by this model, are depicted in Figure 13.1.

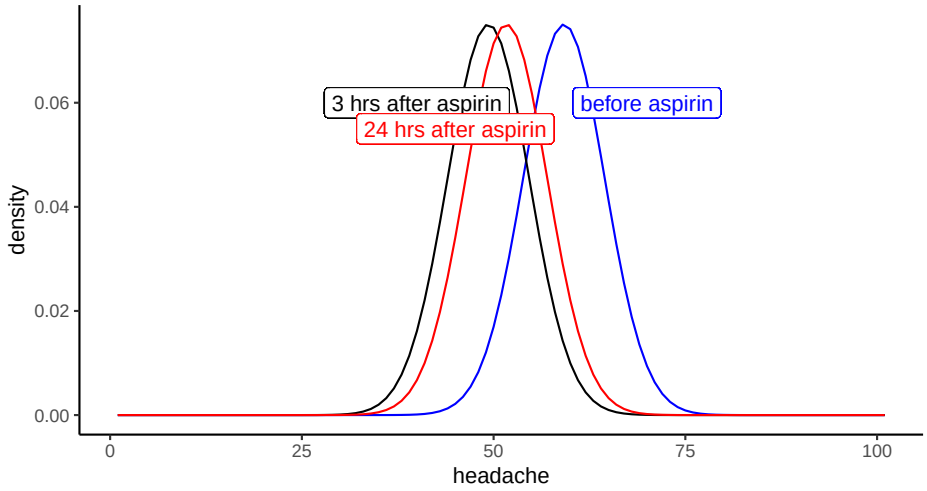


Figure 13.1: Distributions of the three headache levels before aspirin intake, 3 hours after intake and 24 hours after intake, according to the linear mixed model.

So, are these distributions significantly different, in other words, do the means differ significantly before aspirin, 3 hrs after aspirin and 24 hrs after aspirin?

To answer a question about the equality of three means, we need an analysis of variance (ANOVA). Similar to what we did in previous chapters, we specify sum-to-zero coding and type III sums of squares.

```
library(car)
datalong %>%
  lmer(headache ~ measure + (1|patient), data = .,
       contrasts = list(measure = contr.sum)) %>%
  Anova(type = 3, test.statistic = "F")
```

```
## Analysis of Deviance Table (Type III Wald F tests with Kenward-Roger df)
##
## Response: headache
##                    F Df Df.res    Pr(>F)
## (Intercept) 9162.26  1     99 < 2.2e-16 ***
## measure      309.58  2    198 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that instead of an ANOVA table, R plots an Analysis of Deviance table, which you can ignore. Instead of a regular F-statistic, we see the Wald F-test statistic with an estimated Satterthwaite residual degrees of freedom of 198. The degrees of freedom for the measure variable equals 2 (because we have three group means). The F-value is much larger than 1 (remember that the expected value is always 1 if the null-hypothesis is true, see Chapter 6). The p-value shows the significance level. It is less than 0.05, so we can reject the null-hypothesis. We can report:

> "The null-hypothesis that the mean headache level does not change over time was tested with a linear mixed model, with measure entered as a fixed categorical effect ("before", "3hrs after", "24 hrs after") and random effects for patient. An ANOVA showed that the null-hypothesis could be rejected, $F(2, 198) = 309.58, p < .001$."

If one has specific hypotheses regarding short-term and long-term effects, one could perform a planned contrast analysis (see Chapter 10, comparing the first measure with the second measure, and the first measure with the third measure. If one is just interested in whether aspirin has an effect on headache, then the overall F-test should suffice. If apart from this general effect one wishes to explore whether there are significant differences between the three groups of data, without any prior research hypothesis about this, then one could perform a post hoc analysis of the three means. See Chapter 10 on how to perform planned comparisons and post hoc tests.

Now recall that we mentioned an intraclass correlation, or ICC. An intraclass correlation indicates how much clustering there is within the groups, in this case, clustering of headache scores within individual NY Times readers. How much are the three scores alike that come from the same patient? This correlation can be computed using the following formula:

$$ICC = \frac{\sigma^2_{patient}}{\sigma^2_{patient} + \sigma^2_e}$$

Here, the variance of the `patient` random effects is equal to $5.316^2 = 28.260$, and the variance of the residuals e is equal to $2.923^2 = 8.544$, so the intraclass correlation for the headache severity scores is equal to

$$ICC = \frac{28.260}{28.260 + 8.544} = 0.77$$

As this correlation is substantially higher than 0, we conclude there is quite a lot of clustering. Therefore it's a good thing that we used random effects for the individual differences in headache scores among NY Times readers. Had this correlation been 0 or very close to 0, however, then it would not have

mattered to include these random effects. In that case, we might as well use an ordinary linear model, using the `lm()` function. Note from the formula that the correlation becomes 0 when the variance of the random effects for patients is 0. It approaches 0 as the random effects for patients grows small relative to the residual variance. It approaches 1 as the random effects for patients grows large relative to the residual variance. Because variance cannot be negative, ICCs always have values between 0 and 1.

When is an ICC large and when is it small? This is best thought of as being at a birthday party. Most birthday cakes are intended for 8 to 12 guests. If you therefore have one tenth of a birthday cake, that's a substantial piece. It will most likely fill you up. Therefore, an ICC of 0.10 is definitely to be reckoned with. A larger piece than the regular size will fill you up for the rest of the day. A somewhat smaller piece, say 0.05, is also to be taken seriously. It is nice to have a fraction of 0.05 of a whole birthday cake (you don't mind sharing, do you?). However, when the host offers you one percent of the cake (0.01), at least I would be inclined to say no thank you. This implies that when we find an ICC of 0.01, it is fair to simply state that it is very small.

Can we visualise an ICC, in order to better understand what an ICC of 0.01 or 0.50 actually means? Let's try. Imagine a study where we measure bloodpressure in 10 different patients of various ages. Generally, we observe that older people have on average a higher bloodpressure than younger people. In this particular imagined study, we measured bloodpressure 5 times in each patient. This is done because bloodpressure depends very much on time of day, what people are doing, whether they experience stress, etcetera. By measuring bloodpressure 5 times per patient, we get a more reliable insight into people's general bloodpressure level and how it relates to age.

Imagine that we carry out a linear regression where we predict bloodpressure using people's age. We see then a positive relationship: the older the person, the higher the bloodpressure. However, let's look at the residuals of the model. In the interactive app in Figure 13.2, you can change the value of the ICC and see how that affects the residuals. For starters, put the ICC to the value of 0.85. On the left panel of the app, you then see the residuals of the linear regression. It shows the residuals for each patient separately. You see that patient 3 has only positive residuals: based on a prediction using this person's age, we *overestimate* this person's bloodpressure. In contrast, we see that patient 4 has only negative residuals: based on a prediction using age, we *underestimate* this person's bloodpressure. Generally, you can see clear differences in what the residuals look like for each patient separately. This pattern in the residuals shows us that there are other factors besides age that determine a person's bloodpressure. In order to account for this, we can include the categorical variable patient into our linear model. We do that here using random effects. When we do that, and then look at the residuals again, we obtain the residual plot on the right panel in the app. The boxplot shows that the medians are now all very close to 0, which means there are no systematic differences between the

residuals of the 10 different patients anymore.

Because the assumption of linear models is that the residuals are completely random, that is, that there are not systematic effects in the residuals (see Ch. 7), we know that it is very important to include random effects if we observe an ICC of 0.85. Ignoring a factor that causes an ICC of 0.85 can lead to wrong inference (wrong standard errors and confidence intervals).

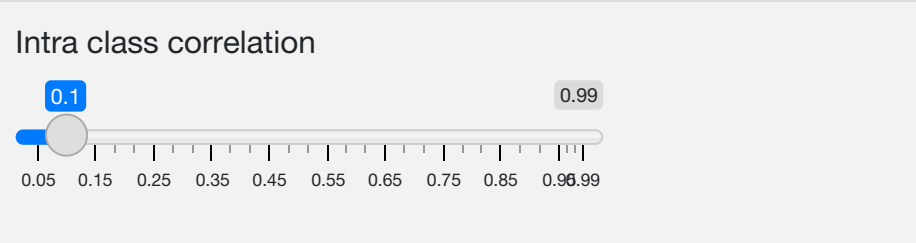


Figure 13.2: [Interactive] A residual plot when the patient variable is not included in a linear model (left panel) and a residal plot when the patient variable is included in the model using random effects (right panel). We see a large difference between the plots, when the ICC has a high value.

However, when the ICC is very low, say 0.01, ignoring a factor does not greatly affect inference. Change the value of the ICC to 0.01 in Figure 13.2, to see what it does to the residuals. The left panel should now resemble the right panel very closely: whether or not you include patient random effects into the model does not visibly affect the pattern in the residuals. Hence, the patient variable can be safely left out of the model.

Play around with different values for the ICC, to check for yourself with what ICC-value you start to notice a difference between the left and the right panel. A difference is indicative that the variable in question (the one on the x-axis) should be included in the linear model.

13.2 Pre-mid-post intervention design: linear effects

In the previous section, we've looked at *categorical* variables: **measure** ("pre-intervention", "3 hours after", and "24 hours after"). We can use the same type of analysis for *numerical* variables. In fact, we could have used a linear effect for time in the headache example: using time of measurement as a numeric variable. Let's look at the headache data again. But now we've created a new variable `time` that is based on the variable **measure**: all first measurements are coded as `time = 0`, all second measurements after 3 hours are coded as `time =`

Table 13.3: Headache measures in NY Times readers in long format with a new variable time.

patient	measure	headache	time
1	pre	52	0
1	post1	45	3
1	post2	47	24
2	pre	59	0
2	post1	50	3
2	post2	55	24
3	pre	65	0
3	post1	56	3
3	post2	58	24
4	pre	51	0

3, and all third measurements after 24 hours are coded as `time = 24`. Part of the data are presented in Table 13.3.

Instead of using a categorical intervention variable, with three levels, we now use a numeric variable, `time`, indicating the number of hours that have elapsed after aspirin intake. At point 0 hours, we measure headache severity, and patients take an aspirin. Next we measure headache after 3 hours and 24 hours. Above, we wanted to know if there were differences in average headache between before intake and 3 hrs and 24 hrs after intake. Another question we might ask ourselves: is there a *linear* reduction in headache severity after taking aspirin?

For this we can do a linear regression type of analysis. We want to take into account individual differences in headache severity levels among patients, so we perform an `lmer()` analysis, using the following code, replacing the categorical variable **measure** by numerical variable `time`:

```
model3 <- datalongnew %>%
  lmer(headache ~ time + (1|patient), data = .)
model3
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: headache ~ time + (1 | patient)
##    Data: .
## REML criterion at convergence: 1993.735
## Random effects:
##  Groups   Name        Std.Dev.
##  patient  (Intercept) 4.561
##  Residual             5.560
## Number of obs: 300, groups:  patient, 100
```

```
## Fixed Effects:
## (Intercept)         time
##     54.7913      -0.1557
```

In the output we see that the model for our data is equivalent to

$$\begin{aligned} \texttt{headache}_{ij} &= 54.79 + patient_i - 0.1557 \times \texttt{time} + e_{ij} \\ patient_i &\sim N(0, \sigma_p = 4.561) \\ e_{ij} &\sim N(0, \sigma_e = 5.560) \end{aligned}$$

This model predicts that at time 0, the average headache severity score equals 54.79, and that for every hour after intake, the headache level drops by 0.1557 points. So it predicts for example that after 10 hours, the headache has dropped 1.557 points to 53.23.

Is this a good model for the data? Probably not. Look at the variance of the residuals: with a standard deviation of 5.56 it is now a lot bigger than in the previous analysis with the same data (see previous section). Larger variance of residuals means that the model explains the data worse: predictions are worse, so the residuals increase in size.

That the model is not appropriate for this data set is also obvious when we plot the data, focusing on the relationship between `time` and **headache** levels, see Figure 13.3.

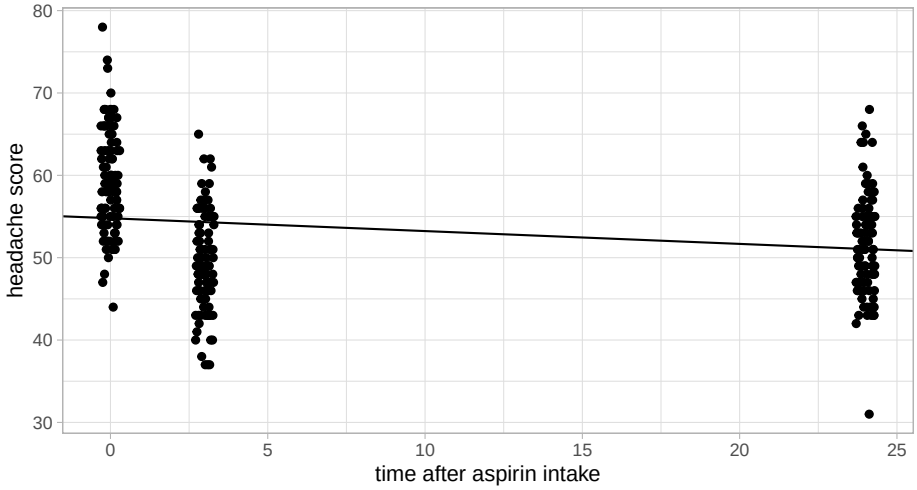


Figure 13.3: Headache levels before aspirin intake, 3 hours after intake and 24 hours after intake.

The line shown is the fitted line based on the output. It can be seen that the prediction for `time = 0` is systematically too low, for `time = 3` systematically too high, and for `time = 24` again too low. So for this particular data set on headache, it would be better to use a categorical predictor for the effect of time on headache, like we did in the previous section.

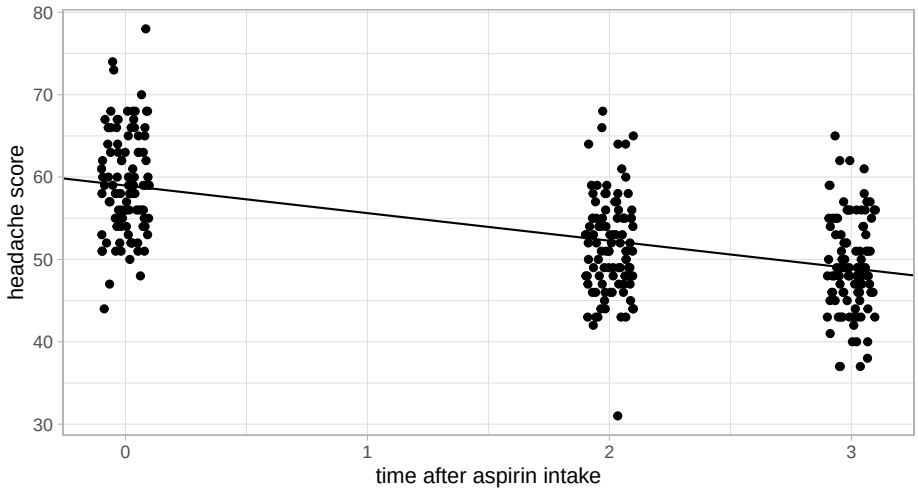


Figure 13.4: Alternative headache levels before aspirin intake, 3 hours after intake and 24 hours after intake.

As an example of a data set where a linear effect would have been appropriate, imagine that we measured headache 0 hours, 2 hours and 3 hours after aspirin intake (and not after 24 hours). Suppose these data would look like those in Figure 13.4. There we see a gradual increase of headache levels right after aspirin intake. Here, a numeric treatment of the time variable would be quite appropriate.

Suppose we would then see the following output.

```
model4 <- datalongnew2 %>%
  lmer(headache ~ time + (1|patient), data = .)
model4 %>% summary()
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: headache ~ time + (1 | patient)
##    Data: .
##
## REML criterion at convergence: 1738.8
##
```

```
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.78038 -0.55298  0.05078  0.55931  2.30389
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  patient  (Intercept) 28.186   5.309
##  Residual              8.776   2.962
## Number of obs: 300, groups:  patient, 100
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  58.9721     0.6028   97.83
## time         -3.3493     0.1371  -24.42
##
## Correlation of Fixed Effects:
##      (Intr)
## time -0.379
```

Because we are confident that this model is appropriate for our data, we can interpret the statistical output. The Satterthwaite error degrees of freedom are 199, so we can construct a 95% confidence interval by finding the appropriate t-value.

```
qt(0.975, df = 199)
```

```
## [1] 1.971957
```

The 95% confidence interval for the effect of time is then from $-3.35-1.97\times0.14$ to $-3.35+1.97\times0.14$, so from -3.63 to -3.07. We can report:

> "A linear mixed model was run on the headache levels, using a fixed effect for the numeric predictor variable time and random effects for the variable patient. We saw a significant linear effect of time on headache level, $t(200) = -24.42, p < .001$. The estimated effect of time based on this analysis is negative, -3.35, so with every hour that elapses after aspirin intake, the predicted headache score decreases with 3.35 points (95% CI: 3.07 to 3.63 points)".

13.3 Linear mixed models and interaction effects

Suppose we carry out the aspirin and headache study not only with a random sample of NY Times readers that suffer from regular headaches, but also with

Table 13.4: Headache measures in NY Times and Wall Street Journal readers in wide format.

patient	group	pre	post
1	NYTimes	55	45
2	WallStreetJ	63	50
3	NYTimes	66	56
4	WallStreetJ	50	37
5	NYTimes	63	50
6	WallStreetJ	65	53

Table 13.5: Headache measures in NY Times and Wall Street Journal readers in long format.

patient	group	measure	headache
1	NYTimes	pre	55
1	NYTimes	post	45
2	WallStreetJ	pre	63
2	WallStreetJ	post	50
3	NYTimes	pre	66
3	NYTimes	post	56

a random sample of readers of the Wall Street Journal that suffer from regular headaches. We'd like to know whether aspirin works, but we are also interested to know whether the effect of aspirin is similar in the two groups of readers. Our null-hypothesis is that the effect of aspirin in affecting headache severity is the same in NY Times and Wall Street Journal readers that suffer from headaches.

H_0: The effect of aspirin is the same for NY Times readers as for Wall Street Journal readers.

Suppose we have the data set in Table 13.4 (we only show the first six patients), and we only look at the measurements before aspirin intake and 3 hours after aspirin intake (pre-post design).

In this part of the data set, patients 2, 4, and 6 read the Wall Street Journal, and patients 1, 3 and 5 read the NY Times. We assume that people only read one of these newspapers. We measure their headache before and after the intake of aspirin (a pre-post design). The data are now in what we call *wide format*: the dependent variable **headache** is spread over two columns, **pre** and **post**. In order to analyse the data with linear models, we need them in *long format*, as in Table 13.5.

```
datalong <- datawide %>%
  pivot_longer(cols = -c(patient, group),
               names_to = "measure",
               values_to = "headache")
```

The new variable **measure** now indicates whether a given measurement of headache refers to a measurement before intake (pre) or after intake (post). Again we could investigate whether there is an effect of aspirin with a linear mixed model, with **measure** as our categorical predictor, but that is not really what we want to test: we only want to know whether the effect of aspirin (being small, large, negative or non-existent) *is the same for both groups*. Remember that this hypothesis states that there is no interaction effect of aspirin (**measure**) and **group**. The null-hypothesis is that **group** is *not* a moderator of the effect of aspirin on headache. There may be an effect of aspirin or there may not, and there may be an effect of newspaper (**group**) or there may not, but we're interested in the *interaction* of aspirin and group membership. Is the effect of aspirin different for NY Times readers than for Wall Street Journal readers?

In our model we therefore need to specify an interaction effect. Since the data are clustered (2 measures per patient), we use a linear *mixed* model. First we show how to analyse these data using dummy variables, later we will show the results using a different approach.

We recode the data into two dummy variables, one for the aspirin intervention (**dummy1**: 1 if **measure** = post, 0 otherwise), and one for group membership (**dummy2**: 1 if **group** = NYTimes, 0 otherwise):

```
datalong <- datalong %>%
  mutate(dummy1 = ifelse(measure == "post", 1, 0),
         dummy2 = ifelse(group == "NYTimes", 1, 0))
datalong %>% head(3)
```

```
## # A tibble: 3 x 6
##   patient group       measure headache dummy1 dummy2
##     <int> <chr>       <chr>      <dbl>  <dbl>  <dbl>
## 1       1 NYTimes     pre           55      0      1
## 2       1 NYTimes     post          45      1      1
## 3       2 WallStreetJ pre           63      0      0
```

Next we need to compute the product of these two dummies to code a dummy for the interaction effect. Since with the above dummy coding, all post measures get a 1, and all NY Times readers get a 1, only the observations that are post aspirin and that are from NY Times readers get a 1 for this product.

```
datalong <- datalong %>%
  mutate(dummy_int = dummy1*dummy2)
datalong %>% head(3)
```

```
## # A tibble: 3 x 7
##   patient group       measure headache dummy1 dummy2 dummy_int
##     <int> <chr>       <chr>      <dbl>  <dbl>  <dbl>     <dbl>
## 1       1 NYTimes     pre           55      0      1         0
## 2       1 NYTimes     post          45      1      1         1
## 3       2 WallStreetJ pre           63      0      0         0
```

With these three new dummy variables we can specify the linear mixed model.

```
model5 <- datalong %>%
  lmer(headache ~ dummy1 + dummy2 + dummy_int + (1|patient),
       data = .)
model5
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: headache ~ dummy1 + dummy2 + dummy_int + (1 | patient)
##    Data: .
## REML criterion at convergence: 1185.5
## Random effects:
##  Groups   Name        Std.Dev.
##  patient  (Intercept) 5.229
##  Residual             2.884
## Number of obs: 200, groups:  patient, 100
## Fixed Effects:
## (Intercept)       dummy1       dummy2    dummy_int
##       59.52       -10.66         0.32         0.60
```

In the output, we recognise the three fixed effects for the three dummy variables. Since we're interested in the interaction effect, we look at the effect of **dummy_int**. The effect is in the order of +0.6. What does this mean?

Remember that all headache measures before aspirin intake are given a 0 for the intervention dummy **dummy1**. A reader from the Wall Street Journal gets a 0 for the group dummy **dummy2**. Since the product of 0×0 equals 0, all measures before aspirin in Wall Street Journal readers get a 0 for the interaction dummy **dummy_int**. Therefore, the intercept of 59.52 refers to the expected headache severity of Wall Street Journal readers *before* they take their aspirin.

Furthermore, we see that the effect of **dummy1** is -10.66. The variable **dummy1** codes for post measurements. So, relative to Wall Street Journal

Table 13.6: Expected headache levels in Wall Street Journal and NY Times readers, before and after aspirin intake.

measure	group	dummy1	dummy2	dummy_int	exp_mean
pre	WallStreetJ	0	0	0	60
post	WallStreetJ	1	0	0	60 + (-11) = 49
pre	NYTimes	0	1	0	60 + 0.3 = 60.3
post	NYTimes	1	1	1	60 + (-11) + 0.3 + 0.6 = 49.9

readers prior to aspirin intake, the level of post intake headache is 10.66 points *lower*.

If we look further in the output, we see that the effect of **dummy2** equals +0.32. This variable **dummy2** codes for NY Times readers. So, relative to Wall Street Journal readers and before aspirin intake (the reference group), NY Times readers score on average 0.32 points higher on the headache scale.

However, we're not interested in a general difference between those two groups of readers, we're interested in the effect of aspirin and whether it is different in the two groups of readers. In the output we see the interaction effect: being a reader of the NY Times AND at the same time being a measure after aspirin intake, the expected level of headache is an extra +0.60. The effect of aspirin is -10.66 in Wall Street Journal readers, as we saw above, but the effect is $-10.66 + 0.60 = -10.06$ in NY Times readers. So in this sample the effect of aspirin on headache is 0.60 *smaller* than in Wall Street Journal readers (note that even while the interaction effect is positive, it is positive on a scale where a high score means more headache).

Let's look at it in a different way, using a table with the dummy codes, see Table 13.6. For each group of data, pre or post aspirin and NY Times readers and Wall Street Journal readers, we note the dummy codes for the new dummy variables. In the last column we use the output estimates and multiply them with the respective dummy codes (1 and 0) to obtain the expected headache level (using rounded numbers):

The exact numbers are displayed in Figure 13.5.

We see that the specific effect of aspirin in NY Times readers is 0.60 smaller than the effect of aspirin in Wall Street Journal readers. This difference in the effect of aspirin between the groups was not significantly different from 0, as we can see when we let R plot a summary of the results.

```
model5 %>% summary()
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: headache ~ dummy1 + dummy2 + dummy_int + (1 | patient)
```

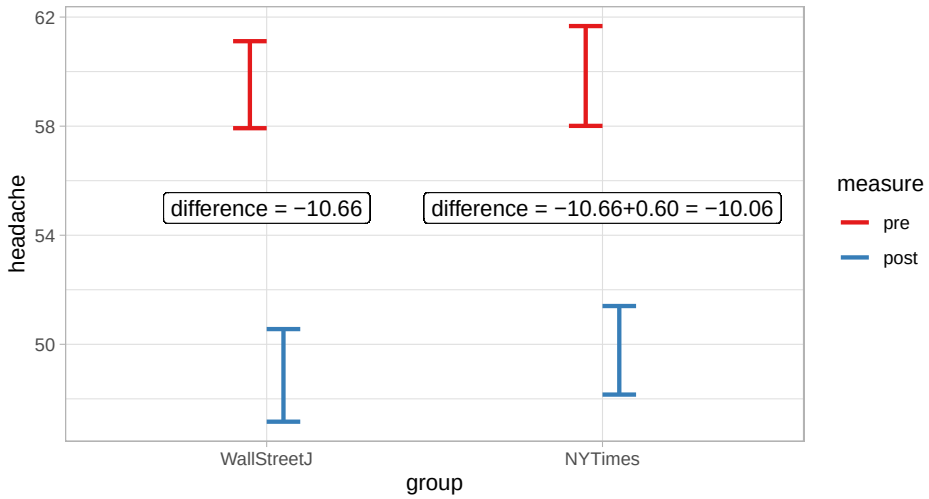


Figure 13.5: Expected headache levels in NY Times readers and Wall Street Journal readers based on a linear mixed model with an interaction effect.

```
##    Data: .
##
## REML criterion at convergence: 1185.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.31732 -0.43239 -0.02912  0.50368  2.20297
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  patient  (Intercept) 27.346   5.229
##  Residual              8.317   2.884
## Number of obs: 200, groups:  patient, 100
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  59.5200     0.8445  70.476
## dummy1      -10.6600     0.5768 -18.482
## dummy2        0.3200     1.1944   0.268
## dummy_int     0.6000     0.8157   0.736
##
## Correlation of Fixed Effects:
##           (Intr) dummy1 dummy2
## dummy1    -0.341
## dummy2    -0.707  0.241
## dummy_int  0.241 -0.707 -0.341
```

> "The null-hypothesis that the effect of aspirin is the same in the two populations of readers cannot be rejected, $t(98) = 0.736, p = .464$. We therefore conclude that there is no evidence that aspirin has a different effect for NY Times than for Wall Street Journal readers."

Note that we could have done the analysis in another way, not recoding the variables into numeric dummy variables ourselves, but by letting R do it automatically. R does that automatically for factor variables like our variable **group**. The code is then:

```
model6 <- datalong %>%
  lmer(headache ~ measure + group + measure:group + (1|patient),
       data = .)
model6 %>% summary()
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: headache ~ measure + group + measure:group + (1 | patient)
##    Data: .
##
## REML criterion at convergence: 1185.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.31732 -0.43239 -0.02912  0.50368  2.20297
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  patient  (Intercept) 27.346   5.229
##  Residual              8.317   2.884
## Number of obs: 200, groups:  patient, 100
##
## Fixed effects:
##                             Estimate Std. Error t value
## (Intercept)                  49.7800     0.8445  58.943
## measurepre                   10.0600     0.5768  17.442
## groupWallStreetJ             -0.9200     1.1944  -0.770
## measurepre:groupWallStreetJ   0.6000     0.8157   0.736
##
## Correlation of Fixed Effects:
##             (Intr) mesrpr grpWSJ
## measurepre  -0.341
## grpWllStrtJ -0.707  0.241
## msrpr:grWSJ  0.241 -0.707 -0.341
```

R has automatically created dummy variables, one dummy `measurepre` that codes 1 for all measurements before aspirin intake, one dummy

`groupWallStreetJ` that codes 1 for all measurements from Wall Street Journal readers, and one dummy `measurepre:groupWallStreetJ` that codes 1 for all measurements from Wall Street Journal readers before aspirin intake. Because the dummy coding is different from the hand coding we did ourselves earlier, the intercept and the main effects of group and measure are now different. We also see that the significance level of the interaction effect is still the same. You are always free to choose to either construct your own dummy variables and analyse them in a quantitative way (using numeric variables), or to let R construct the dummy variables for you (by using a factor variable): the p-value for the interaction effect will always be the same (this is not true for the intercept and the main effects).

Because the two analyses are equivalent (they end up with exactly the same predictions, feel free to check!), we can safely report that we have found a non-significant group by measure interaction effect, $t(98) = 0.736, p = .464$. We therefore conclude that we found no evidence that in the populations of NY Times readers and Wall Street Journal readers, the short-term effect of aspirin on headache is any different.

13.4 Mixed designs

The design in the previous section, where we had both a grouping variable and a pre-post or repeated measures design, is often called a *mixed design*. It is a mixed design in the sense that there are two kinds of variables: one is a *between-individuals* variable, and one variable is a *within-individual* variable. Here the between-individuals variable is **group**: two different populations of readers. It is called *between* because one individual can only be part of one group. When we study the effect of the group effect we are essentially comparing the scores of one group of individuals with the scores of another group of individuals, so the comparison is *between different individuals*. The two groups of data are said to be *independent*, as we knew that none of the readers in this data set reads both journals.

The within-variable in this design is the aspirin intervention, indicated by the variable **measure**. For each individual we have two observations: all individuals are present in both the pre condition data as well as in the post condition data. With this intervention variable, we are comparing the scores of a group of individuals with the scores *of that same group of individuals* at another time point. The comparison of scores is within a particular individual, at time point 1 and at time point 2. So the pre and post sets of data are not independent: the headache scores in both conditions are coming from the same individuals.

Mixed designs are often seen in psychological experiments. For instance, you want to know how large the effect of alcohol intake is on driving performance. You want to know whether the effect of alcohol on driving performance is the same in a Fiat 600 as in a Porsche 918. Suppose you have 100 participants for

your study. There are many choices you can make regarding the design of your study. Here we discuss 4 alternative research designs:

1. One option is to have all participants participate in all four conditions: they all drive a Fiat with and without alcohol, and they all drive a Porsche, with and without alcohol. In this case, both the car and the alcohol are within-participant variables.
2. The second option is to have 50 participants drive a Porsche, with and without alcohol, and to have the other 50 participants drive the Fiat, with and without alcohol. In this case, the car is the between-participants variable, and alcohol is the within-participant variable.
3. The third option is to have 50 participants without alcohol drive both the Porsche and the Fiat, and to have the other 50 participants drive the Porsche and the Fiat with alcohol. Now the car is the within-participant variable, and the alcohol is the between-participants variable.
4. The fourth option is to have 25 participants drive the Porsche with alcohol, 25 other participants drive the Porsche without alcohol, 25 participants drive the Fiat with alcohol, and the remaining 25 participants drive the Fiat without alcohol. Now both the car variable and the alcohol variable are between-participant variables: none of the participants is present in more than 1 condition.

Only the second and the third design described here are mixed designs, having at least one between-participants variable and at least one within-participant variable.

Remember that when there is at least one within variable in your design, you have to use a linear mixed model. If all variables are between variables, one can use an ordinary linear model. Note that the term *mixed* in linear mixed model refers to the effects in the model that can be both random and fixed. The term *mixed* in mixed designs refers to the mix of two kinds of variables: within variables and between variables.

Also note that the within and between distinction refers to the units of analysis. If the unit of analysis is school, then the denomination of the school is a between-school variable. An example of a within-school variable could be time: before a major curriculum reform and after a major curriculum reform. Or it could be teacher: classes taught by teacher A or by teacher B, both teaching at the same school.

13.5 Mixed design with a linear effect

In an earlier section we looked at a mixed design where the between variable was **group** and the within variable was **measure**: pre or post. It was a 2 by

Table 13.7: Cortisol measures over time.

person	group	measure	cortisol
1	Olympian	1	19
1	Olympian	2	20
1	Olympian	3	22
1	Olympian	4	23
1	Olympian	5	24
1	Olympian	6	22

2 design (2×2) design: 2 measures and 2 groups, where we were interested in the interaction effect. We wanted to know whether **group** moderated the effect of **measure**, that is, the effect of aspirin on headache. We used the categorical within-individual variable **measure** in the regression by dummy-coding it.

In an earlier section in this chapter we saw that we can also model linear effects of numeric variables in linear mixed models, where we treated the time variable numerically: 0hrs, 3hrs after aspirin intake and 24 hrs after intake. Here we will give an example of a 3×20 mixed design: we have a categorical group (between-individuals) variable with 3 levels and a numeric time (within) variable with 20 levels. The example is about stress in athletes that are going to take part in the 2018 Winter Olympics. Stress can be revealed in morning cortisol levels. In the 20 days preceding the start of the Olympics, each athlete was measured every morning after waking and before breakfast by letting them chew on cotton. The cortisol level in the saliva was then measured in the lab. Our research question is by how much cortisol levels rise in athletes that prepare for the Olympics.

Three groups were studied. One group consisted of 50 athletes who were selected to take part in the Olympics, one group consisted of 50 athletes that were very good but were not selected (Control group I) and one group consisted of 50 non-athlete spectators that were going to watch the games (Control group II). The research question was about what the differences are in average cortisol increase in these three groups: the Olympians, Control group I and Control group II.

In Table 13.7 you see part of the fictional data, the first 6 measurements on person 1 that belongs to the group of Olympians.

When we plot the data, and use different colours for the three different groups, we already notice that the Olympians show generally higher cortisol levels, particularly at the end of the 20-day period (Figure 13.6).

We want to know to what extent the linear effect of time is moderated by group. Since for every person we have 20 measurements, the data are clustered so we use a linear mixed model. We're looking for a linear effect of time, so we use the **measure** variable numerically (i.e., it is numeric, and we do not transform it into a factor). We also use the categorical variable **group** as a predictor. It

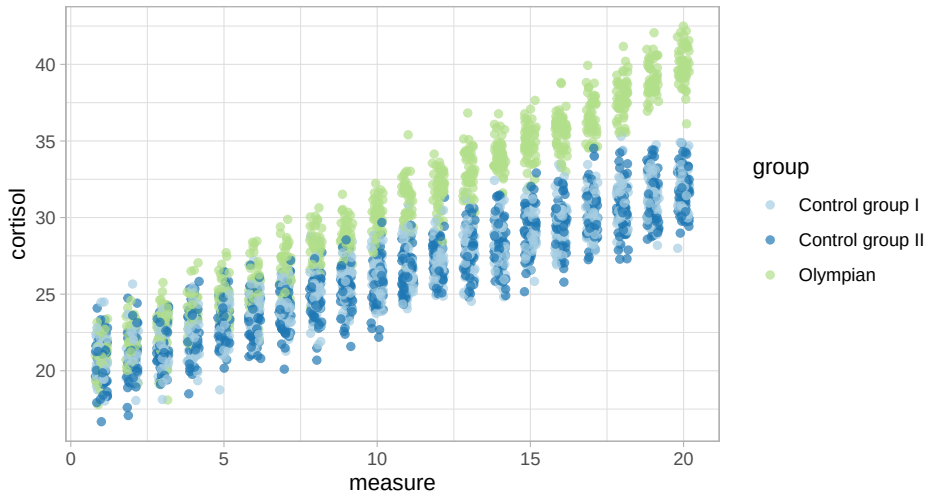


Figure 13.6: Cortisol levels over time in three groups.

is a factor variable with three levels, so R will automatically make two dummy variables. Because we're interested in the interaction effects, we include both main effects of **group** and **measure** as well as their interaction in the model. Lastly, we control for individual differences in cortisol levels by introducing random effects for **person**.

```
model7 <- datalong %>%
  lmer(cortisol ~ measure + group + measure:group + (1|person),
       data = .)
model7 %>% summary()
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: cortisol ~ measure + group + measure:group + (1 | person)
##    Data: .
##
## REML criterion at convergence: 8985.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5807 -0.6414 -0.0112  0.6625  3.3128
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  person   (Intercept) 0.9862   0.9931
##  Residual             0.9970   0.9985
## Number of obs: 3000, groups:  person, 150
```

```
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                    20.094018   0.155004 129.636
## measure                         0.596989   0.005476 109.020
## groupControl group II          -0.226802   0.219208  -1.035
## groupOlympian                  -0.402950   0.219208  -1.838
## measure:groupControl group II   0.006383   0.007744   0.824
## measure:groupOlympian           0.411896   0.007744  53.188
##
## Correlation of Fixed Effects:
##             (Intr) measur grCgII grpOly m:CgII
## measure     -0.371
## grpCntrlgII -0.707  0.262
## groupOlympn -0.707  0.262  0.500
## msr:grpCgII  0.262 -0.707 -0.371 -0.185
## msr:grpOlym  0.262 -0.707 -0.185 -0.371  0.500
```

In the output we see an intercept of 20.09, a slope of about 0.597 for the effect of **measure**, two main effects for the variable **group** (Control group I is the reference group), and two interaction effects (one for Control group II and one for the Olympian group). Let's fill in the linear model equation based on this output:

$$\begin{aligned}\texttt{cortisol}_{ij} = 20.09 + person_i + 0.5970\,\texttt{measure} - 0.2268\,\texttt{ContrGrII} - 0.4029\,\texttt{Olympian} + \\ 0.0063\,\texttt{ContrGrII measure} + 0.4119\,\texttt{Olympian measure} + e_{ij} \\ person_i \sim N(0, \sigma_p^2 = 0.9862) \\ e_{ij} \sim N(0, \sigma_e^2 = 0.9970)\end{aligned}$$

We see a clear intraclass correlation of around $\frac{0.9862}{0.9862+0.9970} = 0.497$ so it's a good thing we've included random effects for `person`. The expected means at various time points and for various groups can be made with the use of the above equation.

It helps interpretation when we look at what linear effects we have for the three different groups. Filling in the above equation for Control group I (the reference group), we get:

$$\texttt{cortisol}_{ij} = 20.09 + person_i + 0.597 \times \texttt{measure} + e_{ij}$$

For Control group II we get:

$$\begin{aligned}\texttt{cortisol}_{ij} &= 20.09 + person_i + 0.597\,\texttt{measure} - 0.2268 + 0.006\,\texttt{measure} + e_{ij} \\ &= 19.8632 + person_i + 0.603 \times \texttt{measure} + e_{ij}\end{aligned}$$

And for the Olympians we get:

$$\begin{aligned}\texttt{cortisol}_{ij} &= 20.09 + person_i + 0.597\ \texttt{measure} - 0.4029 + 0.4119\ \texttt{measure} + e_{ij} \\ &= 19.6871 + person_i + 1.0089 \times \texttt{measure} + e_{ij}\end{aligned}$$

In these three equations, all intercepts are close to 20. The slopes are about 0.6 in both Control groups I and II, whereas the slope is around 1 in the group of Olympian athletes. For illustration, these three implied linear regression lines are depicted in Figure 13.7.

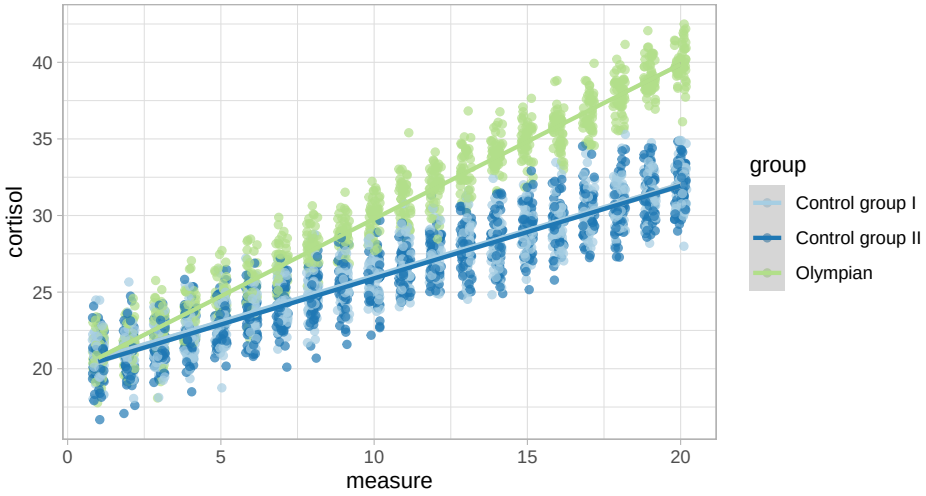


Figure 13.7: Cortisol levels over time in three groups with the group-specific regression lines.

So based on the linear model, we see that in this sample the rise in cortisol levels is much steeper in Olympians than in the two control groups. But is this true for all Olympians and the rest of the populations of high performing athletes and spectators? Note that in the regression table we see two interaction effects: one for `measure:groupControl group II` and one for `measure:groupOlympian`. Here we're interested in the rise in cortisol in the three different groups and to what extent these sample differences extend to the three populations.

A possible answer we find in the F-statistic of an ANOVA. When we run an ANOVA on the results,

```
datalong %>%
  lmer(cortisol ~ measure + group + measure:group + (1|person),
       data = .,
       contrasts = list(group = contr.sum)) %>%
  Anova(type = 3, test.statistic = "F")
```

```
## Analysis of Deviance Table (Type III Wald F tests with Kenward-Roger df)
##
## Response: cortisol
##                         F Df  Df.res Pr(>F)
## (Intercept)   49368.4919  1  197.39 <2e-16 ***
## measure       54256.1350  1 2847.00 <2e-16 ***
## group             1.6984  2  197.39 0.1856
## measure:group  1857.2017  2 2847.00 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

we see a significant measure by group interaction effect, $F(2, 2847) = 1857.20, p < .001$. The null-hypothesis of the same cortisol change in three different populations can be rejected, and we conclude that Olympian athletes, non-Olympian athletes and spectators show a different change in cortisol levels in the weeks preceding the games.

Chapter 14

Non-parametric alternatives for linear mixed models

14.1 Checking assumptions

In previous chapters we discussed the assumptions of linear models and linear mixed models: linearity (in parameters), homoscedasticity (equal variance), normal distribution of residuals, normal distribution of random effects (relevant for linear mixed models only), and independence (no clustering unaccounted for).

The problem of non-linearity can sometimes be solved by introducing quadratic terms, by replacing a linear model $Y = b_0 + b_1X + e$ by another linear model $Y = b_0 + b_1X + b_2X^2 + e$.

If we have non-independence, then you can introduce either an extra fixed effect or a random effect for a clustering variable. For example, if you see that cars owned by low income families have much more mileage than cars owned by high income families, you can account for this by adding a fixed effect of an income variable as predictor. If you see that average mileage is rather similar within municipality but that average mileage can vary quite a lot across municipalities, you can introduce a random effect for municipality (if you have data say from 30 different municipalities).

Unequal variance of residuals and non-normal distribution of residuals are harder to tackle. Non-normality can sometimes be solved by using generalised linear models (see Chapter 15). A combination of non-normality and unequal variance can sometimes be easily solved by using a transformation of the dependent variable, for instance not analysing $Y = b_0 + b_1X + e$ but analysing $\log(Y) = b_0 + b_1X + e$ or $\sqrt{Y} = b_0 + b_1X + e$. There are also more advanced options available that will not be discussed here but that make corrections to

standard errors and p-values that render inference more robust against model violations.

If these data transformations or advanced options don't work (or if you're not acquainted with them), and your data show non-equal variance and/or non-normally distributed residuals, there are non-parametric alternatives. Here we discuss two: Friedman's test and Wilcoxon's signed rank test. We explain them using an imaginary data set on speed skating.

Suppose we have data on 12 speed skaters that participate on the 10 kilometres distance in three separate championships in 2017-2018: the European Championships, the Winter Olympics and the World Championships. Your friend expects that speed skaters will perform best at the Olympic games, so there she expects the fastest times. You disagree and decide to test the null-hypothesis that average times are the same at the three occasions. You regard the data from 2017-2018 to be a random sample of all seasons in the past and future. In Figure 14.1 we see a boxplot of the data.

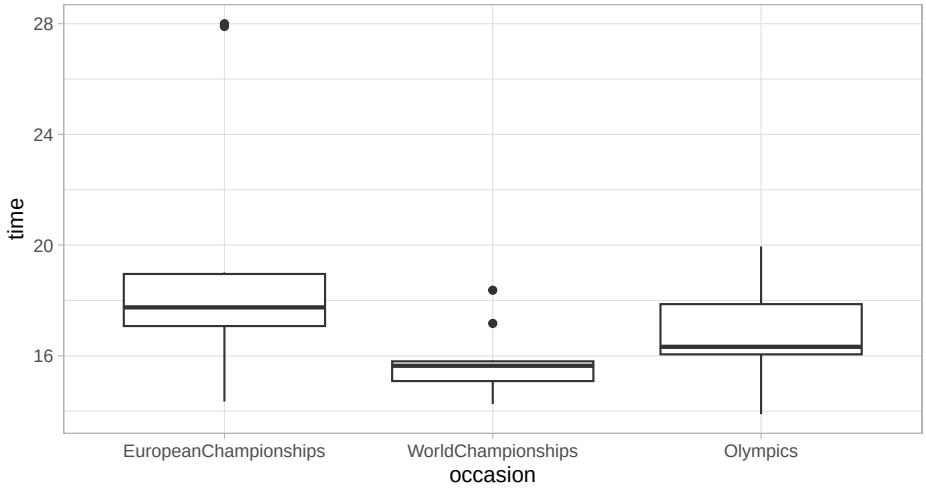


Figure 14.1: Boxplot of the imaginary speed skating data.

In order to test your null-hypothesis, we run a linear mixed model with dependent variable `time`, and independent variable `occasion`. We use random effects for the differences in speed across skaters. In Figure 14.2 we see the residuals. From this plot we clearly see that the assumption of equal variance (homogeneity of variance) is violated: the variance of the residuals in the World Championships condition is clearly smaller than the variance of the European Championships condition. From the histogram of the residuals in Figure 14.3 we also see that the distribution of the residuals is not bell-shaped: it is positively skewed (skewed to the right).

Since the assumptions of homogeneity of variance and of normally distributed

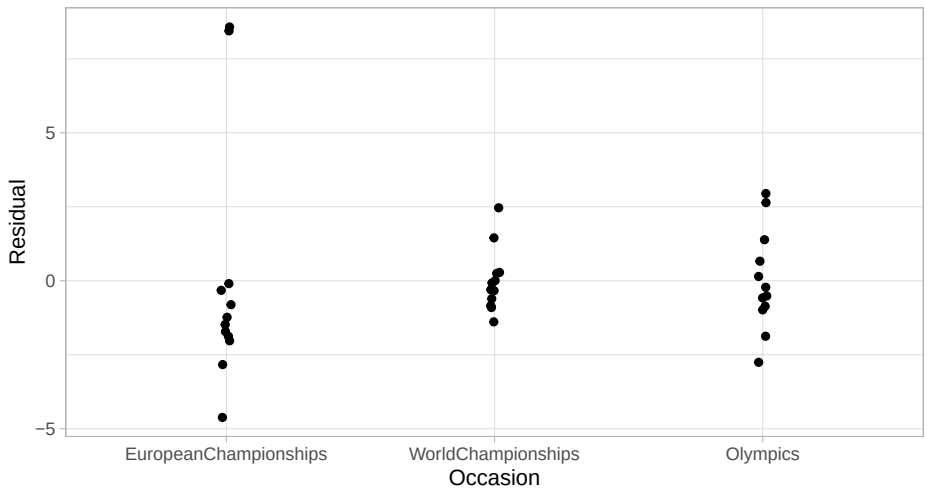


Figure 14.2: Residuals of the speedskating data with a linear mixed model.

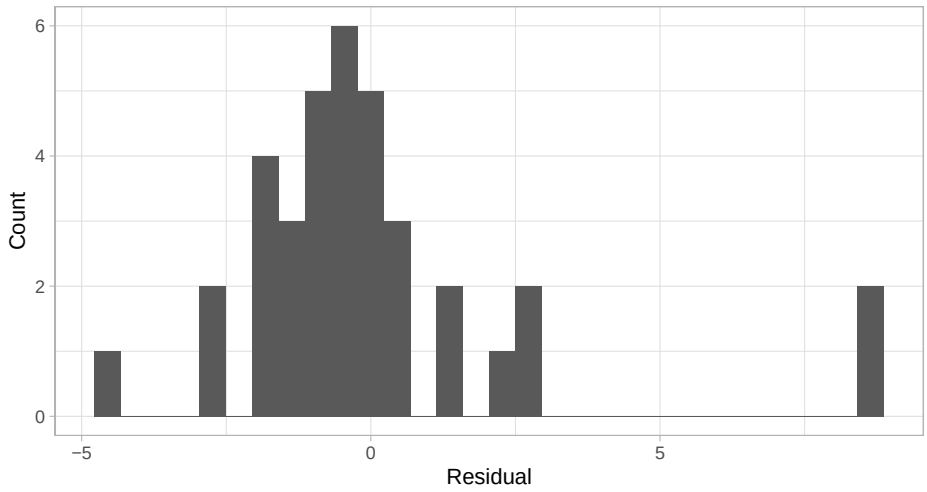


Figure 14.3: Histogram of the residuals of the speedskating data with a linear mixed model.

Table 14.1: The speed skating data in wide format.

athlete	EuropeanChampionships	WorldChampionships	Olympics
1	14.35	15.79	16.42
2	17.36	14.26	18.13
3	19.01	18.37	19.95
4	27.90	15.12	17.78
5	17.67	17.17	16.96
6	17.83	15.30	16.15
7	16.30	15.63	19.44
8	28.00	15.69	16.23
9	18.27	15.65	15.76
10	17.00	14.99	16.18
11	17.10	15.83	13.89
12	18.94	14.77	14.83

residuals are violated[1], the results from the linear mixed model cannot be trusted. In order to answer our research question, we therefore have to resort to another kind of test. Here we discuss Friedman's test, a non-parametric test, for testing the null-hypothesis that the *medians* of the three groups of data are the same (see Chapter 1). This Friedman test can be used in all situations where you have at least 2 levels of the within variable. In other words, you can use this test when you have data from three occasions, but also when you have data from 10 occasions or only 2. In a later section the Wilcoxon signed ranks test is discussed. This test is often used in social and behavioural sciences. The downside of Wilcoxon's test is that it can only handle data sets with 2 levels of the within variable. In other words, it can only be used when we have data from two occasions. Friedman's test is therefore more generally applicable than Wilcoxon's.

14.2 Friedman's test for k measures

Similar to many other non-parametric tests for testing the equality of medians, Friedman's test is based on ranks. Table 14.1 shows the speed skating data in wide format.

We rank all of these time measures by determining the fastest time, then the next to fastest time, etcetera, until the slowest time. But because the data in

[1]Remember that assumptions relate to the population, not samples: often-times your data set is too small to say anything about assumptions at the population level. Residuals for a data set of 8 persons might show very non-normal residuals, or very different variances for two subgroups of 4 persons each, but that might just be a coincidence, a random result because of the small sample size. If in doubt, it is best to use non-parametric methods.

Table 14.2: Row-wise ranks of the speed skating data.

athlete	EuropeanChampionships	WorldChampionships	Olympics
1	1	2	3
2	2	1	3
3	2	1	3
4	3	1	2
5	3	2	1
6	3	1	2
7	2	1	3
8	3	1	2
9	3	1	2
10	3	1	2
11	3	2	1
12	3	1	2

each row belong together (we compare individuals with themselves), we do the ranking *row-wise*. For each athlete separately, we determine the fastest time (1), the next fastest time (2), and the slowest time (3) and put the ranks in a new table, see Table 14.2. There we see for example that athlete 1 had the fastest time at the European Championships (14.35, rank 1) and the slowest at the Olympics (16.42, rank 3).

Next, we compute the sum of the ranks column-wise: the sum of the ranks for the European Championships data is 31, for the Olympic data it's 15 and for the World Championships data it is 26. We call these sums S_j, where the j indicates the column.

From these sums S_j we can gather that in general, these athletes showed their best times (many rank 1s) at the World Championships, as the sum of the ranks is lowest. We also see that in general these athletes showed their worst times (many rank 2s and 3s) at the European Championships, as the relevant column showed the highest sum of ranks.

In order to know whether these sums of ranks are significantly different from each other, we may compute an F_r-value based on the following formula:

$$F_r = \left[\frac{12}{nk(k+1)}\Sigma_{j=1}^{k} S_j^2\right] - 3n(k+1)$$

In this formula, n stands for the number of rows (12 athletes), k stands for the number of columns (3 occasions), and S_j^2 stands for the squared sum of column j (31^2, 15^2, and 26^2, respectively). If we fill in these numbers, we get:

Table 14.3: The raw skating data in random order.

athlete	EuropeanChampionships	WorldChampionships	Olympics
1	15.76	14.26	17.78
2	19.44	17.10	17.83
3	16.18	14.83	15.63
4	15.30	17.00	16.42
5	15.79	16.23	17.36
6	14.77	28.00	15.65
7	14.35	16.15	16.30
8	27.90	15.12	18.94
9	16.96	19.01	19.95
10	15.83	17.17	13.89
11	14.99	18.27	15.69
12	18.37	17.67	18.13

$$\begin{aligned} F_r &= \left[\frac{12}{12 \times 3(3+1)} \times (31^2 + 15^2 + 26^2)\right] - 3 \times 12(3+1) \\ &= \left[\frac{12}{144} \times 1862\right] - 144 = 11.167 \end{aligned}$$

What can we tell from this F_r-statistic? In order to say something about significance, we have to know what values are to be expected under the null-hypothesis that there are no differences across the three groups of data. Suppose we randomly mixed up the data by taking all the speed skating times and randomly assigning them to the three contests and the twelve athletes, until we have a newly filled data matrix, for example the one in Table 14.3.

If we then compute F_r for this data matrix, we get a different value. If we do this mixing up the data and computing F_r say 1000 times, we get 1000 values for F_r, summarised in the histogram in Figure 14.4.

So if the data are just randomly distributed over the three columns (and 12 rows) in the data matrix, we expect no systematic differences across the three columns and so the null-hypothesis is true. So now we know what the distribution of F_r looks like when the null-hypothesis is true: more or less like the one in Figure 14.4. Remember that for the true data that we actually gathered (in the right order that is!), we found an F_r-value of 11.167 . From the histogram, we see that only very few values of 11.167 or larger are observed when the null-hypothesis is true. If we look more closely, we find that only 0.2% of the values are larger than 11.167 , so we have a two-tailed p-value of 0.004. The 95th percentile of these 1000 F_r-values is 6.1666667, meaning that of the 1000 values for F_r, 5% are larger than 6.1666667. So if we use a significance level of 5%, our observed

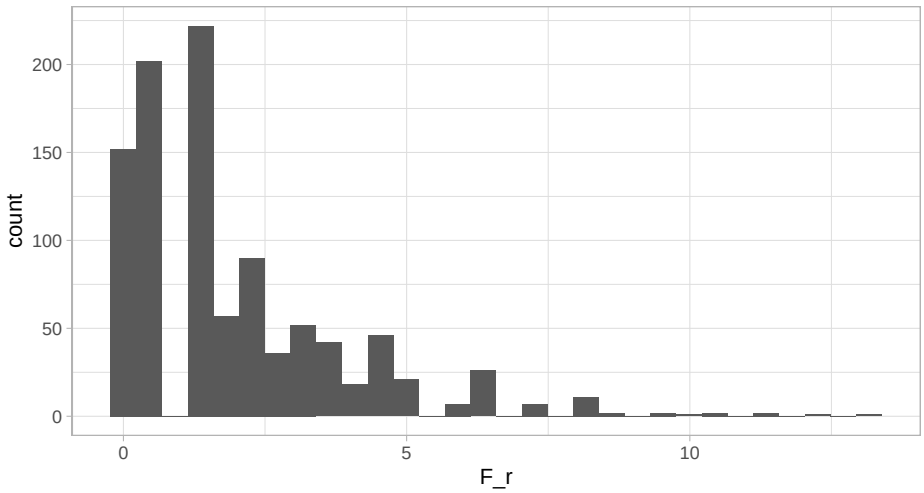


Figure 14.4: Histogram of 1000 possible values for F_r given that the null-hypothesis is true, for 12 speed skaters.

value of 11.167 is larger than the critical value for F_r, and we conclude that the null-hypothesis can be rejected.

Now this p-value of 0.004 and the critical value of 6.1666667 are based on our own computations[2]. Actually there are better ways. One is to look up critical values of F_r in tables, for instance in Kendall M.G. (1970) *Rank correlation methods*. (fourth edition). The p-value corresponding to this F_r-value depends on k, the number of groups of data (here 3 columns) and n, the number of rows (12 individuals). If we look up that table, we find that for $k = 3$ and $n = 12$ the critical value of F_r for a type I error rate of 0.05 equals 6.17. Our observed F_r-value of 11.167 is larger than that, therefore we can reject the null-hypothesis that the median skating times are the same at the three different championships. So we have to tell your friend that there are general differences in skating times at different contests, $F_r = 11.167, p < 0.05$, but it is not the case that the fastest times were observed at the Olympics.

A third way to do null-hypothesis testing is to make an approximation of the distribution of F_r under the null-hypothesis. Note that the distribution in the histogram in Figure 14.4 is very strangely shaped. The reason is that the data set is quite limited. Suppose we have data not on 12 speed skaters, but on 120. If we then randomly mix up data again and compute 1000 different values for F_r, we get the histogram in Figure 14.5.

[2]What we have actually done is a very simple form of *bootstrapping* or *permutation testing*: jumbling up the data set many times and in that way determining the distribution of a test-statistic under the null-hypothesis, in this case the distribution of F_r. For more on bootstrapping, see Davison, A.C. & Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge, UK: Cambridge.

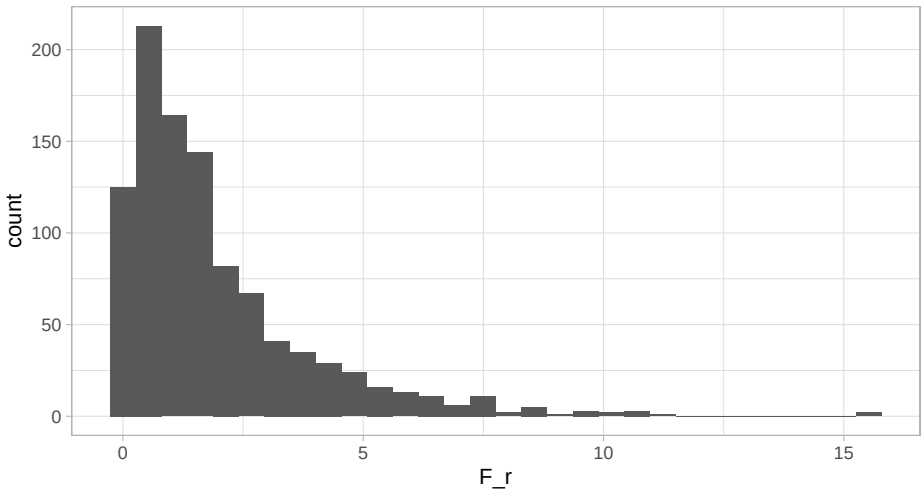


Figure 14.5: Histogram of 1000 possible values for F_r given that the null-hypothesis is true, for 120 speed skaters.

The shape becomes more regular. It also starts to resemble another distribution, that of the χ^2 (chi-square). It can be shown that the distribution of the F_r for a large number of rows in the data matrix, and at least 6 columns, approaches the shape of the χ^2-distribution with $k-1$ degrees of freedom. This is shown in Figure 14.6.

The density of the χ^2-distribution with 2 degrees of freedom approaches the histogram quite well, but not perfectly. In general, for large n and $k > 5$, the approximation is very good. In that way it gets easier to look up p-values for certain F_r-values, because the χ^2-distribution is well-known[3], so we don't have to look up critical values for F_r in old tables. For a significance level of 5%, the critical value of a χ^2 with 2 degrees of freedom is 5.991. We can see that in R with the following code:

```
# critical value for chi-square distribution
# with 2 degrees of freedom and alpha = 0.05:
qchisq(p = 0.95, df = 2)
```

```
## [1] 5.991465
```

This is close to the value in the table for F_r in old books: 6.17. The part of the χ^2-distribution with 2 degrees of freedom that is larger than the observed 11.167 is 0.004, so our approximate p-value for our null-hypothesis is 0.004.

[3]The χ^2-distribution is based on the normal distribution: the χ^2-distribution with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.

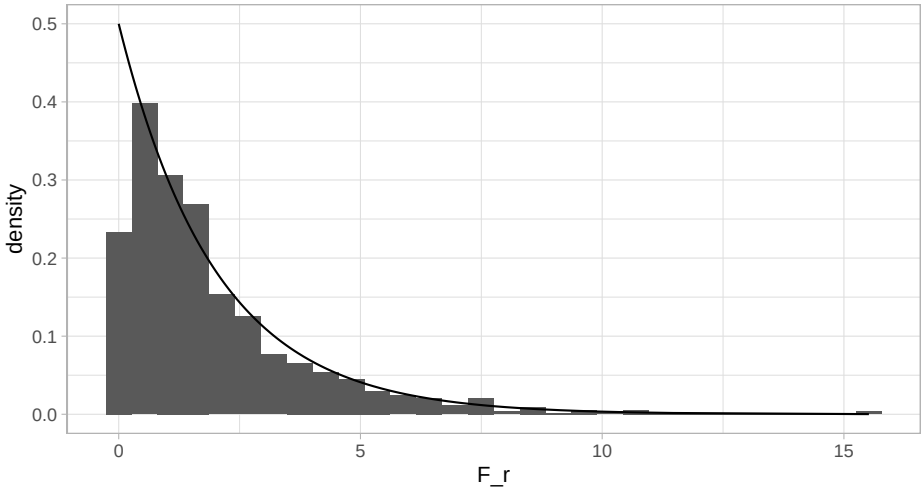


Figure 14.6: The distribution of F_r under the null-hypothesis, overlain with a chi-square distribution with 2 degrees of freedom.

```
# p-value for statistic of 11.167 based on
# chi-square distribution with 2 degrees of freedom:
pchisq(11.167, df = 2, lower.tail = F)
```

```
## [1] 0.003759385
```

14.3 Comparing Friedman's test with linear mixed model on ranks (advanced)

Remember that for many non-parametric tests, they are equivalent to applying a linear model on ranks. This is not true for this Friedman test. For this data set, we could ignore potential problems with the data and carry out a linear mixed model analysis, using the actual finishing times, and predict them with a fixed effect for the type of championship and a random effect for each individual skater. If you have doubts about the assumptions, apply a non-parametric test like the Friedman test.

Alternatively, we could apply ranking on all the finish times, not row-wise as we do for Friedman's test but lumping all times together. Next, we apply the same linear mixed model analysis with a fixed effect for the type of championship and a random effect the skaters, now using the ranks as dependent variable. The result would be a perfectly acceptable analysis, however, it is not equivalent to

Table 14.4: The raw skating data in long data format.

athlete	occasion	time
1	EuropeanChampionships	14.35
1	Olympics	16.42
1	WorldChampionships	15.79
2	EuropeanChampionships	17.36
2	Olympics	18.13
2	WorldChampionships	14.26

Friedman's test.[4]

14.4 How to perform Friedman's test in R

In order to let R do the calculations for you, you need first of all your data to be in long format. If your data happen to be in wide format, use the `pivot_longer()` function to get the data in long format. Suppose your data is in wide format, as in Table 14.1. Then the following code turns the data into long format:

```
datawide %>%
  pivot_longer(cols = -athlete,
               names_to = "occasion",
               values_to = "time")
```

This creates the long format data matrix in Table 14.4:

We can then specify that we want Friedman's test by using the `friedman.test()` function and indicating which variables we want to use:

```
datalong %>%
  friedman.test(time ~ occasion | athlete, data = .)
```

```
##
##  Friedman rank sum test
##
## data:  time and occasion and athlete
## Friedman chi-squared = 11.167, df = 2, p-value = 0.00376
```

[4] See for instance https://seriousstats.wordpress.com/2012/02/14/friedman/

This code says, "use `time` as dependent variable, `occasion` as independent variable, and the data are clustered within athletes (data with the same `athlete` ID number belong together)".

In the output we see a chi-squared statistic, degrees of freedom, and an asymptotic (approximated) p-value. Why don't we see an F_r-statistic?

The reason is, as discussed in the previous section, that for a large number of measurements (in wide format: columns) and a large number of individuals (in wide format: rows), the F_r statistic tends to have the same distribution as a chi-square, χ^2, with $k-1$ degrees of freedom. So what we are looking at in this output is really an F_r-value of 11.167 (exactly the same value as we computed by hand in the previous section). In order to approximate the p-value, this value of 11.167 is interpreted as a chi-square (χ^2), which with 2 degrees of freedom has a p-value of 0.004.

This asymptotic (approximated) p-value is the correct p-value if you have a lot of rows (large n) and at least 6 variables ($k > 5$). If you do not have that, as we have here, this asymptotic p-value is only what it is: an approximation. However, this is only a problem when the approximate p-value is close to the pre-selected significance level α. If α equals 0.05, an approximate p-value of 0.002 is much smaller than that, and we do not hesitate to call it significant, whatever its true value may be. If a p-value is very close to α, it might be a good idea to look up the exact critical values for F_r in online tables[5]. If your F_r is larger than the critical value for a certain combination of n, k and α, you may reject the null-hypothesis.

In the above case we can report:

> "A Friedman's test showed that speed skaters have significantly different median times on the 10 kilometre distance at the three types of contests, $F_r = 11.167, p = .004$."

Linear mixed model version on rank data

Althought it is not equivalent to Friedman's test, the same data could be analysed using a linear mixed model on the ranks. We start from the long data format with the original data and create new variable for the ranks:

```
datalong <- datalong %>%
  mutate(rank = rank(time))
datalong %>% head()
```

```
##   athlete              occasion  time rank
## 1       1 EuropeanChampionships 14.35    3
```

[5] https://www.jstor.org/stable/3315656?seq=1

```
## 2       1              Olympics 16.42   19
## 3       1    WorldChampionships 15.79   13
## 4       2 EuropeanChampionships 17.36   24
## 5       2              Olympics 18.13   28
## 6       2    WorldChampionships 14.26    2
```

Note that we now have large values for the ranks: the data is ranked across all skaters and all championships, in contrast to Friedman's test where the ranking is done within individual skaters.

Next we apply a linear mixed model and ask for an ANOVA, testing the null-hypothesis that the three championships show the same average rank in the population.

```
library(lmerTest)
library(car)
datalong %>%
  lmer(rank ~ occasion + (1|athlete), data = .) %>%
  Anova(type = 3, test.statistic = "F")
```

```
## Analysis of Deviance Table (Type III Wald F tests with Kenward-Roger df)
##
## Response: rank
##                    F Df Df.res    Pr(>F)
## (Intercept) 92.3181  1 32.409 5.246e-11 ***
## occasion     7.6717  2 22.000  0.002967 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can report:

> "A linear mixed model on the ranked times showed that speed skaters have significantly different median times on the 10 kilometre distance across the three types of contests, $F(2, 22) = 7.67, p = .003$."

14.5 Wilcoxon's signed ranks test for 2 measures

Friedman's test can be used for 2 measures, 3 measures or even 10 measures. As stated earlier, the well-known Wilcoxon's test can only be used for 2 measures. For completeness, we also discuss that test here. For that test we compare the measures at the Olympics and at the World Championships.

Table 14.5: The raw skating data and the computations for Wilcoxon signed ranks test

athlete	WorldChampionships	Olympics	d	rank_d	ranksign
1	15.79	16.42	0.63	5	5
2	14.26	18.13	3.87	12	12
3	18.37	19.95	1.58	8	8
4	15.12	17.78	2.66	10	10
5	17.17	16.96	-0.21	3	-3
6	15.30	16.15	0.85	6	6
7	15.63	19.44	3.81	11	11
8	15.69	16.23	0.54	4	4
9	15.65	15.76	0.11	2	2
10	14.99	16.18	1.19	7	7
11	15.83	13.89	-1.94	9	-9
12	14.77	14.83	0.06	1	1

For each athlete, we take the difference in skating times (Olympics - WorldChampionships) and call it d, see Table 14.5. Next we rank these d-values, irrespective of sign, and call these ranks `rank_d`. From Table 14.5 we see that athlete 12 shows the smallest difference in skating times ($d = 0.06$, rank $= 1$) and athlete 2 the largest difference ($d = 3.78$, rank $= 12$).

Next, we indicate for each rank whether it belongs to a positive or a negative difference d and call that variable `ranksign`.

Under the null-hypothesis, we expect that some of the larger d-values are positive and some of them negative, in a fairly equal amount. If we sum the ranks having plus-signs and sum the ranks having minus-signs, we would expect that these two sums are about equal, but only if the null-hypothesis is true. If the sums are very different, then we should reject this null-hypothesis. In order to see if the difference in sums is too large, we compute them as follows:

$$T^+ = 5 + 12 + 8 + 10 + 6 + 11 + 4 + 2 + 7 + 1 = 66$$
$$T^- = 3 + 9 = 12$$

To know whether T^+ is significantly larger than T^-, the value of T^+ can be looked up in a table, for instance in Siegel & Castellan (1988). There we see that for T^+, with 12 rows, the probability of obtaining a T^+ of at least 66 is 0.0171. For a two-sided test (if we would have switched the columns of the two championships, we would have gotten a T^- of 66 and a T^+ of 12!), we have to double this probability. So we end up with a p-value of $2 \times 0.0171 = 0.0342$.

In the table we find no critical values for large sample size n, but fortunately,

similar to the Friedman test, we can use an approximation using a different distribution, here the normal distribution. It can be shown that for large sample sizes, the statistic T^+ is approximately normally distributed with mean

$$\mu = \frac{n(n+1)}{4}$$

and variance:

$$\sigma^2 = \frac{n(n+1)(2n+1)}{24}$$

If we therefore standardise the T^+ by subtracting the μ and then dividing by the square root of the variance $\sqrt{(\sigma^2)} = \sigma$, we get a z-value with mean 0 and standard deviation 1. To do that, we use the following formula:

$$z = \frac{T^+ - \mu}{\sigma} = \frac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

Here T^+ is 66 and n equals 12, so if we fill in the formula we get $z = 2.118$. From the standard normal distribution we know that 5% of the observations lie above 1.96 and below -1.96. So a value for z larger than 1.96 or smaller than -1.96 is enough evidence to reject the null-hypothesis. Here our z-statistic is larger than 1.96, therefore we reject the null-hypothesis that the median skating times are the same at the World Championships and the Olympics. The p-value associated with a z-score of 2.118 is 0.034.

The analysis for a Wilcoxon test looks complicated, but is actually equivalent to a linear mixed model applied to ranked data for moderate to large data sets.[6] This will be further discussed at the end of the next section.

14.6 How to perform Wilcoxon's signed ranks test in R

If you want R to perform the Wilcoxon test, your data needs to be in wide format. If your data are in long format, you can transform them with the function `pivot_wider()`:

```
datawide <- datalong %>%
  pivot_wider(id_cols = athlete, values_from = "time", names_from = "occasion")
```

[6] see https://seriousstats.wordpress.com/2012/02/14/friedman/

Next, you use the `wilcox.test()` function. You select the two variables (occasions) that you would like to compare, and indicate that the Olympic and World Championship data are paired within rows (i.e., the Olympic and World championship data in one row belong to the same individual).

```
wilcox.test(datawide$Olympics, datawide$WorldChampionships,
            paired = TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  datawide$Olympics and datawide$WorldChampionships
## V = 66, p-value = 0.03418
## alternative hypothesis: true location shift is not equal to 0
```

The output shows a V-statistic, which corresponds to our T^+ above. The standard output yields an approximate p-value. This p-value is by default two-sided. Wilcoxon's T^+-statistic (or V) under the null-hypothesis approaches a normal distribution in case we have a large number of observations (many rows in wide format). If $n > 15$, the approximation is good enough so that if we standardise this statistic, it can be interpreted as a z-score (standardised score with a normal distribution). That means that a z-score of 1.96 or larger or -1.96 or smaller can be regarded as significant at the 5% significance level. Since the standard normal distribution is only an approximation, and we have $n = 12$, it is safer here to look at the exact significance level. That can be done with the option `exact = TRUE`:

```
wilcox.test(datawide$Olympics, datawide$WorldChampionships,
            paired = TRUE,
            exact = TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  datawide$Olympics and datawide$WorldChampionships
## V = 66, p-value = 0.03418
## alternative hypothesis: true location shift is not equal to 0
```

In this case we see that the exact p-value is equal, to the fifth decimal, to the approximate p-value. Note that we use a two-sided test, to allow for the fact that random sampling could lead to a higher median for the Olympic Games or a higher median for the World Championships. We just want to know whether the null-hypothesis that the two medians differ can be rejected (in whatever direction) or not.

Let's compare the output with the Friedman test, but then only use the relevant occasions in your code:

```
datalong %>%
  filter(occasion != "EuropeanChampionships") %>%
  mutate(occasion = as.integer(occasion)) %>%
  friedman.test(time ~ occasion | athlete,
                data = .)
```

```
##
##  Friedman rank sum test
##
## data:  time and occasion and athlete
## Friedman chi-squared = 5.3333, df = 1, p-value = 0.02092
```

Note that the `friedman.test()` function does not perform well if some variables are factors and you make a selection of the levels. Here we have the factor `occasion` with originally 3 levels. If the `friedman.test()` function only finds two of these in the data it is supposed to analyse, it throws an error. Therefore we turn factor variable `occasion` into an integer variable first. The `friedman.test()` function then turns this integer variable into a new factor variable with only 2 levels before the calculations.

In the output we see that the null-hypothesis of equal medians at the World Championships and the Olympic Games can be rejected, with an approximate p-value of 0.02.

Note that both the Friedman and Wilcoxon tests come up with very similar p-values, even though their rationales are different: Friedman's test is based on ranks and Wilcoxon's test is based on the size of positive and negative differences between measures 1 and 2. Wilcoxon's test therefore uses more information than Friedman's test. Both can be used in the case you have two measures. Friedman's test can be used in all situations where you have 2 or more measures per row, whereas Wilcoxon's test can only be used if you have 2 measures per row.

In sum, we can report in two ways on our hypothesis regarding similar skating times at the World Championships and at the Olympics:

1. "A Friedman test showed a significant difference between the 10km skating times at the World Championships and at the Olympics, $F_r = 5.33, p = .02$."

2. "A Wilcoxon signed ranks test showed a significant difference between the 10km skating times at the World Championships and at the Olympics, $T^+ = 66, p = .03$."

How do we know whether the fastest times were at the World Championships or at the Olympics? If we look at the raw data in Table 14.1, it is not immediately obvious. We have to inspect the $T^+ = 66$ and $T^- = 12$ and consider what they represent: there is more positivity than negativity. The positivity is due to positive ranksigns that are computed based on $d = Olympics - WorldChampionships$, see Table 14.5. A positive difference d means that the Olympics time was larger than the WorldChampionships time. A large value for time stands for slower speed. A positive ranksign therefore means that the Olympics time was larger (slower!) than the WorldChampionships time. A large `rank_d` means that the difference between the two times was relatively large. Therefore, you get a large value of T^+ if the Olympic times are on the whole slower than the World Championships times, and/or when these positive differences are relatively large. When we look at the values of `ranksign` in Table 14.5, we notice that only two values are negative: one relatively large value and one relatively small value. The rest of the values are positive, both small and large, and these all contribute to the T^+ value. We can therefore state that the pattern in the data is that for most athletes, the Olympic times are slower than the times at the World Championships.

Of course, this is better visualised with a graph, see Figure 14.7. Easy to see that for 2 skaters, the times were fastest at the Olympics. The rest had their best time at the World Championships.

Wilcoxon equals linear mixed model applied to ranks

Wilcoxon's test can also be conceived of as the analog of a linear model on ranks. If we turn the skating data from the World Championships and the Olympic Games into ranks, we can apply a linear mixed model with a fixed effect for occasion and a random effect for the skater. In R we start from the data in long format:

```
library(lmerTest)
datalong %>%
  filter(occasion != "EuropeanChampionships") %>%
  mutate(rank = rank(time)) %>%
  lmer(rank ~ occasion + (1|athlete), data = .) %>%
  summary()
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: rank ~ occasion + (1 | athlete)
```

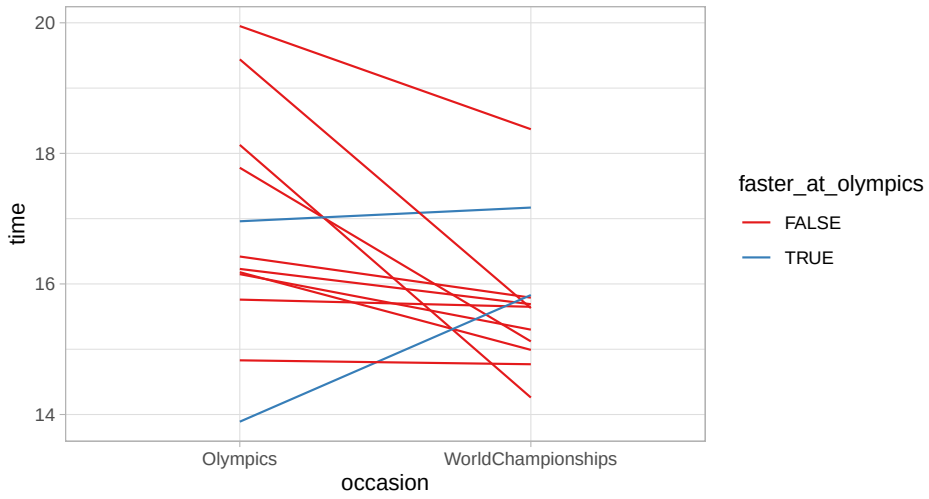


Figure 14.7: Skating times at the Olympics and the World Championships. Only two of the skaters show their fastest time at the Olympics.

```
##    Data: .
##
## REML criterion at convergence: 149.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.11878 -0.59704  0.05797  0.78980  1.48883
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  athlete  (Intercept)  8.939   2.99
##  Residual             34.576   5.88
## Number of obs: 24, groups:  athlete, 12
##
## Fixed effects:
##                  Estimate Std. Error     df t value Pr(>|t|)
## (Intercept)         9.667      1.904 21.109   5.076 4.93e-05 ***
## occasionOlympics    5.667      2.401 11.000   2.361   0.0378 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## occsnOlympc -0.630
```

From the results we see that the average rank is higher in the Olympics data. The p-value is lower than 0.05, so we conclude:

> "A linear mixed model on the ranked data (Wilcoxon's test) showed a significant difference between the 10km skating times at the World Championships and at the Olympics, $t = 2.36, p = .04$."

Note that the p-value is slightly different compared to the Wilcoxon test result earlier. This is because the data set is relatively small. For moderate to large sample sizes, the difference will disappear.

14.7 Ties

Many non-parametric tests are based on ranks. For example, if we have the data sequence 0.1, 0.4, 0.5, 0.2, we give these values the ranks 1, 3, 4, 2, respectively. But in many data cases, data sequences cannot be ranked unequivocally. Let's look at the sequence 0.1, 0.4, 0.4, 0.2. Here we have 2 values that are exactly the same. We say then that we have *ties*. If we have ties in our data like the 0.4 in this case, one very often used option is to arbitrarily choose one of the 0.4 values as smaller than the other, and then average the ranks. Thus, we rank the data into 1, 3, 4, 2 and then average the tied observations: 1, 3.5, 3.5, 2. As another example, suppose we have the sequence 23, 54, 54, 54, 19, we turn this into ranks 2, 3, 4, 5, 1 and take the average of the ranks of the tied observations of 54: 2, 4, 4, 4, 1. These ranks corrected for ties can then be used to compute the test statistic, for instance Friedman's F_r or Wilcoxon's z. However, in many cases, because of these corrections, a slightly different formula is to be used. So the formulas become a little bit different. This is all done in R automatically. If you want to know more, see Siegel and Castellan (1988). *Non-parametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

14.8 Take-away points

- When a distribution of residuals looks very far removed from a normal distribution and/or shows heterogeneity of variance, consider either a data transformation or using a non-parametric method of analysis.
- Friedman's and Wilcoxon's tests are non-parametric alternatives for linear mixed modelling with one categorical predictor variable, in a within-subjects design.
- Wilcoxon's can be used for an independent variable with only two categories.

- Friedman's test can also be used for an independent variable with more than two categories.
- Alternatively to Wilcoxon's test and Friedman's test, one can apply linear mixed models to ranked data. Wilcoxon's test is actually equivalent to a linear mixed model on ranks, and will yield the same p-values for large enough samples.

Key concepts

- Friedman's test
- Wilcoxon's signed ranks test
- Ties

Chapter 15

Generalised linear models: logistic regression

15.1 Introduction

In previous chapters we were introduced to the linear model, with its basic form

$$\begin{aligned} Y &= b_0 + b_1 X_1 + \cdots + b_n X_n + e \\ e &\sim N(0, \sigma_e^2) \end{aligned}$$

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,

$$\begin{aligned} Y &= b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2 + e \\ e &\sim N(0, \sigma_e^2) \end{aligned}$$

or when we use a quadratic term to account for another type of non-linearity in the data:

$$\begin{aligned} Y &= b_0 + b_1 X_1 + b_2 X_1 X_1 + e \\ e &\sim N(0, \sigma_e^2) \end{aligned}$$

In all these models, the assumption is that the effects of the parameters (b_0, b_1, b_2) can be summed.

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 $N(\mu, \sigma^2)$, in principle all values between $-\infty$ and $+\infty$ are possible, but they tend to concentrate around the mean μ, in the shape of the bell-curve.

A normal distribution is suitable for continuous dependent variables. Most measured variables are however not continuous at all. Think for example of temperature: if a thermometer gives degrees Celsius with a precision of 1 decimal, the 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 15.1 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 of the residuals seems well approximated by a normal curve.

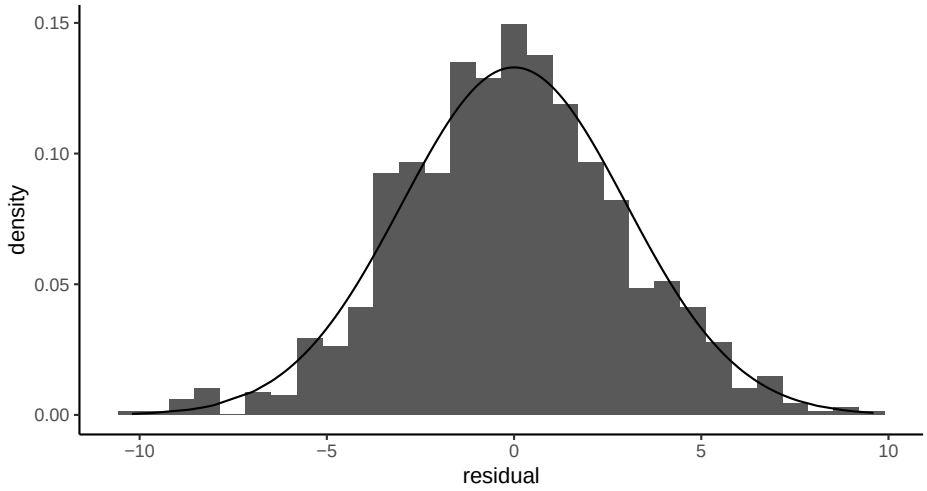


Figure 15.1: A histogram of residuals with a normal curve (black line). Even if measures are in fact discrete, the normal distribution can be a good approximation of the distribution of the residuals.

But let's look at an example where the discreteness is more prominent. In Figure 15.2 we see the residuals of an analysis of test results. Students had to do an assignment that had to meet four criteria: 1) originality, 2) language, 3) structure, and 4) literature review. Each criterion was registered as either fulfilled (1) or not fulfilled (0). The total 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. We call such a variable a *count variable*. In an analysis, this score was predicted on the basis of the average test score on

previous assignments, using a linear model.

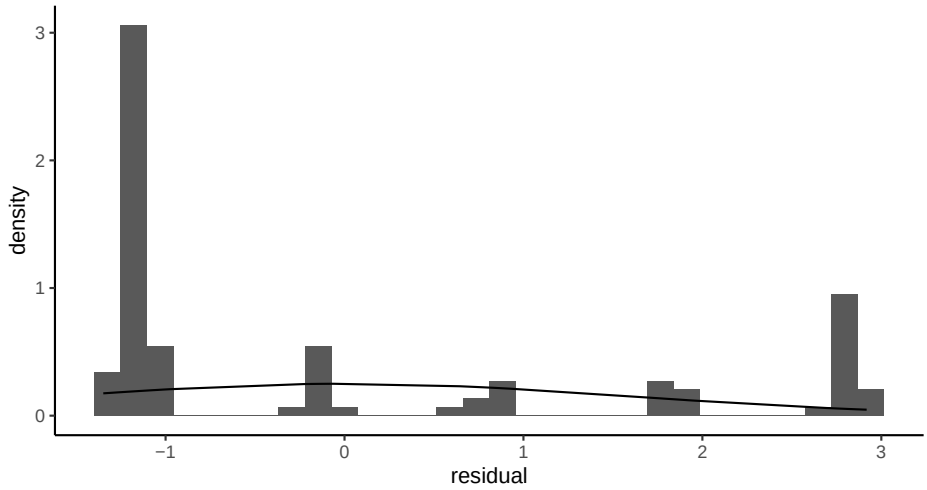


Figure 15.2: Count data example where the best fitting normal distribution is not a good approximation of the distribution of the residuals.

Figure 15.2 shows that the residuals are very discrete, and that a normal distribution with the same mean and variance is a very bad approximation of the distribution. We often see this phenomenon when our dependent variable has only a limited number of possible values.

An even more extreme case we observe when our dependent variable consists of only two values. For example, suppose our dependent variable is whether or not students passed the assignment: only those assignments that fulfilled all four criteria are regarded as sufficient. If we score all students who passed as 1 and score all students who failed as 0 and we predict this 0/1 score by the average test score on previous assignments using a linear model, we get the residuals displayed in Figure 15.3.

Here it is also evident that a normal approximation of the residuals will not do. When the dependent variable has only two 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 two possible values (dichotomous dependent variables like yes/no or pass/fail, heads/tails, 1/0). In Chapter 16, we will discuss the case where the dependent variable consists of counts $(0, 1, 2, 3, 4, 5, ...)$.

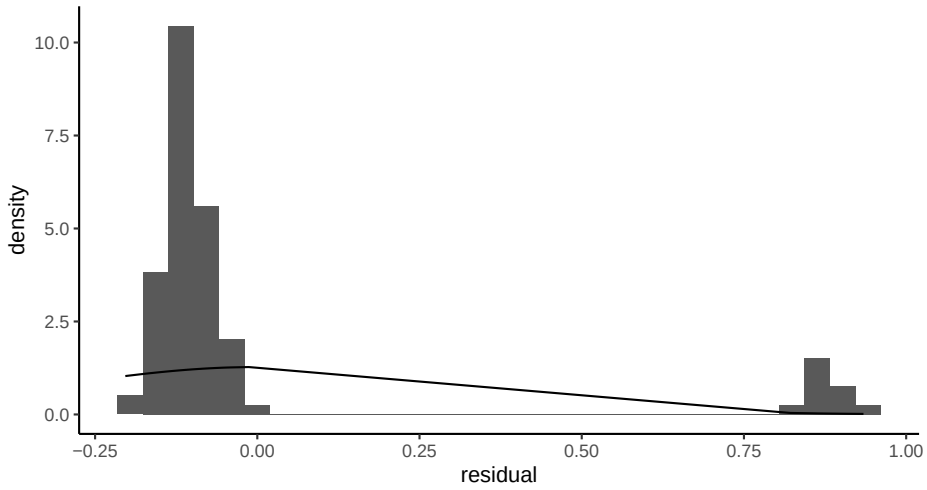


Figure 15.3: Dichotomous data example where the best fitting normal distribution is not a good approximation of the distribution of the residuals.

15.2 Example data with dichotomous outcome

Imagine that we analyse results on a simple reading test for children in first grade. These children are usually either 6 or 7 years old, depending on what month they were born in. The test is every year around 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 collects data in a random sample of first grade pupils. She computes the age of the child in months. Each child that passes the test gets a score of 1 and all the others get a score of 0. Figure 15.4 plots the data.

She wants to use the following linear model:

$$\texttt{score} = b_0 + b_1 \texttt{age} + e$$
$$e \sim N(0, \sigma_e^2)$$

Figure 15.5 shows the data with the estimated regression line and Figure 15.6 shows the distribution of the residuals as a function of **age**.

Clearly, a linear model is not appropriate. The assumption of the linear model is that the residuals are scattered randomly around 0. This is not the case here as we see a clear pattern in the residuals. The main problem is that the dependent variable **score** has only two possible values: 0 and 1. When we have a dependent variable that is *dichotomous* or *binary* (only two values), we generally use *logistic regression*. The basic idea is that instead of assuming the normal distribution for the residuals, we assume the Bernoulli distribution.

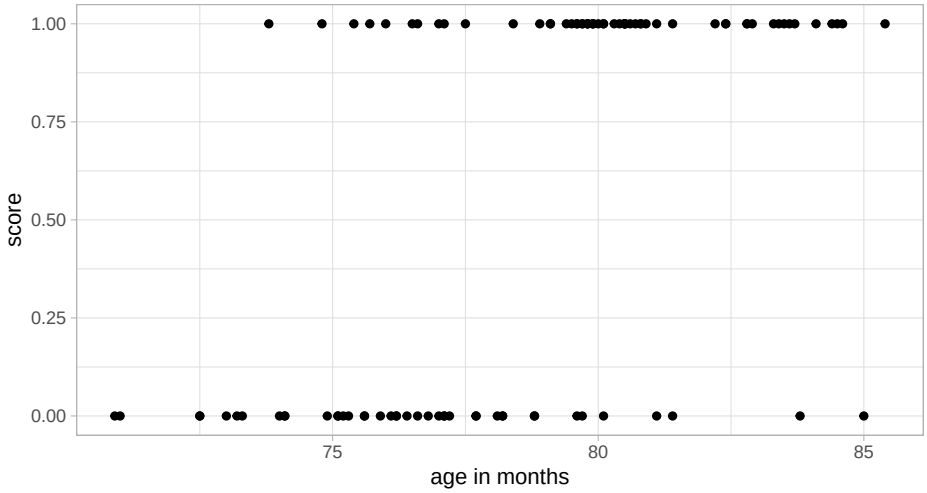


Figure 15.4: Data example: Test outcome (score) as a function of age, where 1 means pass and 0 means fail.

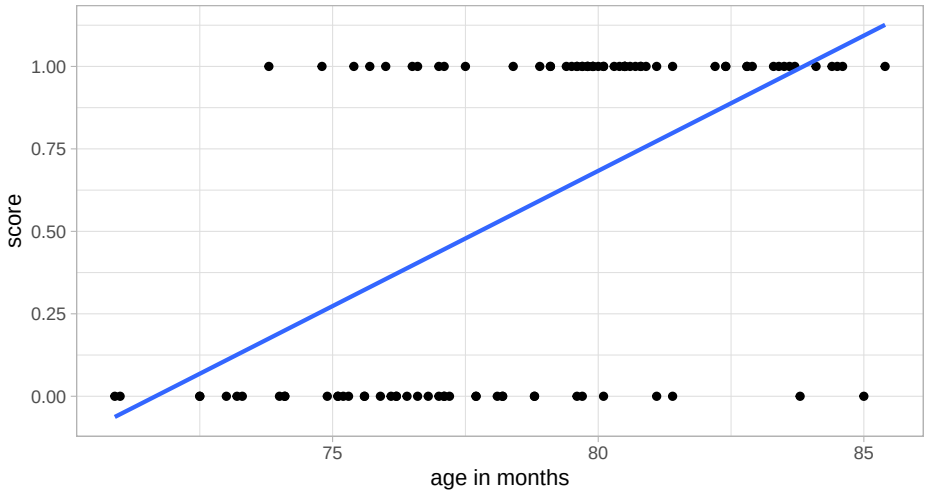


Figure 15.5: Example test data with a linear regression line.

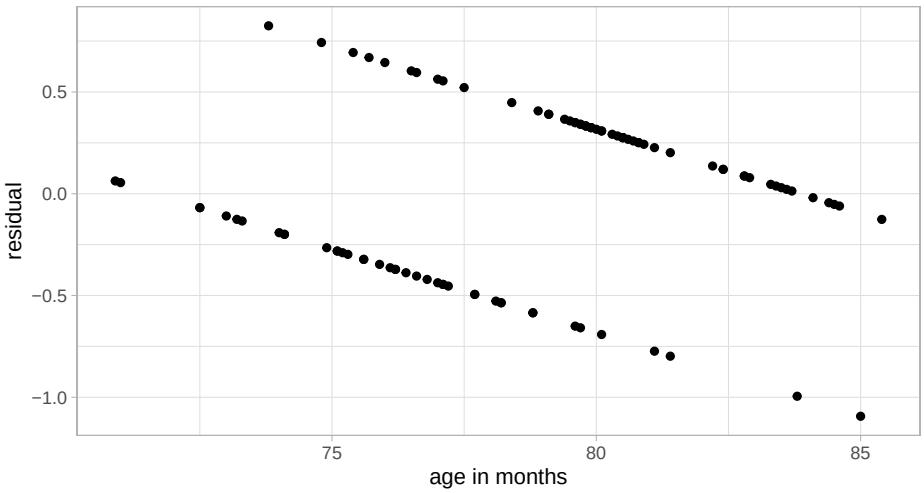


Figure 15.6: Residuals as a function of age, after a linear regression analysis of the test data.

15.3 Alternative: the Bernoulli distribution

Before we introduce the Bernoulli distribution, we go back to the regular linear model and rewrite it a bit. First we write the linear equation that gives us the expected value of the dependent variable, $\hat{Y}$.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 \tag{15.1}$$

Next, we state that the actual observed value, Y, depends on the expected or predicted value $\hat{Y}$, and is normally distributed around that expected value, with a a certain variance σ^2:

$$Y \sim N(\hat{Y}, \sigma^2) \tag{15.2}$$

In this case, we omit the residual e, but of course it is still there in the form of the difference between the observed value Y and the expected value $\hat{Y}$, and this difference will have a normal distribution. The important thing to note here is that we split the linear model into two parts: a linear part, consisting of the linear equation to model $\hat{Y}$, and the distribution part, consisting of a normal distribution for the observed value Y.

This division into linear equation and distribution is the basic idea of generalised linear models. These models are all linear in that they all have the same type of linear equation part, but they differ in what kind of distribution they use. Here we model Y has having a normal distribution. Instead, we could use a

distribution that is more suitable for a Y variable that only has two outcomes: the Bernoulli distribution.

A typical example where the Bernoulli distribution is appropriate is the flip of a coin. For a coin flip there are only two possible outcomes: heads or tails. For a fair coin, the probability of heads is 50%. That means that if we flip the coin 100 times, we expect to see heads in 50% of the cases. If the coin is not fair, for example when the probability of heads is 0.4, we can expect that if we flip the coin 100 times, on average we expect to see 40 times heads and 60 times tails. Our best bet then is that the outcome is tails. However, if we actually flip the coin, we might see heads anyway. Even though tails is more likely than heads, we never know when we will observe heads and when we will observe tails. There is some randomness. This is the kind of randomness that is modelled with the Bernoulli distribution. Let Y be the outcome of a coin flip: heads or tails. We model a Bernoulli distribution for variable Y with probability p for heads, in the following way:

$$Y \sim Bern(p_{heads})$$

If we regard Y to be the outcome of a coin flip, the p_{heads} parameter tells us how often on average we can see heads rather than tails.

The same kind of randomness and unpredictability 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, $\hat{Y}$, but we *observe* a value for Y that is usually different from $\hat{Y}$. This difference between predicted and observed is the residual. For a whole group of observations with the same predicted value for Y, we know that the whole group of data points will show normal distribution around this predicted value, but we're completely unsure what the residual will be for each individual data point.

$$Y \sim N(\hat{Y}, \sigma_e^2)$$

In our example of passing the test by the first graders, passing the test or not 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 test on the basis of age? And then for every predicted probability, we allow for the fact that actually the observed success for each individual child can differ. Our linear model could then look like this:

$$p_i = b_0 + b_1 \texttt{age}_i \tag{15.3}$$

$$\texttt{score}_i \sim Bern(p_i) \tag{15.4}$$

So for each child i, we predict the probability of passing the test, p_i, on the basis of their age. Next, the randomness in the data comes from the fact that a

probability is only a probability, so that the observed passing or failing of a child $\texttt{score}_i$, is like a coin toss with probability of p_i for success. This can account for the observation that children with the exact same age, can still score differently on the test.

If we would apply such a model to the data, we end up with the following equation:

$$p = -5.876 + 0.082 \times \texttt{age}$$

This equation means that for a child with an age of 80 months, we have a probability of passing the test equal to $-5.876 + 0.082 \times 80 = 0.68$. Now let's look at the prediction for a child aged 71 months (a month before their sixth birthday). We then end up with $-5.876 + 0.082 \times 71 = -0.05$. This can't however be a probability: a probability is defined to be a number between 0 and 1. A similar problem we get when we do the calculations for a child aged 84 months (shortly after their seventh birthday): $-5.876 + 0.082 \times 84 = 1.02$. Again, a value that can't be a probability.

Therefore, in general, a linear equation that yields probabilities directly does not work. With a linear equation we can always end up with values less than 0 and more than 1, which can't be probabilities. That means that in addition to a Bernoulli distribution, we need a trick that ensures we always get a probability between 0 and 1. The trick that is used is using *log-odds*.

15.4 Log-odds

To solve the problem with probabilities is that instead of using probabilities in the linear equation, we use *log-odds*. Log-odds are defined as the natural logarithm of the odds. First we briefly explain what odds are, and next we briefly explain what the logarithm is.

15.4.1 Odds

When talking about probabilities, one often also talks about odds, most often in the context of sports and gambling. In the context of our example data, the odds of passing the test is the ratio of the probability passing a test, and the probability of *not* passing the test.

$$\texttt{odds}_{pass} = \frac{p_{pass}}{p_{fail}} \tag{15.5}$$

In general, we calculate the odds by dividing the probability p by its *complement*:

$$\texttt{odds} = \frac{p}{1-p} \tag{15.6}$$

More on odds

Odds are an alternative way of expressing the likelihood of a particular event. They are used a lot in betting and gambling settings. Take for example the rolling of a dice. The *probability* of rolling a 6 is equal to $\frac{1}{6}$ because there are 6 different types of outcome, and only one of them is "rolling a 6". We say the probability is "one out of six". For probabilities therefore, we calculate the number of events that lead to success and we divide by the *total* number of events.

In contrast, with odds, we count the number of outcomes that lead to success and we divide that by the number of outcomes that lead to failure. There is one outcome for success ("6"), and five outcomes for failure ("1, 2, 3, 4, or 5"). The odds ratio is then "one to five", which is usually written as 1:5, but is equal to $\frac{1}{5}$.

Instead of counting events you can also use the probability directly to come up with odds. The probability of a six is $\frac{1}{6}$ and the probability of not a six is $\frac{5}{6}$. The odds can then be computed by taking the ratio of these two probabilities: $\frac{1/6}{5/6} = \frac{1}{5}$. In general, the odds of an event is the probability of that event p divided by probability of that event *not* happening, $q = 1 - p$,

$$\texttt{odds} = \frac{p}{1-p} \tag{15.7}$$

Suppose the probability of winning the lottery is 1%. Then the probability of losing is 99%. A probability of 1% means that if I play the lottery a total of 100 times, I expect to win 1 time and lose 99 times. That means once in a total of 100 times: $\frac{1}{100}$. In contrast, the odds are the number of events that are successful divided by the number of events that are unsuccessful, therefore 1:99, or $\frac{1}{99}$.

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 (success), there will be 1 tails (failure). In other words, if there were 11 coin tosses in total, 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 $1/11 = 0.09$.

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 $\frac{4}{9} = 0.44$.

When transforming odds into probabilities, you can use the following equation:

$$p = \frac{\texttt{odds}}{\texttt{odds} + 1} \tag{15.8}$$

15.4.2 Natural logarithm

The *logarithm* is closely related to the concept of the *exponent*. The exponent tells us how many times to use a number in a multiplication. In 4^2 the exponent "2" says to use 4 twice in a multiplication, so $4^2 = 4 \times 4 = 16$. Similarly, 4^3 means we have to use 4 three times: $4^3 = 4 \times 4 \times 4 = 64$.

When using the logarithm we go in reverse: How many of one number do we have to use in a multiplication to make another number. For example, if we want to know how often we have to use 4 in a multiplication to obtain 16, this is the same as asking about the logarithm of 16 with base 4, $\texttt{log}_4(16)$. We know that if we multiply 4×4, we obtain 16. Therefore,

$$\texttt{log}_4(16) = 2 \tag{15.9}$$

Instead of base 4, we can also use other numbers. For example, we can use base 2:

$$\texttt{log}_2(16) = 4 \tag{15.10}$$

because if we multiply 2 four times, we obtain 16: $2 \times 2 \times 2 \times 2 = 2^4 = 16$.

A special case is using the number e as the base of our logarithm. The value of e is called Euler's number and is about 2.7. When using e as the base we often write ln instead of $\log_e$. For example,

$$\log_e(16) = \texttt{ln}(16) \approx 2.77 \tag{15.11}$$

because if we raise e to the power 2.77, we get 16: $e^{2.77} \approx 16$. An alternative way of writing $e^{2.77}$ is $\exp(2.77)$.

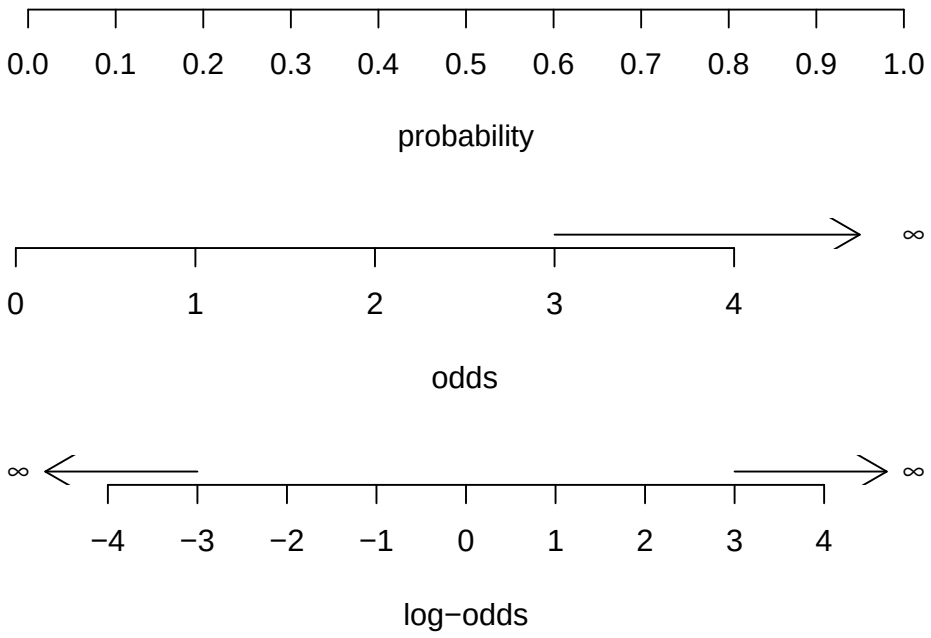
When using e as the base of the logarithm, we call it the *natural* logarithm.

15.4.3 Transforming probability into log-odds

A probability p is always between 0 and 1. If we transform a probability p into odds, we always end up with a number somewhere between 0 and infinity. For example if we use a small probability like 0.1, the corresponding odds is $\frac{0.1}{0.9} \approx 0.11$. If we use a large probability like 0.9, the corresponding odds is $\frac{0.9}{0.1} \approx 9$.

If we next take the *natural logarithm of the odds*, we end up with values that can be both positive and negative, in fact all values between $-\infty$ and $+\infty$ are possible. For example, with a probability of 0.1, the odds are 0.11 and the natural logarithm of 0.11 equals $\ln(0.11) = -2.2$. With a probability of 0.9, the odds are 9 and the natural logarithm of 9 equals $\ln(9) = +2.2$. The natural logarithm of the odds is called the *log-odds*.

The three different scales are illustrated below. A probability close to 0 is equivalent to a log-odds close to $-\infty$, and a probability close to 1 is equivalent to a log-odds close to $+\infty$. A probability of 0.5 is equivalent to a log-odds of 0.



The transformation of a probability into log-odds, by first calculating odds and then taking the natural logarithm, is called the so-called *logit function*:

$$\texttt{logit}(p) = \ln \frac{p}{1-p}$$

This logit function is plotted in Figure 15.7.

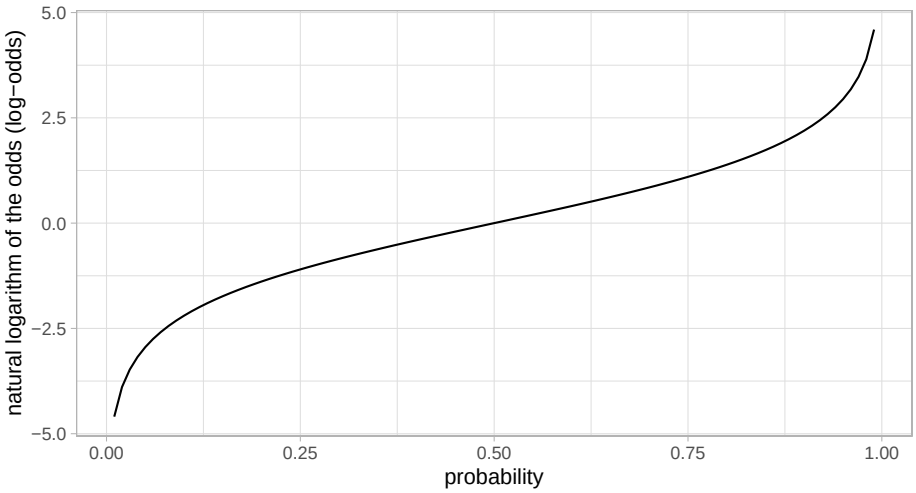


Figure 15.7: The logit function transforms probabilities into log-ods.

Because log-odds can be any value between $-\infty$ and $+\infty$, they are very useful for linear equations. It is therefore in logistic regression that instead of probabilities we have a linear equation for the log-odds.

15.5 Logistic link function

In previous pages we have seen that log-odds 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 test data in children might then look like this:

$$\begin{aligned} \texttt{log-odds}_{pass} &= b_0 + b_1 \texttt{age} \\ Y &\sim Bern(p_{pass}) \end{aligned}$$

Note the strange relationship between the probability parameter p_{pass} for the Bernoulli distribution, and the dependent variable for the linear equation `log-odds`. The linear model predicts the log-odds, but for the Bernoulli distribution, we use the probability.

Note that earlier, we learned how a probability can be transformed into a log-odds using the logit function $\texttt{log-odds} = \ln \frac{p}{1-p}$. We also need a way to translate the log-odds that we compute using the linear equation back into probabilities. For that we use the so-called *logistic* function.

$$p = \frac{e^{\texttt{log-odds}}}{1 + e^{\texttt{log-odds}}}$$

Here we see Euler's number e again. As e^x can also be written as $\exp(x)$, we can therefore write

$$p = \frac{\exp(\texttt{log-odds})}{1 + \exp(\texttt{log-odds})}$$

The logistic function is plotted in Figure 15.8. Its shape is S-like and resembles the cumulative normal distribution closely (see Ch. 1).

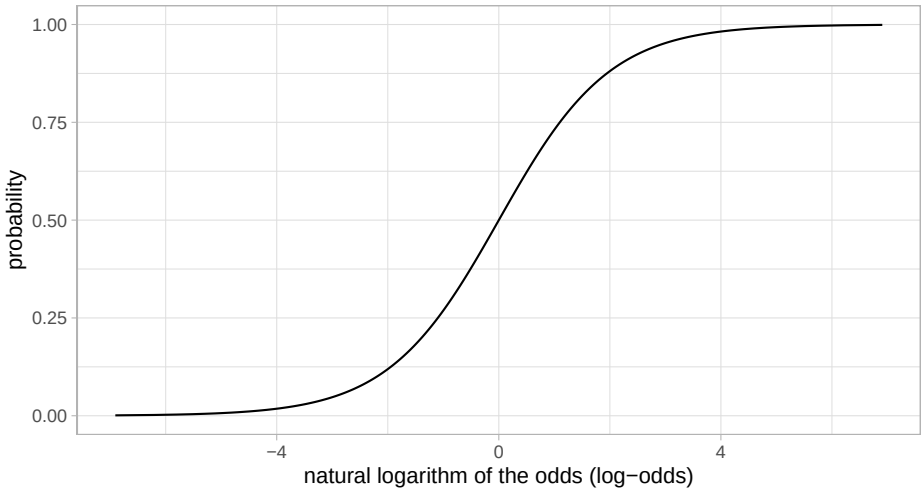


Figure 15.8: The relationship between the natural logarithm of the odds and the probability.

Overview probabilities, odds and log-odds

From probability p to odds, use

$$\texttt{odds} = \frac{p}{1-p} \tag{15.12}$$

From probability to log-odds, use the logit function:

$$\texttt{log-odds} = \text{logit}(p) = \ln\frac{p}{1-p} \tag{15.13}$$

From log-odds to back to p, use the logistic function:

$$p = \frac{\exp(\texttt{log-odds})}{1+\exp(\texttt{log-odds})} \tag{15.14}$$

15.6 Logistic regression applied to example data

So what does it look like, a linear model for log-odds?

In Figure 15.9 we show the linear model for the log-odds of passing the test for the first-graders. These log-odds are predicted by age using a straight, linear regression line.

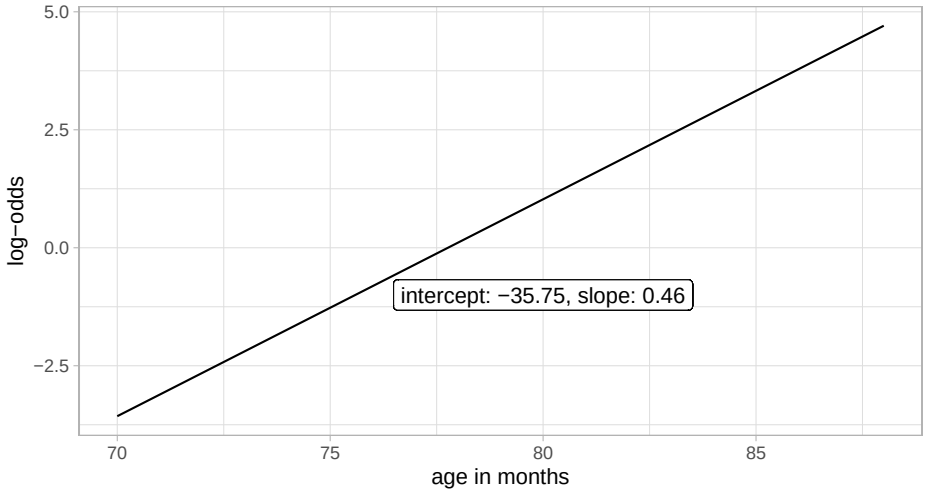


Figure 15.9: Example of a linear model for the log-odds of passing the test.

When we take all these predicted log-odds and convert them back into probabilities using the logistic function, we obtain the plot in Figure 15.10.

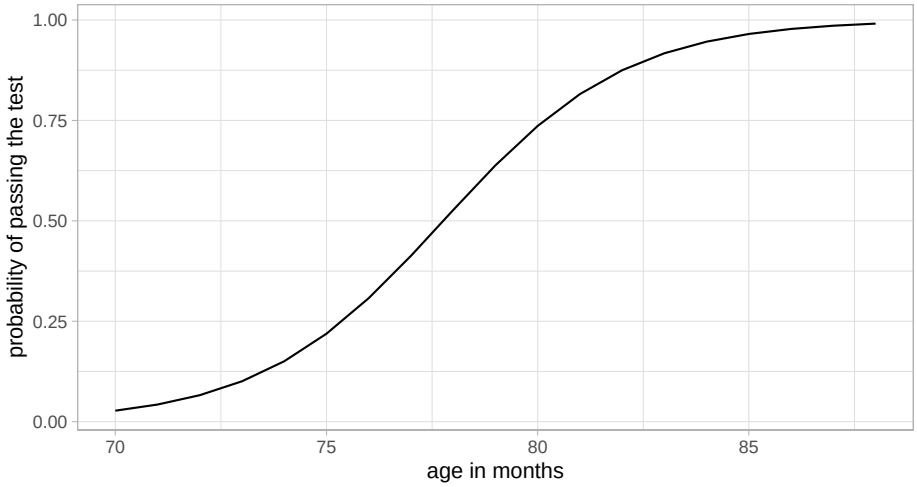


Figure 15.10: Example with log-odds transformed into probabilties (vertical axis).

Here we see an S-shape relationship between **age** and the probability. It has the same shape as the logistic function in Figure 15.8. We see that our model predicts probabilities close to 0 for very young ages, and probabilities close to 1 for relatively old ages. There is a clear positive effect of age on the probability of passing the test. Note that the relationship is not linear on the scale of the probabilities, see Figure 15.10, but linear on the scale of the logit of the probabilities (the log-odds), see Figure 15.9.

The curvilinear shape we see in Figure 15.10 is a logistic function of **age** and can be described as:

$$p = \text{logistic}(b_0 + b_1\texttt{age}) = \frac{\exp(b_0 + b_1\texttt{age})}{1 + \exp(b_0 + b_1\texttt{age})} \tag{15.15}$$

$$= \frac{\exp(-35.75 + 0.46 \times \texttt{age})}{1 + \exp(-35.75 + 0.46 \times \texttt{age})} \tag{15.16}$$

In summary, if we go from log-odds to probabilities, we use the logistic function, $\text{logistic}(x) = \frac{\exp(x)}{1+\exp(x)}$. If we go from probabilities to log-odds, 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 first-grade children that either passed or failed the test, we see that this curve gives a description of our data, see Figure 15.11. The model predicts that around the age of 78 months, the probability of passing the test is around 0.50. We indeed see in Figure 15.11 that around this age about as many children pass the test (**score** = 1) as children that don't (**score** = 0). We see that for younger ages, say around the age of 72 months, more children fail the test than pass the test. This fits with a lower probability of passing the test for such an age. For older children, say around 84 months, we see more children passing the test than failing the test, again in line with the predicted high probability.

On the basis of this analysis there seems to be a positive relationship between age in first-grade children and the probability of passing the test in this sample.

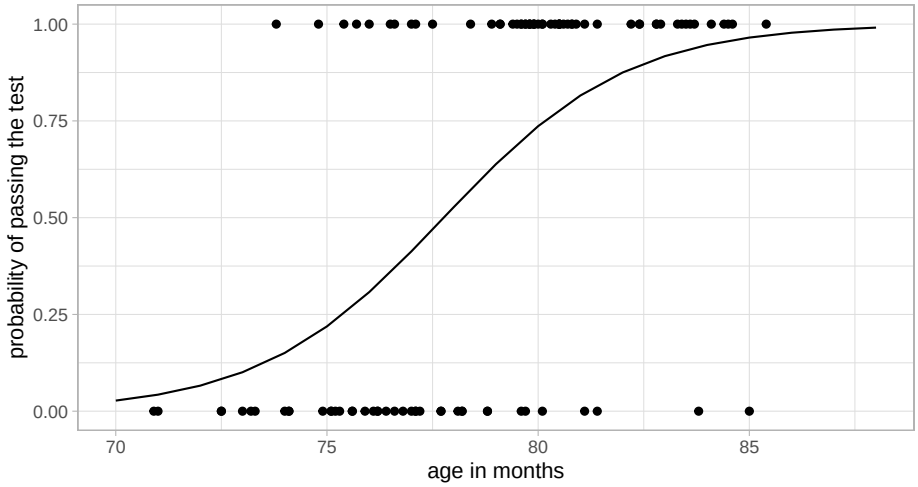


Figure 15.11: Transformed regression line and raw data points.

What we have done here is a *logistic regression* of passing the test on age. It is called logistic because the curve in Figure 15.11 has a logistic shape. Logistic regression is one specific form of a *generalised linear model*. Logistic regression is a generalised linear model with a Bernoulli distribution and a so-called *logit link function*: instead of modelling the probability directly, we have modelled *the logit of the probabilities of obtaining a Y-value of 1* (the log-odds).

Overview

1) In a regular linear model for continuous outcomes, an equation gives the predicted Y-value, and the observed Y-values are normally distributed around $\hat{Y}$.
2) For dependent variables that have only two possible outcomes, usually a Bernoulli distribution is used instead of a normal distribution.
3) The probability cannot be modelled with an equation directly because probability values need to be between 0 and 1.
4) Instead of the probablity p, a transformation of the probability is used: the natural logarithm of the odds, also called the logit function $\texttt{log-odds} = \text{logit}(p) = \ln\frac{p}{1-p}$. We then get the equation:

$$\texttt{log-odds} = b_0 + b_1 X_1 + b_2 X_2$$

5) A link is needed between the log-odds that is predicted by the equation and the probability p for the Bernoulli distribution $Y \sim Bern(p)$. For that we use a logistic transformation:

$$p = \frac{\exp(\texttt{log-odds})}{1 + \exp(\texttt{log-odds})} \tag{15.17}$$

In the next section we will see how logistic regression can be performed in R, and how we should interpret the output.

The linear model model as a special case

The regular linear model can be seen as a generalised linear model. Any generalised model has three ingredients:

1) a *linear equation* to model predictions,
2) a *distribution* for the actual observed outcome,
3) a *link function* between what is predicted and the distribution.

For logistic regression we have a linear equation that predicts log-odds, and we have a Bernoulli distribution with a logit link function between the log-odds and the p-parameter of the Bernoulli distribution. For the regular linear model, we have a linear equation that gives us $\hat{Y}$, and we use a normal distribution $N(\hat{Y}, \sigma^2)$. Since both the equation and the distribution refer to the same $\hat{Y}$, the link is very simple: the link function is $\hat{Y} = \hat{Y}$. This is referred to as an *identity* link function ($\hat{Y}$ is identical to $\hat{Y}$). The regular linear models from previous chapters are therefore special cases of generalised linear models: generalised linear models with normal distributions and an identity link.

Table 15.1: Imaginary test results first-graders and their age in months (part of the data).

result	age
pass	80
fail	80
fail	78
fail	77
pass	80
fail	72

Besides the logit link function and the identity link function, there are several other link functions possible. One of them, the exponential link, we will discuss in the next chapter on generalised linear models for count data (Ch. 16). Apart from the Bernoulli distribution and the normal distribution, there are also alternatives, like the Poisson distribution, also discussed in Chapter 16.

15.7 Logistic regression in R

Let's analyse the test results from the random sample of first-graders. We see part of the data in Table 15.1.

The dependent variable is **result**, and we want to study whether it is related to **age**. The dependent variable is dichotomous (two possible outcomes) so we choose logistic regression. The function that we use in R is the `glm()` function, which stands for Generalised Linear Model. We can use the following code:

```
model.test <- data.test %>%
  glm(result ~ age,
      data = .,
      family = binomial(link = logit))
```

```
## Error in eval(family$initialize): y values must be 0 <= y <= 1
```

We get an error message: R tells us that our dependent variable needs to have values between 0 and 1. We therefore create a dummy variable **score**, that is 1 if the child passed the test, and 0 if it did not pass:

```
data.test <- data.test %>%
  mutate(score = ifelse(result == "pass", 1, 0))
```

Next we run the logistic regression using the dummy variable as the outcome variable:

```
model.test <- data.test %>%
  glm(score ~ age,
      data = .,
      family = binomial(link = logit))
```

Note in the code that we specify that we want to use the *binomial* distribution with a logit link function. But why do we tell R to use a binomial distribution when actually we want to use a Bernoulli? 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.test <- data.test %>%
  glm(score ~ age,
      data = .,
      family = binomial)
```

Below, we see the parameter estimates from this generalised linear model run on the test data.

```
model.test %>%
  tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic    p.value
##   <chr>          <dbl>     <dbl>     <dbl>      <dbl>
## 1 (Intercept)  -35.7      7.58       -4.71 0.00000243
## 2 age            0.460    0.0968      4.75 0.00000203
```

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, 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:

$$\begin{aligned} \texttt{log-odds}_{\texttt{pass}} &= b_0 + b_1 \texttt{age} \\ \texttt{score} &\sim Bern(p_{\texttt{pass}}) \end{aligned}$$

If we fill in the values from the R output, we get

$$\begin{aligned} \texttt{log-odds}_{\texttt{pass}} &= -35.7 + 0.46 \times \texttt{age} \\ \texttt{score} &\sim Bern(p_{\texttt{pass}}) \end{aligned}$$

We can interpret these results by making some predictions. Imagine a child with an age of 72 months (on its sixth birthday). Then the predicted log-odds equals $-35.7 + 0.46 \times 72 = -2.58$. When we transform the log-odds back to a probability,

```
exp(-2.58)/(1+ exp(-2.58))
```

```
## [1] 0.07043673
```

we get $\frac{\exp(-2.58)}{1+\exp(-2.58)} = 0.07$. Note that the probability as well as the odds refer to the outcome that is coded as 1. Here we coded pass as "1" and fail as "0".

To make predictions using this model easier, we can use the `predict()` function. In order to get the predicted log-odds we can use

```
predict(model.test, newdata = data.frame(age = 72))
```

```
##         1
## -2.649915
```

Note that the result -2.649915 is different from our hand-calculated -2.58. The difference is the result of the rounding in the R model output.

If we want to predict probabilities instead of the log-odds, we can use

```
predict(model.test, newdata = data.frame(age = 72), type = "response")
```

```
##          1
## 0.06599423
```

So this model predicts that for children on their sixth birthday, about 7% will pass the test.

Now imagine a child on its seventh birthday (84 months): what does the model predict for the test result?

The predicted log-odds equals $-35.7 + 0.46 \times 84 = 2.94$. When we transform this back to a probability, we get $\frac{\exp(2.94)}{1+\exp(2.94)} = 0.95$.

Letting R do the work for us, we get the same probability:

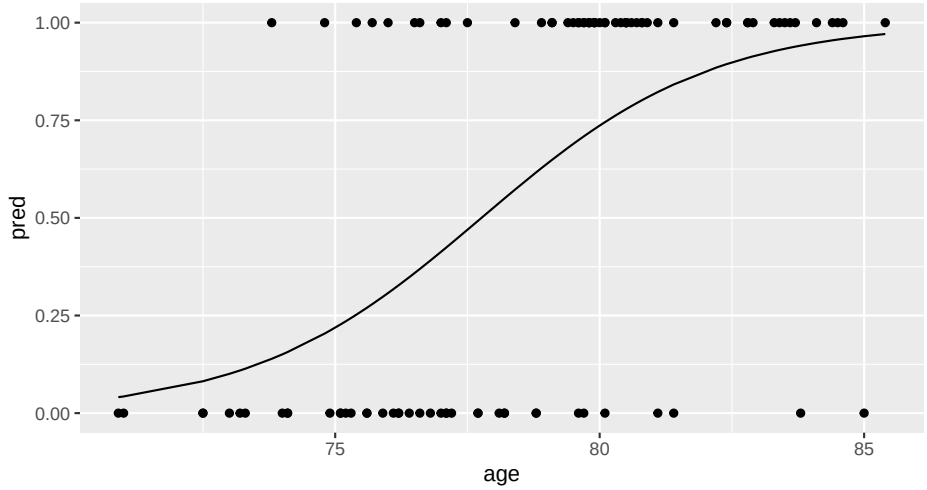
```
predict(model.test, newdata = data.frame(age = 84), type = "response")
```

```
##         1
## 0.9461369
```

We observe that this model predicts that older children have a higher probability that they pass the test than younger children. Six year olds have a very low probability of only 7%, and seven year olds have a high probability of 95%.

A visualisation is often more helpful than a lot of calculations. The relationship between age and the probability of passing the test an be visualised using `ggplot()` and the `add_predictions()` function from the `modelr` package:

```
data.test %>%
  add_predictions(model.test, type = "response") %>%
  ggplot(aes(x = age, y = pred)) +
  geom_line() +
  geom_point(aes(y = score)) # show data points
```



We found that the probability of passing the test is related to age in this data set of randomly selected children, but is there also an effect of age in the entire population of first-graders? The regression table shows us that the effect of age, +0.46, is statistically significant at an α of 5%, $z = 4.75, p = .00000203$.

> "In a logistic regression with dependent variable passing the test and independent variable age, the null-hypothesis was tested that age is not related to passing the test. Results showed that the null-hypothesis could be rejected, $b_{age} = 0.46, z = 4.75, p < 0.001$. We conclude that in the population of first graders, a higher age is associated with a higher probability of passing the test."

Note that similar to other linear models, the intercept can be interpreted as the predicted log-odds for children that have values 0 for all other variables in the model. Therefore, an intercept of -35.7 means in this case that the predicted log-odds for children with age 0 months equals -35.7. This is equivalent to a probability of $\frac{\exp(-35.7)}{1+\exp(-35.7)} \approx 0$.

15.8 Take-away points

- Logistic regression is in place when the dependent variable is dichotomous (yes/no, 1/0, TRUE/FALSE).
- Logistic regression is a form of a generalised linear model.
- Any generalised model has three properties: 1) a *linear equation* to model predictions, 2) a *distribution* for the actual observed outcome, and 3) a *link function* between what is predicted and the distribution.
- In a regular linear model for continuous outcomes, an equation gives the predicted Y-value, and the observed Y-values are normally distributed around $\hat{Y}$.
- For dependent variables that have only two possible outcomes, usually a Bernoulli distribution is used instead of a normal distribution. A Bernoulli distribution depends on probability p.
- The probability cannot be modelled with an equation directly because the values need to be between 0 and 1.
- Instead of the probablity p, a transformation of the probability is used, the natural log of the odds: $\texttt{log-odds} = \ln\frac{p}{1-p}$. This is called the logit function. We then get the equation:

$$\texttt{log-odds} = b_0 + b_1\texttt{X}$$

- A link is needed between the log-odds that is predicted by the equation and the probability p for the Bernoulli distribution. We call that the logit link function. We can transform a log-odds into a probability using the logistic function:

$$p = \frac{\exp(\texttt{log-odds})}{1 + \exp(\texttt{log-odds})}$$

Key concepts

- Logit
- Bernoulli distribution
- Odds
- Oddsratio
- log-odds
- Logistic link function

Chapter 16

Generalised linear models for count data: Poisson regression

16.1 Poisson regression

Count data are very typical in that they are always positive, and inherently discrete. Often when using linear models on count data, we see non-normal distributions of residuals. A better way to handle count data is using the generalised linear model. In the previous chapter we discussed a generalised linear model for situations where the dependent variable consists of zeros and ones (binary or dichotomous data): the logistic regression model. In this chapter we focus on the Poisson version of the generalised linear model that is appropriate when the dependent variable is a count variable. To look at some example data, we go to the movies.

For every weekend, the British Film Institute publishes the box office figures on the top 15 films released in the UK, all other British releases and newly released films. Table 16.1 shows part of the data from the last weekend of November, 2023[1].

The films in Table 16.1 are ranked in such a way that the top film (rank 1) is the film with the highest weekend gross revenue. The second film (rank 2) is the film with the second highest weekend gross revenue. See Chapter 8 on a discussion of ranks.

[1]Data downloadable from https://www.bfi.org.uk/education-research/film-industry-statistics-research/weekend-box-office-figures

Table 16.1: Box office data last weekend of November 2023. The films are ranked in terms of weekend gross revenue.

Rank	Film	Country of Origin	Number of cinemas	Weekend Gross
1	Napoleon	UK/USA	716	5235706
2	The Hunger Games: The Ballad Of Songbirds And Snakes	USA	663	2689643
3	Wish	USA	617	2432228
4	Saltburn	UK/USA	477	572728
5	The Marvels	UK/USA	556	485099
6	Cliff Richard: The Blue Sapphire Tour 2023 (Concert)	UK	411	329826

Here we model in how many cinemas a film is shown, as a function of the variable **rank**. The relationship is shown in Figure 16.1. It seems there is a relation between how much money a film generates and the number of cinemas that show it. However, the relationship is clearly not linear: a simple regression model will not work here.

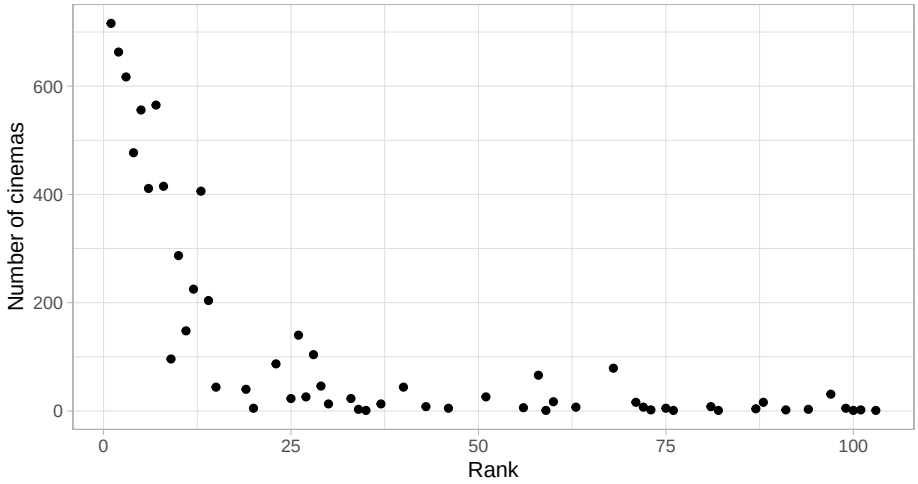


Figure 16.1: The number of cinemas showing a film in a particular weekend, as a function of the film's rank in terms of gross weekend revenue.

As the dependent variable is a count variable, we choose a Poisson model.

Whereas the normal distribution has two parameters, the mean and the variance, the Poisson distribution has only one parameter, λ (Greek letter 'lambda'). λ is a parameter that indicates *tendency*. Figure 16.2 shows a Poisson distribution with a tendency of 2.

What we see is that the values tend to cluster around the lambda parameter value of 2 (therefore we call λ a tendency parameter). We see only discrete values, and no values below 0. The distribution is not symmetrical, and we

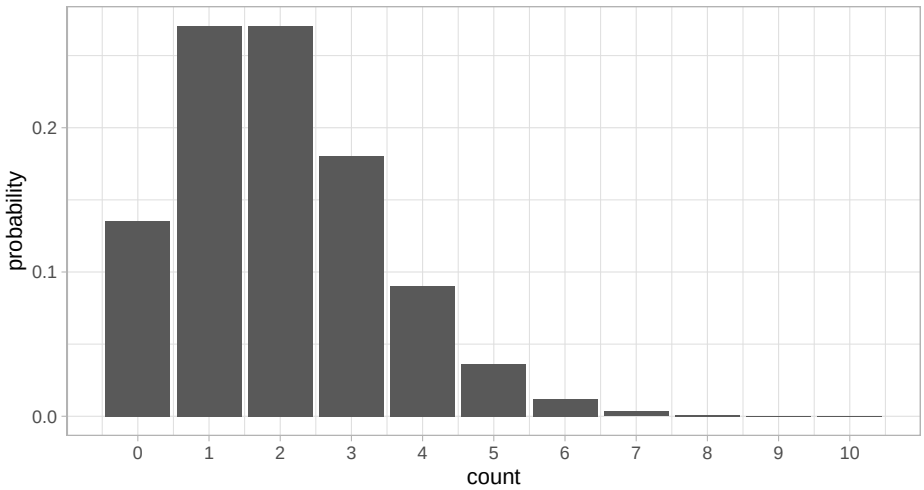


Figure 16.2: The Poisson probability distribution with lambda = 2.

see a few extreme values such as those higher than 6. If we would take the mean of the distribution, we would find a value of 2. If we would compute the variance of the distribution we would also find 2. In general, if we have a Poisson distribution with a tendency parameter λ, we know from theory that both the mean and the variance will be equal to λ.

A Poisson model could be suitable for our data: a linear equation could predict the parameter λ and then the actual data show a Poisson distribution.

$$\lambda = b_0 + b_1 X$$
$$Y \sim Poisson(\lambda)$$

However, because of the additivity assumption, the equation $b_0 + b_1 X = \lambda$ can lead to negative values for λ if either b_0 or b_1 is negative, and/or if X has negative values. A negative value for λ is not logical, because we then would have a tendency to observe values like -2 and -4 in our data, which is contrary to having count data. A Poisson distribution always shows integers of at least 0, so one way or another we have to make sure that we always have a λ of at least 0.

Remember that we saw the reverse problem with logistic regression: there we wanted to have the possibility of negative values for our dependent variable logodds ratio, so therefore we used the logarithm. Here we want to always have positive values for our dependent variable, so we can use the inverse of the logarithm function: the exponential. Then we have the following model:

$$\lambda = \exp(b_0 + b_1 X) = e^{b_0 + b_1 X}$$
$$Y \sim Poisson(\lambda)$$

This is a generalised linear model, now with a Poisson distribution and an exponential link function. The exponential function makes any value positive, for instance $\exp(0) = 1$ and $\exp(-10) = 0.00005$, so that we always have a positive λ parameter.

Let's analyse the film revenues data with this generalised linear model. Our dependent variable is **Number of cinemas**, that is, the number of cinemas that are running a particular film, and the independent variable is **Rank**, which indicates the relative amount of weekend gross revenu for that film (rank 1 meaning the most money). When we run the analysis, the result is as follows:

$$\begin{aligned} \lambda &= \exp(6.450 - 0.075 \times \texttt{Rank}) \\ \texttt{N}_{cinemas} &\sim Poisson(\lambda) \end{aligned}$$

What does it mean? Well, similar to logistic regression, we can understand such equations by making some predictions for interesting values of the independent variable. For instance, a value of 1 for **Rank** means that we are looking at the film with the largest gross revenu. If we fill in that value, we get the equation $\lambda = \exp(6.450 - 0.075 \times 1) = \exp(6.375) = 587$. Thus, for the top film of the week, we expect to see that the film runs in 587 cinemas.

Another value of **Rank** might be 10, representing the film with the 10th largest gross revenue. If we fill in that value, we get: $\lambda = \exp(6.450 - 0.075 \times 10) = \exp(5.70) = 299$. Thus, for the film with the 10th largest revenue, we expect to see that it runs in 299 cinemas.

If we look at the pattern in these expected scores, we see that the number of cinemas is negatively related to the gross revenue. That is what we find in this data set. In the next section we will see how to perform the analysis in R.

16.2 Poisson regression in R

Poisson regression is a form of a generalised linear model analysis, similar to logistic regression, so we can use the `glm()` function to analyse the cinema data. Instead of using a Bernoulli distribution, we use a Poisson distribution. The code is as follows.

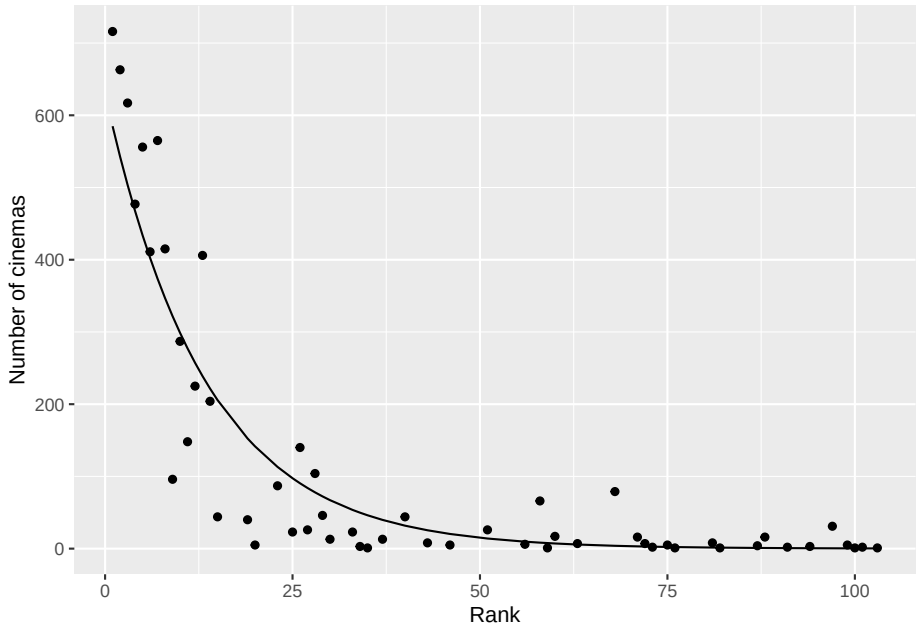
```
model <- bfi %>%
  glm(`Number of cinemas` ~ Rank, family = poisson, data = .)
model %>% tidy(conf.int = T)
```

```
## # 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)   6.45     0.0173      372.        0   6.41      6.48
## 2 Rank         -0.0747   0.00114     -65.7       0  -0.0769   -0.0725
```

We see the same values for the intercept and the effect of **Rank** as in the previous section. We now also see 95% confidence intervals for these parameter values.

For interpretation and checking model assumptions, it is a good idea to make a visualisation of the model. We can let R plot the predicted counts next to the observed data using the following code:

```
bfi %>%
  add_predictions(model, type = "response") %>%
  ggplot(aes(x = Rank, y = `Number of cinemas`)) +
  geom_point() +
  geom_line(aes(y = pred))
```



Note that the line with the predicted counts is curvy. The model equation $b_0 + b_1\text{Rank}$ is linear, that's why we call the Poisson model a linear model, but through the exponential link function, $\lambda = \exp(b_0 + b_1\text{Rank})$, we get a line that is not straight anymore.

Do we think the model is a good description of the data? It seems overall rather OK. We only see that for the highest ranked films (the first six films or so), the predicted numbers are generally too high (on the line or above it), and for the films that come next (ranks 7 to 25 or so), the predicted numbers are generally too low (all under the line). With a perfect model, the dots would all be on the line, or randomly distributed around. Some deviation from randomness is observed here.

Going back to the regression table, we see that for **Rank**, the confidence interval runs from -0.077 to -0.073. The value 0 is clearly not included in the confidence interval. We also see a test statistic for **Rank**, which is a z-statistic similar to that seen in logistic regression. Remember that the z-statistic is computed by b/SE. For large enough samples, the z-statistic follows a standard normal distribution. From that distribution we know that a z-value of -65.7 is significant at the 5% level. The output tells us it has an associated p-value of less than 0.001.

However, what should we do with this information? Remember that the statistical test for a null-hypothesis only works if you have a random sample from population data. The first question that we should ask ourselves is whether we have a random sample of data here. The second question is, if we have a random sample of data, from what population was it randomly sampled?

We already get stuck at the first question, since clearly we did not obtain a random sample of data. All our data came from the last weekend of November 2023. The data can clearly be used to say something about that weekend, as most of the films were included in the data, but can they be used to say something about other weekends? Possibly not. Other weekends see different numbers of visitors and different kinds of visitors. It makes a difference if we are looking at a dark and gloomy weekend in November, a sunny weekend in July, or a weekend during the Christmas holidays when whole families go to the cinema. Some weekends on end there may be a huge blockbuster, that might also significantly alter the pattern in the data. It is therefore safest, when reporting this analysis, to stick to a description of the data set itself, and not generalise.

We can write:

> "UK cinema data from 24-26 November 2023 were modelled using a generalised linear model with a Poisson distribution (Poisson regression), with independent variable rank and dependent variable the number of cinemas. The numbers of cinemas could reasonably be described by the equation $\lambda = \exp(6.45 - 0.075 \times \texttt{rank})$, where the films were ranked in terms of gross weekend revenu."

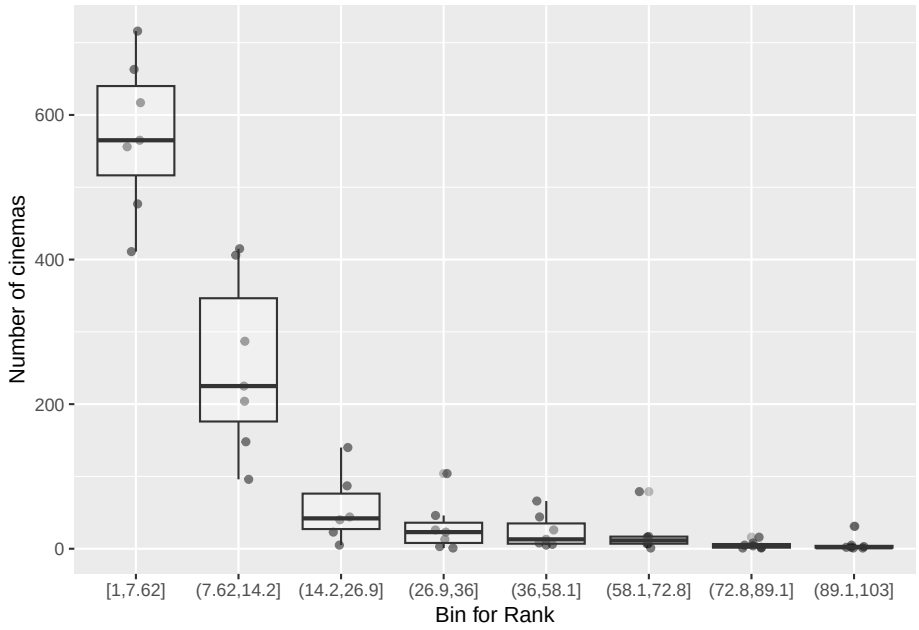
16.3 Overdispersion (advanced)

As we saw, with the Poisson distribution there is only one parameter, λ. That means that if we model data with a Poisson distribution, the variance of the residuals should be exactly the same as the expected number.

Let's look at the cinema data. We saw that a higher rank number was associated with a lower expected number of cinemas. Since a lower expected number should go together with a lower variance, according to the Poisson model, we should see that also in the data, otherwise the model is wrong. But also, the variance

in numbers should be more or less equal to the average number. Let's look at the distribution of the data, by first binning the data. We put the 8-top ranking films in the first bin (group), the next 8 films in the second bin, etcetera. Then for every bin we plot the distribution.

```
bfi %>%
  mutate(bin = cut_number(Rank, n = 8, boundary = 1)) %>%
  ggplot(aes(x = bin, y = `Number of cinemas`)) +
  geom_jitter(width = .1, height = 0, alpha = 0.5)  +
  geom_boxplot(aes(x = bin), alpha = .3) +
  xlab("Bin for Rank")
```



We see that the variance is larger when the values are larger, and becomes smaller when numbers become smaller. However, do the variances actually resemble the expected numbers? For every bin of 8 films, we compute the mean and the variance.

```
bfi %>%
  mutate(bin = cut_number(Rank, n = 8, boundary = 1)) %>%
  group_by(bin) %>%
  summarise(var = var(`Number of cinemas`),
            mean = mean(`Number of cinemas`)) %>%
  dplyr::select(bin, mean, var)
```

```
## # A tibble: 8 x 3
```

```
##   bin           mean     var
##   <fct>        <dbl>   <dbl>
## 1 [1,7.62]    572.   11049.
## 2 (7.62,14.2] 254.   14942.
## 3 (14.2,26.9]  56.5   2421.
## 4 (26.9,36]    30.9   1275.
## 5 (36,58.1]    24      538.
## 6 (58.1,72.8]  21.2    839.
## 7 (72.8,89.1]   5.29    28.6
## 8 (89.1,103]    6.43   119.
```

For most of the bins, we see an average number that is lower than the corresponding variance. Clearly the variance does not match the mean. In statistics this is called overdispersion. There is a test to detect overdispersion in your data, available in the `performance` package.

```
library(performance)
check_overdispersion(model)
```

```
## # Overdispersion test
##
##        dispersion ratio =  108.223
##   Pearson's Chi-Squared = 5627.584
##                 p-value =  < 0.001
```

Extensive treatment of dispersion is outside the scope of this book. The interested reader is referred to more specialised resources on generalised linear models.

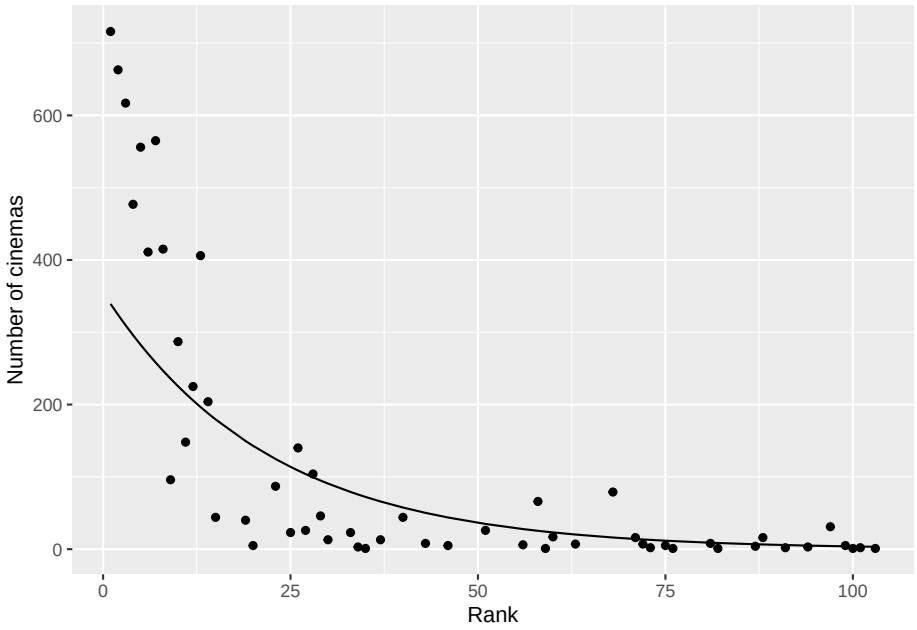
But one option to deal with violations of the Poisson model assumptions that we will mention here is to try out alternative models. A well-known alternative to the Poisson model is the negative binomial model. The negative binomial model is similar to the Poisson model but more flexible (the Poisson distribution is a special case of the more general negative binomial distribution). It can be estimated with the `glm.nb()` function from the `MASS` package.

```
library(MASS)
model2 <- bfi %>%
  glm.nb(`Number of cinemas` ~ Rank, data = .)
check_overdispersion(model2)
```

```
## # Overdispersion test
##
##        dispersion ratio =  1.321
##   Pearson's Chi-Squared = 68.693
##                 p-value =   0.06
```

The negative-binomial model solves most of the overdispersion problem as the dispersion ratio drops dramatically. Below we see the expected counts for this model. Comparing this figure with that for the Poisson model, we see that it makes different predictions, especially for the top-ranked films. The predictions seem actually much worse for the top-ranked films, as the model underestimates the number of cinemas for the top 8 films.

```
bfi %>% add_predictions(model2, type = "response") %>%
  ggplot(aes(x = Rank, y = `Number of cinemas`)) +
  geom_point() +
  geom_line(aes(y = pred))
```

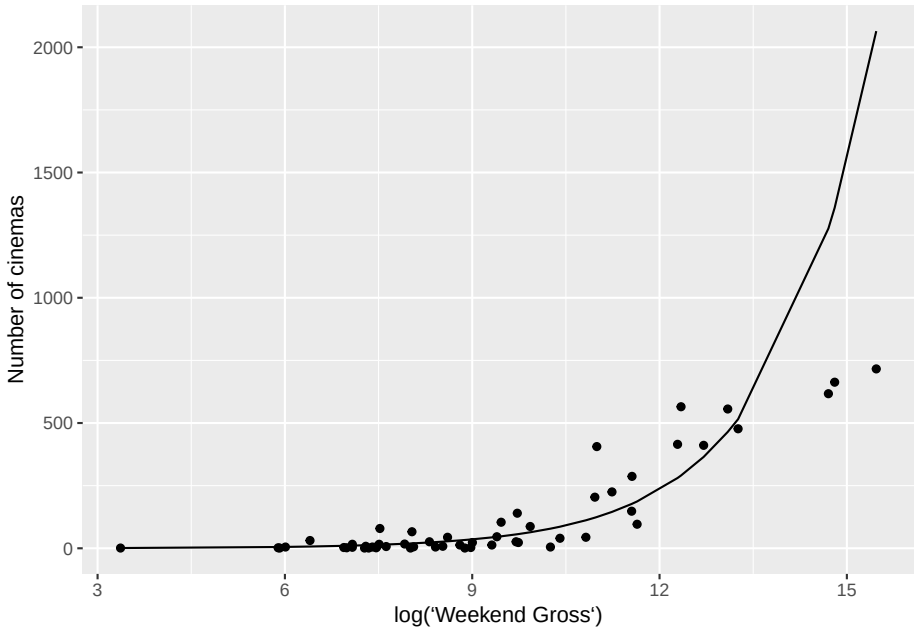


This problem is not easily solved. Instead of the ranks, we could work with the original data: the actual gross weekend revenue. Because this variable is extremely skewed (a few very large outliers), it makes sense to work with the logarithm of that variable and use that in the negative binomial model.

```
model3 <- bfi %>%
  glm.nb(`Number of cinemas` ~ log(`Weekend Gross`),
         data = .)

bfi %>%
  add_predictions(model3, type = "response") %>%
  ggplot(aes(x = log(`Weekend Gross`),
             y = `Number of cinemas`)) +
```

```
  geom_point() +
  geom_line(aes(y = pred))
```



The negative binomial model shows a good fit for most of the data. The model is sadly not able to predict the three top films correctly. This model seems better than the model that used the ranks, since there the top 8 films were badly predicted.

16.4 Association between two categorical variables

Poisson models are used when you count things, and when you are interested in how two or more variables are associated. Let's go back to Chapter 1. There we discussed the situation where we have two categorical variables, and you want to see how they co-vary. We made a crosstable that showed the counts of each combination of the levels of the categorical variables.

Imagine we do a study into the gender balance in companies. In one particularly large, imaginary company, there are about 100,000 employees. For a random sample of 1000 employees, we know their gender (female or male), and we know whether they have a position in senior management or not (leadership: yes or no). It turns out that there are 22 female senior managers, and only 8 male senior managers. If we would randomly pick a senior manager from this sample,

Table 16.2: The number of employees in a senior management role (no/yes), as a function of gender

	Leadership		
gender	**no**	**yes**	**Total**
female	778	22	800
male	192	8	200
Total	970	30	1000

they would more likely be a woman than a man, right? Can we therefore conclude that there is a positive gender bias, in that senior managers in this company are more often female? Think about it carefully, before you read on.

The problem is, you should first know how many men and how many women there are in the first place. Are there *relatively* many female leaders, given the gender distribution in this company? In Table 16.2 you find the crosstable of the data. From this table we see that there are 800 women in the sample and 200 men.

Thus, we see see more female leaders than male leaders, but we also see many more women than men working there in the first place.

It is clear that the data in the table are count data. They can therefore be analysed using Poisson regression. Later in this chapter we will see how we can do that. But first we have to realise that the counts are a bit different from the counts we observed in the cinema example where we counted the number of cinemas than a particular film. The dependent variable there was a count variable, but it also had a clear meaning: the count reflected the choice of cinema directors to show a film.

In the example here, we simply count the number of employees in a data set that fit a set of criteria (i.e. being a woman and being a senior manager). Technically, both variables are count variables, but conceptually they might feel different for you.

Therefore, before we dive deeper into Poisson models, we will look at a very traditional way of analysing count data in crosstables: computing the Pearson chi-square statistic.

16.5 Cross-tabulation and the Pearson chi-square statistic

Table 16.2 shows that there were 800 women and 200 men in the data set, and of these there were in total 30 employees in senior management. Now if 30 people

out of 1000 are in senior management, it means that a proportion of $\frac{30}{1000} = 0.03$ are in senior management. You might argue that there is only a gender balance if that same proportion is observed in both men and women.

Let's look at the proportion of women in senior management. We observe 22 female leaders out of a total of 800 women. That's equal to a proportion of $\frac{22}{800} = 0.0275$. That is quite close but not equal to 0.03.

Now let's look at the men. We observe 8 male leaders out of a total of 200 men. That's equal to a proportion of $\frac{8}{200} = 0.04$. That is quite close but not equal to 0.03.

Relatively speaking, therefore, there are more men in a senior management role (4%) than there are women in a senior management role (2.75%). Overall there are more women leaders in this data set, but taking into account the gender ratio, there are relatively fewer women than men in the leadership.

Now the question remains whether this difference in leadership across gender is real. These data come from only one particular month, from only 1000 employees instead of all 100,000. Every month new people get hired, get fired, or simply leave. With a huge turn-over, the numbers change all the time. The question is whether we can generalise from these data to the whole company: is there a general gender bias in this company, or are the results we look at due to the random sampling?

The Pearson chi-square test is based on comparing that what you expect given a hypothesis, with what you actually observed. Suppose our hypothesis is that there is a perfect gender balance: the proportion of leaders among the women is equal to the proportion of leaders among the men.

Before we continue, we assume that the total numbers of both genders and the leadership positions are reflective of the actual situation. We therefore assume the proportion of women is 0.8 and the proportion of men is 0.2. We also assume the proportion of senior managers is 3% and the proportion of other employees is 97%.

Assuming gender balance, we expect that a proportion of 0.03 of the women are in a leadership position, and we also expect a proportion of 0.03 of the men are in a leadership position. Given there are 800 women in the data set, under the gender balance assumption we expect $0.03 \times 800 = 24$ women with a leadership role. Given there are 200 men, under the gender balance assumption we expect $0.03 \times 200 = 6$ men with a leadership role.

Now that we have the number of people that we expect based on gender balance, we can compare them with the actual observed numbers. Let's calculate the deviations of what we observed from what we expect.

We observed 22 women in a leadership position, where we expected 24, so that is a deviation of $22 - 24 = -2$ (observed minus expected).

We observed 8 men in the leadership, where we expected 6, which results in a deviation of $8 - 6 = +2$ (observed minus expected).

We also have the observed counts of the employees not in the leadership. Also for them we can calculate expected numbers based on gender balance. A proportion of .97 are not in the leadership. We therefore expect $0.97 \times 800 = 776$ women. We observed 778, so then we have a deviation of $778 - 776 = +2$.

For the men we expect $0.97 \times 200 = 194$ to be without a leadership role, which results in a deviation of $192 - 194 = -2$.

If we would add these deviations up, we should be able to say something about how much the observed numbers deviate from the expected numbers. However, if we do that, we get a total of 0, since half of the deviations are positive and half of them are negative. In order to avoid that and get something that is very large when the deviations are very different from 0, we could instead take the squares of these differences and then add them up (the idea is similar to that of computing the variance as an indicator of how much values deviate from each other). Another reason for taking the squares is that the sign of the deviation doesn't really matter; it is only the distance from 0 that is relevant.

If we take the deviations, square them, and add them up, we get:

$$(-2)^2 + 2^2 + 2^2 + (-2)^2 = 16$$

However, we should not forget that a deviation of 2 is only a relatively small deviation when the expected number is large, like 776. Observing 778 instead of the expected 776 is not such a big deal: it's only a 0.26% increase. In contrast, if you expect only 6 people and suddenly you observe 8 people, that is relatively speaking a large difference (a 33% increase). We should therefore off-set the observed deviations from what we expected. We then get:

$$\frac{(-2)^2}{24} + \frac{2^2}{6} + \frac{2^2}{776} + \frac{(-2)^2}{194} = 0.86$$

What we actually did here was computing a test statistic that helps us to decide whether we have evidence in this data set that speaks against gender balance. It is called the Pearson chi-square statistic and it indicates to what extent the proportions observed in one categorical variable (say leadership membership vs non-membership) are different for the different categories of another categorical variable (say males and females). We observed different leadership proportions for males and females and this test statistic helps us decide if the proportions were also significantly different.

In formula form, the Pearson chi-square statistic looks like this:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

For every cell i in the crosstable, here 4 cells, we compute the difference between the observed count (O) and expected count (E). We square these difference first, then divide them by the expected count, and the resulting numbers we add up (summation). As we saw, we ended up with $X^2 = 0.86$.

If the null-hypothesis is true, the sampling distribution of X^2 is a chi-square distribution. The shape of this distribution depends on the degrees of freedom, similar to the t-distribution. The degrees of freedom for Pearson's chi-square are determined by the number of categories for variable 1, K_1, and the number of categories for variable 2, K_2: $\text{df} = (K_1 - 1)(K_2 - 1)$. Thus, with a 2×2 crosstable there is only one degree of freedom: $(2-1) \times (2-1) = 1$.

The chi-square distribution with one degree of freedom is plotted in Figure 16.3.

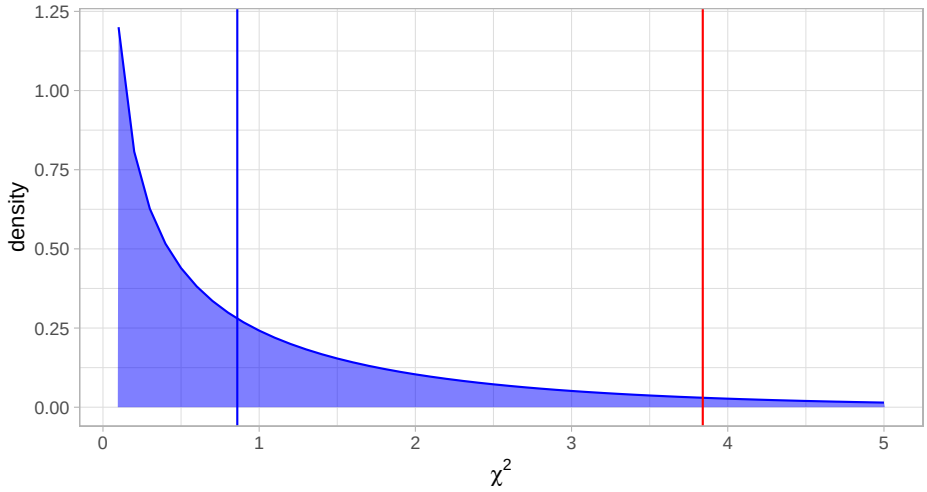


Figure 16.3: The chi-square distribution with one degree of freedom. Under the null-hypothesis we observe values larger than 3.84 only 5% of the time, indicated by the red line. The blue line represents the value based on the data.

The observed value of the test statistic is 0.86 and plotted in blue. The critical value for the null-hypothesis to be rejected at an alpha of 5% is 3.84 and depicted in red. The null-hypothesis can clearly not be rejected. There is no evidence that there is any gender imbalance in this large company of 100,000 employees.

16.6 Pearson chi-square in R

The Pearson chi-square test is based on data from a crosstable. The data that you have are most often not in such a format, so you have to create that table first.

Suppose you have the dataframe with the 1000 employees. We have the variable ID of the employee, a variable for the gender, and a variable for whether they are in senior management.

The top rows of the data might look like this:

```
companydata %>% head(3)
```

```
## # A tibble: 3 x 3
##      ID gender leader
##   <int> <chr>  <chr>
## 1     1 male   no
## 2     2 male   no
## 3     3 male   no
```

As we saw in Chapter 1, we can create a table with the `tabyl()` function from the janitor package.

```
library(janitor)
companytable <- companydata %>%
  tabyl(gender, leader)
companytable
```

```
##  gender  no yes
##  female 778  22
##    male 192   8
```

Next we plug the crosstable into the function `chisq.test()`.

```
chisq.test(companytable)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  companytable
## X-squared = 0.48325, df = 1, p-value = 0.487
```

Two things strike us: the value for the X-squared is not equal to the 0.86 that we computed by hand. But we also see that R used Yates' continuity correction. This continuity correction is important if some of your expected cell counts are rather low. Here we saw that we expected only 6 men in senior management.

If we don't let R apply this continuity correction, we obtain our manually computed statistic of 0.86:

```
chisq.test(companytable, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  companytable
## X-squared = 0.85911, df = 1, p-value = 0.354
```

Generally you should use the corrected version. We then report:

> "Using data on 1000 employees the observed proportion of women in a leadership role was 0.0275. The observed proportion of men in a leadership role was 0.04. A Pearson chi-square test with continuity correction showed that this difference was not significant, $X2(1) = 0.48$, $p = 0.487$. The null-hypothesis that the proportions are equal in the population could not be rejected."

16.7 Analysing crosstables with Poisson regression in R

If we want to analyse the same data with a Poisson regression, we need to create a dataframe with a count variable. We start from the original dataframe `companydata` and count the number of men and women with and without a senior management role.

```
companydata %>% head(3)
```

```
## # A tibble: 3 x 3
##      ID gender leader
##   <int> <chr>  <chr>
## 1     1 male   no
## 2     2 male   no
## 3     3 male   no
```

```
companycounts <- companydata %>%
  group_by(gender, leader) %>%
  summarise(count = n())
companycounts
```

```
## # A tibble: 4 x 3
## # Groups:   gender [2]
```

```
##   gender leader count
##   <chr>  <chr>  <int>
## 1 female no       778
## 2 female yes       22
## 3 male   no       192
## 4 male   yes        8
```

We now have a dataframe with the count data in long format, exactly what we need for the Poisson linear model. We take the count variable as dependent variable, and let gender and leader be the predictors:

```
model1 <- companycounts %>%
  glm(count ~ gender + leader, family = poisson, data = .)
model1
```

```
##
## Call:  glm(formula = count ~ gender + leader, family = poisson, data = .)
##
## Coefficients:
## (Intercept)   gendermale    leaderyes
##       6.654       -1.386       -3.476
##
## Degrees of Freedom: 3 Total (i.e. Null);  1 Residual
## Null Deviance:       1503
## Residual Deviance: 0.8003    AIC: 31.27
```

In the output we see that the effect of being male is negative, saying that there are fewer men than women in the data. The leadership effect is also negative: there are fewer people in a leadership role than people not in that role.

We can look at what the model predicts regarding the counts:

```
library(modelr)
companycounts %>%
  add_predictions(model1, type = "response")
```

```
## # A tibble: 4 x 4
## # Groups:   gender [2]
##   gender leader count  pred
##   <chr>  <chr>  <int> <dbl>
## 1 female no       778 776.
## 2 female yes       22  24.0
## 3 male   no       192 194.
## 4 male   yes        8   6
```

You see that the model makes the same predictions as under the assumption that there is a gender balance. The deviations in the observed compared to the predicted (expected) are 2 again, similar as we saw when computing the chi-square test. We want to know however, whether the ratio of leaders is the same in males and females: in other words whether the effect of leadership on counts is different for men and women. We therefore check if there is a significant gender by leader interaction effect.

```
model2 <- companycounts %>%
  glm(count ~ gender + leader + gender:leader,
      family = poisson, data = .)
model2 %>% tidy()
```

```
## # A tibble: 4 x 5
##   term                 estimate std.error statistic  p.value
##   <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)             6.66     0.0359   186.    0
## 2 gendermale             -1.40     0.0806   -17.4   1.55e-67
## 3 leaderyes              -3.57     0.216    -16.5   4.12e-61
## 4 gendermale:leaderyes    0.388    0.421      0.921 3.57e- 1
```

We see a positive interaction effect: the specific combination of being a male *and* a leader, has an extra positive effect of 0.388 on the counts in the data. The effect is however not significant, since the p-value is .357.

We see that the p-value is very similar to the p-value for the uncorrected Pearson chi-square test (.354) and different from the corrected one (.487). Every p-value with these kinds of analyses is only an approximation. The p-values are dissimilar here because the expected number of males in a senior management role is pretty low. For data sets with higher counts, the differences between the methods will disappear. Generally, the Pearson chi-square with continuity correction gives you best results for relatively low expected counts (counts of 5 or less).

If you would like to report the results form this Poisson analysis instead of a Pearson chi-square test, you could write:

> "Using data on 1000 employees the observed proportion of women in a leadership role was 0.0275. The observed proportion of men in a leadership role was 0.04. A Poisson regression with number of employees as the dependent variables and main effects of gender, leadership and their interaction showed a non-signficant interaction effect, $B = 0.388$, $SE = 0.421$, $z = 0.92$, $p = .357$. The null-hypothesis that the proportions are equal in the population could not be rejected."

If the data set is large enough and the numbers are not too close to 0, the same conclusions will be drawn, whether from a z-statistic for an interaction effect in a Poisson regression model, or from a cross-tabulation and computing a Pearson chi-square. The advantage of the Poisson regression approach is that you can do much more with them, for instance more than two predictors, and including also numeric variables as predictors. In the Poisson regression, you also make it more explicit that when computing the z-statistic, you take into account the main effects of the variables (the row and column totals). You do that also for the Pearson chi-square, but it is less obvious: we did that by first calculating the proportion of males and females and then calculating the overall proportion of senior management roles, before calculating the expected numbers. The next section shows that you can more easily answer advanced research questions with Poisson regression than with Pearson's chi-squares and crosstables.

16.8 Going beyond 2 by 2 crosstables (advanced)

If your categorical variables have more than 2 categories you can still do a Pearson chi-square test. You only run into trouble when you have more than two categorical variables, and you want to study the relationship between them. Let's turn to an example with three categorical variables. Suppose in the employee data, we have data on gender, leadership role, and also whether people work either at the head office in central London, or whether they work at one of the dozens of distribution centres across the world. Do we see a different gender bias in the head office compared to the distribution centres?

Let's have a look at the data again. Now we make separate tables, one for the head office and one for the distribution centres, see Figure 16.4.

gender	no	yes	Total
female	706	1	707
male	141	4	145
Total	847	5	852

gender	no	yes	Total
female	72	21	93
male	51	4	55
Total	123	25	148

Figure 16.4: The crosstables of gender and leadership, separately for the distribution centres (left table) and the head office (right table).

Remember that in the complete data set, we observed 4% of leadership in the men and 2.75% in the women. Now let's look only at the data from the distribution centres. Out of 707 women, 1 of them is in senior management, leading to a proportion of 0.001. Out of 145 men, 4 of them are a senior manager, leading to a proportion of 0.028. Thus, at the distribution centres there seems to be a gender bias in that men tend to be more often in senior

management positions than women, taking into account overall differences in gender representation.

Contrast this with the situation at the head office in London. There, 21 out of 93 women are in senior management, a proportion of 0.23. For the men, 4 out of 55 are in senior management, a proportion of 0.07. Thus, at the head office, senior management positions are relatively speaking more prevalent among the women than the men. The opposite of what we observed in the distribution centres.

The next question is then, can we generalise this conclusion to the population of the entire company: is the gender bias different at the two types of location? For that we can do a Poisson analysis. First we prepare the data in such a way that for every combination of gender, leadership and location, we get a count. We do that in the following way.

```
companydata_location %>% head(3)
```

```
## # A tibble: 3 x 4
##      ID gender leader location
##   <int> <chr>  <chr>  <chr>
## 1   182 female yes    distribution
## 2   197 female yes    headoffice
## 3   231 female yes    headoffice
```

```
companycounts <- companydata_location %>%
  group_by(gender, leader, location) %>%
  summarise(count = n()) %>%
  arrange(location)
companycounts
```

```
## # A tibble: 8 x 4
## # Groups:   gender, leader [4]
##   gender leader location     count
##   <chr>  <chr>  <chr>        <int>
## 1 female no     distribution   706
## 2 female yes    distribution     1
## 3 male   no     distribution   141
## 4 male   yes    distribution     4
## 5 female no     headoffice      72
## 6 female yes    headoffice      21
## 7 male   no     headoffice      51
## 8 male   yes    headoffice       4
```

Next, we apply a Poisson regression model. We predict the counts by taking the gender totals, the leader totals and the location totals.

```
companycounts %>%
  glm(count ~ gender + leader + location, family = poisson, data = .)
```

```
##
## Call:  glm(formula = count ~ gender + leader + location, family = poisson,
##     data = .)
##
## Coefficients:
##        (Intercept)          gendermale           leaderyes  locationheadoffice
##              6.494              -1.386              -3.476              -1.750
##
## Degrees of Freedom: 7 Total (i.e. Null);  4 Residual
## Null Deviance:       2168
## Residual Deviance: 117.9     AIC: 166.4
```

Generally, there are fewer males, fewer leaders, and fewer people in the head office compared to the other categories. In the previous analysis we saw that a gender bias in leadership was modelled using an interaction effect. Now that we have data on location, there could in addition be a gender bias in location (relatively more women in the head office), and a location bias in leadership (relatively more leaders in the head office). We can therefore add these three two-way interaction effects. But most importantly, we want to know whether the gender bias in management (the interaction between gender and leadership) is moderated by location. This we can test by adding a three-way interaction effect: gender by leader by location. Putting it all together we get the following R code:

```
companycounts %>%
  glm(count ~ gender + leader + location +
        gender:leader + gender:location + leader:location +
        gender:leader:location,
      family = poisson, data = .)
```

The exact same model can be achieved with less typing, by using the * operator: then you get all main effects, two-way interaction effects and the three-way interaction effect automatically:

```
companycounts %>% glm(count ~ gender*leader*location,
                      family = poisson, data = .) %>%
  tidy()
```

```
## # A tibble: 8 x 5
##   term                                estimate std.error statistic  p.value
```

```
##   <chr>                                    <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                               6.56    0.0376    174.   0
## 2 gendermale                               -1.61    0.0922    -17.5  2.73e-68
## 3 leaderyes                                -6.56    1.00       -6.56 5.56e-11
## 4 locationheadoffice                       -2.28    0.124     -18.5  4.90e-76
## 5 gendermale:leaderyes                      3.00    1.12        2.67 7.55e- 3
## 6 gendermale:locationheadoffice             1.27    0.205       6.18 6.53e-10
## 7 leaderyes:locationheadoffice              5.33    1.03        5.17 2.37e- 7
## 8 gendermale:leaderyes:locationheadoffice  -4.31    1.26       -3.42 6.29e- 4
```

The only parameter relevant for our research question is the last one concerning the gender by leader by location interaction effect. Its value -4.31 is negative, indicating that the particular combination of being male, being a leader and working at the head office leads to a relatively lower count. There is thus a gender by location bias in leadership, in that the males are relatively more present in leadership positions at the distribution centres than at the head office. The effect is significant, so we can conclude that the gender bias shows a different pattern at the head office than at the distribution centres.

We report:

> "Using data on 1000 employees the observed proportion of women in a leadership role was 0.0275. The observed proportion of men in a leadership role was 0.04. Taking into account differences across location we saw that women at distribution centres were relatively less often in senior management position than men (0.1% vs. 2.8%). At the head office, women were relatively more often in a senior management position than men (23% vs. 7%). A Poisson regression with number of employees as the dependent variables and main effects of gender, leadership, location and all their two-way and three-way interaction effects showed a significant three-way interaction effect, $B = -4.31$, $SE = 1.26$, $z = -3.42$, $p < .001$. The null-hypothesis that the gender bias is equal in the two types of locations of the company can be rejected."

16.9 Take-away points

- A Poisson regression model is a form of a generalised linear model.
- Poisson regression is appropriate in situations where the dependent variable is a count variable $(0, 1, 2, 3, ...)$.
- Whereas the normal distribution has two parameters (mean μ and variance σ^2), the Poisson distribution has only one parameter: λ.

- A Poisson distribution with parameter λ has mean equal to λ and variance equal to λ.
- λ can take any real value between 0 and ∞.
- Poisson models can be used to analyse crosstable data. This can also be done using Pearson's chi-square, but Poisson models allow you to easily go beyond two categorical variables, and include numerical predictors.

Key concepts

- Poisson regression
- Poisson distribution
- Tendency parameter λ
- Pearson's chi-square
- χ^2-distribution

Appendix A

Cumulative probabilities for the standard normal distribution

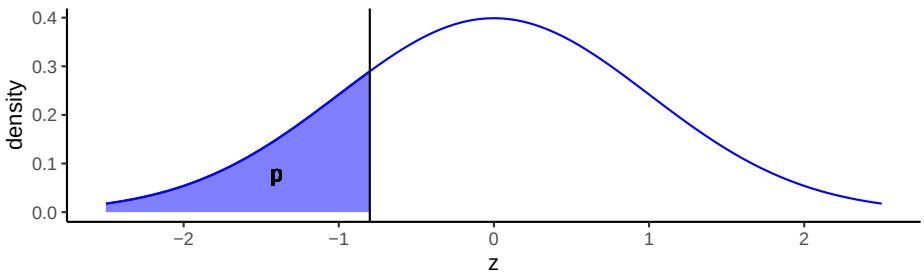


Table A.1: Cumulative proportions p for the standard normal distribution.

z	p	z	p	z	p	z	p	z	p	z	p
-4.00	0.0000	-1.43	0.0764	-0.71	0.239	0.01	0.504	0.73	0.7673	1.45	0.9265
-3.80	0.0001	-1.42	0.0778	-0.70	0.242	0.02	0.508	0.74	0.7704	1.46	0.9279
-3.60	0.0002	-1.41	0.0793	-0.69	0.245	0.03	0.512	0.75	0.7734	1.47	0.9292
-3.40	0.0003	-1.40	0.0808	-0.68	0.248	0.04	0.516	0.76	0.7764	1.48	0.9306
-3.20	0.0007	-1.39	0.0823	-0.67	0.251	0.05	0.520	0.77	0.7794	1.49	0.9319
-3.00	0.0013	-1.38	0.0838	-0.66	0.255	0.06	0.524	0.78	0.7823	1.50	0.9332
-2.90	0.0019	-1.37	0.0853	-0.65	0.258	0.07	0.528	0.79	0.7852	1.51	0.9345
-2.80	0.0026	-1.36	0.0869	-0.64	0.261	0.08	0.532	0.80	0.7881	1.52	0.9357
-2.70	0.0035	-1.35	0.0885	-0.63	0.264	0.09	0.536	0.81	0.7910	1.53	0.9370
-2.60	0.0047	-1.34	0.0901	-0.62	0.268	0.10	0.540	0.82	0.7939	1.54	0.9382
-2.50	0.0062	-1.33	0.0918	-0.61	0.271	0.11	0.544	0.83	0.7967	1.55	0.9394
-2.40	0.0082	-1.32	0.0934	-0.60	0.274	0.12	0.548	0.84	0.7995	1.56	0.9406

-2.30	0.0107	-1.31	0.0951	-0.59	0.278	0.13	0.552	0.85	0.8023	1.57	0.9418
-2.20	0.0139	-1.30	0.0968	-0.58	0.281	0.14	0.556	0.86	0.8051	1.58	0.9429
-2.10	0.0179	-1.29	0.0985	-0.57	0.284	0.15	0.560	0.87	0.8078	1.59	0.9441
-2.00	0.0228	-1.28	0.1003	-0.56	0.288	0.16	0.564	0.88	0.8106	1.60	0.9452
-1.99	0.0233	-1.27	0.1020	-0.55	0.291	0.17	0.567	0.89	0.8133	1.61	0.9463
-1.98	0.0239	-1.26	0.1038	-0.54	0.295	0.18	0.571	0.90	0.8159	1.62	0.9474
-1.97	0.0244	-1.25	0.1056	-0.53	0.298	0.19	0.575	0.91	0.8186	1.63	0.9484
-1.96	0.0250	-1.24	0.1075	-0.52	0.302	0.20	0.579	0.92	0.8212	1.64	0.9495
-1.95	0.0256	-1.23	0.1093	-0.51	0.305	0.21	0.583	0.93	0.8238	1.65	0.9505
-1.94	0.0262	-1.22	0.1112	-0.50	0.309	0.22	0.587	0.94	0.8264	1.66	0.9515
-1.93	0.0268	-1.21	0.1131	-0.49	0.312	0.23	0.591	0.95	0.8289	1.67	0.9525
-1.92	0.0274	-1.20	0.1151	-0.48	0.316	0.24	0.595	0.96	0.8315	1.68	0.9535
-1.91	0.0281	-1.19	0.1170	-0.47	0.319	0.25	0.599	0.97	0.8340	1.69	0.9545
-1.90	0.0287	-1.18	0.1190	-0.46	0.323	0.26	0.603	0.98	0.8365	1.70	0.9554
-1.89	0.0294	-1.17	0.1210	-0.45	0.326	0.27	0.606	0.99	0.8389	1.71	0.9564
-1.88	0.0301	-1.16	0.1230	-0.44	0.330	0.28	0.610	1.00	0.8413	1.72	0.9573
-1.87	0.0307	-1.15	0.1251	-0.43	0.334	0.29	0.614	1.01	0.8438	1.73	0.9582
-1.86	0.0314	-1.14	0.1271	-0.42	0.337	0.30	0.618	1.02	0.8461	1.74	0.9591
-1.85	0.0322	-1.13	0.1292	-0.41	0.341	0.31	0.622	1.03	0.8485	1.75	0.9599
-1.84	0.0329	-1.12	0.1314	-0.40	0.345	0.32	0.626	1.04	0.8508	1.76	0.9608
-1.83	0.0336	-1.11	0.1335	-0.39	0.348	0.33	0.629	1.05	0.8531	1.77	0.9616
-1.82	0.0344	-1.10	0.1357	-0.38	0.352	0.34	0.633	1.06	0.8554	1.78	0.9625
-1.81	0.0351	-1.09	0.1379	-0.37	0.356	0.35	0.637	1.07	0.8577	1.79	0.9633
-1.80	0.0359	-1.08	0.1401	-0.36	0.359	0.36	0.641	1.08	0.8599	1.80	0.9641
-1.79	0.0367	-1.07	0.1423	-0.35	0.363	0.37	0.644	1.09	0.8621	1.81	0.9649
-1.78	0.0375	-1.06	0.1446	-0.34	0.367	0.38	0.648	1.10	0.8643	1.82	0.9656
-1.77	0.0384	-1.05	0.1469	-0.33	0.371	0.39	0.652	1.11	0.8665	1.83	0.9664
-1.76	0.0392	-1.04	0.1492	-0.32	0.374	0.40	0.655	1.12	0.8686	1.84	0.9671
-1.75	0.0401	-1.03	0.1515	-0.31	0.378	0.41	0.659	1.13	0.8708	1.85	0.9678
-1.74	0.0409	-1.02	0.1539	-0.30	0.382	0.42	0.663	1.14	0.8729	1.86	0.9686
-1.73	0.0418	-1.01	0.1562	-0.29	0.386	0.43	0.666	1.15	0.8749	1.87	0.9693
-1.72	0.0427	-1.00	0.1587	-0.28	0.390	0.44	0.670	1.16	0.8770	1.88	0.9699
-1.71	0.0436	-0.99	0.1611	-0.27	0.394	0.45	0.674	1.17	0.8790	1.89	0.9706
-1.70	0.0446	-0.98	0.1635	-0.26	0.397	0.46	0.677	1.18	0.8810	1.90	0.9713
-1.69	0.0455	-0.97	0.1660	-0.25	0.401	0.47	0.681	1.19	0.8830	1.91	0.9719
-1.68	0.0465	-0.96	0.1685	-0.24	0.405	0.48	0.684	1.20	0.8849	1.92	0.9726
-1.67	0.0475	-0.95	0.1711	-0.23	0.409	0.49	0.688	1.21	0.8869	1.93	0.9732
-1.66	0.0485	-0.94	0.1736	-0.22	0.413	0.50	0.691	1.22	0.8888	1.94	0.9738
-1.65	0.0495	-0.93	0.1762	-0.21	0.417	0.51	0.695	1.23	0.8907	1.95	0.9744
-1.64	0.0505	-0.92	0.1788	-0.20	0.421	0.52	0.698	1.24	0.8925	1.96	0.9750
-1.63	0.0516	-0.91	0.1814	-0.19	0.425	0.53	0.702	1.25	0.8944	1.97	0.9756
-1.62	0.0526	-0.90	0.1841	-0.18	0.429	0.54	0.705	1.26	0.8962	1.98	0.9761
-1.61	0.0537	-0.89	0.1867	-0.17	0.433	0.55	0.709	1.27	0.8980	1.99	0.9767
-1.60	0.0548	-0.88	0.1894	-0.16	0.436	0.56	0.712	1.28	0.8997	2.00	0.9772
-1.59	0.0559	-0.87	0.1922	-0.15	0.440	0.57	0.716	1.29	0.9015	2.10	0.9821
-1.58	0.0571	-0.86	0.1949	-0.14	0.444	0.58	0.719	1.30	0.9032	2.20	0.9861
-1.57	0.0582	-0.85	0.1977	-0.13	0.448	0.59	0.722	1.31	0.9049	2.30	0.9893

-1.56	0.0594	-0.84	0.2005	-0.12	0.452	0.60	0.726	1.32	0.9066	2.40	0.9918
-1.55	0.0606	-0.83	0.2033	-0.11	0.456	0.61	0.729	1.33	0.9082	2.50	0.9938
-1.54	0.0618	-0.82	0.2061	-0.10	0.460	0.62	0.732	1.34	0.9099	2.60	0.9953
-1.53	0.0630	-0.81	0.2090	-0.09	0.464	0.63	0.736	1.35	0.9115	2.70	0.9965
-1.52	0.0643	-0.80	0.2119	-0.08	0.468	0.64	0.739	1.36	0.9131	2.80	0.9974
-1.51	0.0655	-0.79	0.2148	-0.07	0.472	0.65	0.742	1.37	0.9147	2.90	0.9981
-1.50	0.0668	-0.78	0.2177	-0.06	0.476	0.66	0.745	1.38	0.9162	3.00	0.9987
-1.49	0.0681	-0.77	0.2206	-0.05	0.480	0.67	0.749	1.39	0.9177	3.20	0.9993
-1.48	0.0694	-0.76	0.2236	-0.04	0.484	0.68	0.752	1.40	0.9192	3.40	0.9997
-1.47	0.0708	-0.75	0.2266	-0.03	0.488	0.69	0.755	1.41	0.9207	3.60	0.9998
-1.46	0.0721	-0.74	0.2296	-0.02	0.492	0.70	0.758	1.42	0.9222	3.80	0.9999
-1.45	0.0735	-0.73	0.2327	-0.01	0.496	0.71	0.761	1.43	0.9236	4.00	1.0000
-1.44	0.0749	-0.72	0.2358	0.00	0.500	0.72	0.764	1.44	0.9251	4.01	1.0000

Appendix B

Critical values for the t-distribution

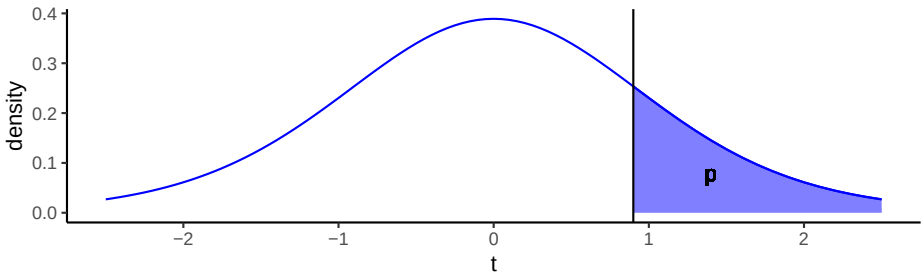


Table B.1: Critical values for the t-distribution, given the degrees of freedom (rows) and tail probability p (columns). These can be used for critical values for a given confidence level.

	Confidence level								
	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Right-tail probability p								
df	**0.1**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
1	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
120	1.289	1.658	1.980	2.076	2.358	2.617	2.860	3.160	3.373
10000	1.282	1.645	1.960	2.054	2.327	2.576	2.808	3.091	3.291

Appendix C

Some basic algebra for linear models

If you want to understand linear models better, some *linear algebra* (a.k.a. *matrix algebra*) comes in handy. For linear algebra, you need to understand how to work with matrices: how to multiply them (calculating the product of two matrices), how to *invert* them (a kind of determining the reciprocal), and how to *transpose* them. We briefly explain those operations here, and then show how these tricks relate to ordinary least squares (OLS) estimation in practice. We also show how to perform linear algebra in R.

If you want to compute the product of matrix $\mathbf{A}$ and matrix $\mathbf{B}$, the order of the matrices is important. Suppose we have the matrix $\mathbf{A}$

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

and matrix $\mathbf{B}$

$$\mathbf{B} = \begin{bmatrix} e & f \\ g & h \end{bmatrix}$$

If you want to calculate the product $\mathbf{AB}$, you take each row of $\mathbf{A}$ and you take the sum of the crossproducts with each column of $\mathbf{B}$. That is, if the product of $\mathbf{AB} = \mathbf{C}$, then the first element of $\mathbf{C}$, c_{11} (row 1, column 1) equals $a \times e + b \times g$. In general, if $\mathbf{C} = \mathbf{AB}$, and c_{ij} is the element in the i-th row and the j-th column, we have

$$c_{ij} = A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{iJ}B_{Ij}$$

where I is the number of rows in $\mathbf{B}$ and J is the number of columns in $\mathbf{A}$.

If we do that for each row of $\mathbf{A}$ and each column of $\mathbf{B}$, then we get matrix $\mathbf{C}$:

$$\mathbf{C} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{bmatrix}$$

Example C.1 (Matrix Multiplication)**.** Let

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

and

$$\mathbf{B} = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}$$

Then the product $\mathbf{AB}$ equals

$$\mathbf{AB} = \begin{bmatrix} 2 & 4 \\ 3 & 7 \end{bmatrix}$$

whereas the product $\mathbf{BA}$ equals

$$\mathbf{BA} = \begin{bmatrix} 6 & 4 \\ 4 & 3 \end{bmatrix}$$

In R, multiplying matrices is done with the operator `%*%`:

```
A <- rbind(c(1, 0),
           c(1, 1))
B <- rbind(c(2, 4),
           c(1, 3))
A%*%B  # A x B
```

```
##      [,1] [,2]
## [1,]    2    4
## [2,]    3    7
```

```
B%*%A  # B x A
```

```
##      [,1] [,2]
## [1,]    6    4
## [2,]    4    3
```

Let $\mathbf{y}$ be a vector containing the values of the dependent variable Y. Let $\mathbf{X}$ be the design matrix (see Chapter 10), where each row represents a unit of observation, and each column represents a *numeric* independent variable (remember that each categorical variable is always coded into one or more numeric variables, e.g. dummy variables).

When you do some algebra with matrix $\mathbf{X}$ and vector $\mathbf{y}$, you get the parameters (intercept and slopes) for a linear model:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

where $\mathbf{X}^\top$ is the transpose of $\mathbf{X}$, which means the matrix on its side, where the first row becomes the first column, the second row becomes the second column, etcetera. For instance, suppose matrix $\mathbf{X}$ is

$$\mathbf{X} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$$

then the transpose of $\mathbf{X}$ is $\mathbf{X}^\top$

$$\mathbf{X}^\top = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

In R, the transpose can be calculated using `t()`:

```
M <- rbind(c(0, 1, 1, 0),
           c(1, 1, 1, 1))
M
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    0    1    1    0
## [2,]    1    1    1    1
```

```
t(M)  # the transpose
```

```
##      [,1] [,2]
## [1,]    0    1
## [2,]    1    1
## [3,]    1    1
## [4,]    0    1
```

$\mathbf{X}^{-1}$ is the *inverse* of matrix $\mathbf{X}$. The inverse of a matrix $\mathbf{X}$ is a matrix for which we know that $\mathbf{X}\mathbf{X}^{-1} = \mathbf{I}$, where $\mathbf{I}$ is the *identity matrix*.

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

that is, a matrix with all 0s except on the diagonal from top-left to bottom-right.

In R, the inverse of a matrix can be calculated using the `ginv()` function in the `MASS` package (Chapter 10). The function `ginv()` can be used for all matrices (it actually computes the *generalised inverse*). If the matrix has as many rows as it has columns, a matrix is called "square" and then the function `solve()` can also be used.

```
X = rbind(c(1, 1),  # a square matrix X
          c(0, 1))
X
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    0    1
```

```
solve(X) # the inverse of X
```

```
##      [,1] [,2]
## [1,]    1   -1
## [2,]    0    1
```

```
X %*% solve(X) # a product of a square matrix with its inverse is identity matrix
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
```

Example C.2 (Obtaining linear model parameters)**.** Suppose we have the data set in Table C.1, where

Let response vector $\mathbf{y}$ contain the observed values of the dependent variable Y:

$$\mathbf{y} = \begin{bmatrix} 1 \\ 0 \\ 2 \\ -4 \\ -1 \\ 1 \end{bmatrix}$$

Table C.1: Small data example.

y	group
1	1
0	1
2	2
-4	2
-1	3
1	3

Table C.2: Small data example with the columns of the design matrix.

y	group	intercept	group2	group3
1	1	1	0	0
0	1	1	0	0
2	2	1	1	0
-4	2	1	1	0
-1	3	1	0	1
1	3	1	0	1

Let design matrix $\mathbf{X}$ contain the values for three independent variables (three columns). The first is by default a vector of 1s. The second and third variables are dummy variables coding for **group**. These new variables are given in Table C.2 to show how they relate to the original data.

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Then the vector with the OLS parameter values (intercept and slopes) is obtained by calculating

$$(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 0.5 \\ -1.5 \\ -.5 \end{bmatrix}$$

Let R do the algebra for you:

```
y <- c(1, 0, 2, -4, -1, 1)
group2 = c(0, 0, 1, 1, 0, 0)
```

```
group3 = c(0, 0, 0, 0, 1, 1)
X <- cbind(intercept = 1,
           group2,
           group3)
X
```

```
##      intercept group2 group3
## [1,]         1      0      0
## [2,]         1      0      0
## [3,]         1      1      0
## [4,]         1      1      0
## [5,]         1      0      1
## [6,]         1      0      1
```

```
solve(t(X)%*%X) %*% t(X) %*% y
```

```
##           [,1]
## intercept  0.5
## group2    -1.5
## group3    -0.5
```

You obtain the same values with the function `lm()`:

```
tibble(y = c(1, 0, 2, -4, -1, 1), # dependent variable
       group = factor(c(1, 1, 2, 2, 3, 3))) %>%  # categorical indep variable
  lm(y ~ group, data = .) # run a linear model with ordinary least squares
```

```
##
## Call:
## lm(formula = y ~ group, data = .)
##
## Coefficients:
## (Intercept)       group2       group3
##         0.5         -1.5         -0.5
```

What the design matrix looks like can be obtained using `model.matrix()`.

```
tibble(y = c(1, 0, 2, -4, -1, 1), # dependent variable
       group = factor(c(1, 1, 2, 2, 3, 3))) %>%  # categorical indep variable
  lm(y ~ group, data = .) %>% # run a linear model with ordinary least squares %>%
  model.matrix() # show design matrix X
```

```
##   (Intercept) group2 group3
## 1           1      0      0
## 2           1      0      0
## 3           1      1      0
## 4           1      1      0
## 5           1      0      1
## 6           1      0      1
## attr(,"assign")
## [1] 0 1 1
## attr(,"contrasts")
## attr(,"contrasts")$group
## [1] "contr.treatment"
```

You see the intercept variable and the two dummy variables that are created by default for the factor variable **group**.