

Advanced Statistics Remix

David Schuster

2021-10-11

Contents

About this Book	9
Attribution	9
License	9
1 Statistics for Research	11
1.1 Introduction	11
1.2 Measurement	12
1.3 Descriptive Statistics: Summarizing our observations	15
1.4 Inferential Statistics: Generalizing from our observations	16
1.5 The cautionary tale of Simpson’s paradox	19
1.6 A brief introduction to research design	22
1.7 Causality, Research, and Statistics	28
2 Getting started with R	35
2.1 Videos	35
2.2 Introduction	35
2.3 Installing R	37
2.4 Typing commands at the R console	42
2.5 Doing simple calculations with R	49
2.6 Storing a number as a variable	52
2.7 Using functions to do calculations	56
2.8 Letting RStudio help you with your commands	61
2.9 Storing many numbers as a vector	64

2.10 Storing text data	68
2.11 Storing “true or false” data	70
2.12 Indexing vectors	78
2.13 Quitting R	81
2.14 Summary	83
2.15 Additional R concepts	84
2.16 Using comments	85
2.17 Installing and loading packages	86
2.18 Managing the workspace	96
2.19 Navigating the file system	100
2.20 Loading and saving data	106
2.21 Useful things to know about variables	115
2.22 Factors	120
2.23 Data frames	123
2.24 Lists	127
2.25 Formulas	129
2.26 Generic functions	130
2.27 Getting help	131
2.28 Summary	136
3 Descriptive statistics	139
3.1 Videos	139
3.2 Introduction	139
3.3 Measures of central tendency	142
3.4 Measures of variability	153
3.5 Skew and kurtosis	165
3.6 Getting an overall summary of a variable	173
3.7 Descriptive statistics separately for each group	178
3.8 Good descriptive statistics are descriptive!	181
3.9 Drawing graphs	182

CONTENTS	5
4 Inferential statistics: The Central Limit Theorem	211
4.1 Videos	211
4.2 Introduction	211
4.3 How are probability and statistics different?	212
4.4 What does probability mean?	213
4.5 Samples, populations and sampling	219
4.6 The law of large numbers	230
4.7 Sampling distributions and the central limit theorem	232
4.8 Estimating population parameters	252
4.9 Estimating a confidence interval	260
4.10 Summary	266
5 Hypothesis testing	269
5.1 Videos	269
5.2 Introduction	269
5.3 A menagerie of hypotheses	270
5.4 Two types of errors	274
5.5 Test statistics and sampling distributions	276
5.6 Making decisions	278
5.7 The p value of a test	281
5.8 Reporting the results of a hypothesis test	285
5.9 Running the hypothesis test in practice	287
5.10 Effect size, sample size and power	288
5.11 Some issues to consider	296
5.12 Summary	299
6 Issues in Hypothesis Testing	301
6.1 Videos	301
6.2 Introduction	301
6.3 The researcher affects NHST outcomes	302
6.4 NHST Misunderstandings	306
6.5 NHST Issues	307
6.6 Conclusions	308

7 Data Cleaning and Missing Values Analysis	309
7.1 Videos	309
7.2 Introduction: Dealing with the Unexpected	309
7.3 Data Cleaning	310
7.4 A General Plan for Data Cleaning	311
7.5 Step 0. Design your Research to Minimize Data Problems	312
7.6 Step 1. Examine Your Data	313
7.7 Step 2. Outlier Analysis	339
7.8 Step 3. Missing values analysis	341
7.9 Step 4. Test-specific assumption checking	349
7.10 Communicate results of data cleaning in APA style	349
8 Regression	351
8.1 Videos	351
8.2 Introduction	352
8.3 The General Linear Model (GLM)	352
8.4 Correlations	355
8.5 Linear regression	374
8.6 Estimating a linear regression model	378
8.7 Multiple linear regression	381
8.8 Quantifying the fit of the regression model	384
8.9 Hypothesis tests for regression models	388
8.10 Testing the significance of a correlation	393
8.11 Regarding regression coefficients	396
8.12 Assumptions of regression	399
8.13 Model checking	400
8.14 Model selection	423
8.15 Practical Issues in Correlation and Regression	432
8.16 Summary	438

CONTENTS	7
----------	---

9 Statistics Reference	439
9.1 One-Sample z -Test	439
9.2 Correlation	443
9.3 Linear Regression	450

About this Book

This is a textbook for my advanced statistics course, first used in Fall 2021.

It is a remix of existing open source educational materials. I am contributing very little text. The primary source of content is Navarro (2018).

Attribution

Where authors are indicated throughout this text, the content has been copied verbatim with no more than minor editorial changes. Note that this differs from an APA style manuscript in which all verbatim text is typically quoted or blockquoted.

Navarro, D. (2018). Learning statistics with R: A tutorial for psychology students and other beginners (version 0.6). Retrieved from <https://learningstatisticswithr.com>

Crump, M. J. C., Navarro, D., & Suzuki, J. (2019, June 5). Answering Questions with Data (Textbook): Introductory Statistics for Psychology Students. <https://doi.org/10.17605/OSF.IO/JZE52>

Further, text copied from Navarro (2018) is from the Bookdown translation by Emily Kothe.

License

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Chapter 1

Statistics for Research

1.1 Introduction

Text by David Schuster

Video: Applied Statistics

Statistics is a rich and diverse field with endless theories and application. To call this a “statistics course” may be too vague to be useful. Will we study the theory behind the statistics or study how statistics are used? Let’s narrow our scope. This course is primarily concerned with statistical methods used by researchers. That is, researchers systematically (deliberately and consistently) gather evidence in order to generate new knowledge. Statistics provide an important tool to help researchers be more systematic in their discovery of new knowledge. Researchers are to statisticians as video game players are to video game designers. Most video game players enjoy playing games but don’t necessarily care about how the game is constructed, coded, developed, and sold. Similarly, most researchers don’t necessarily care about how mathematical theory supports statistical concepts; instead, they want to use statistics to answer their research questions. Unfortunately, unlike playing a video game, statistical methods provide very little feedback (and usually none at all) about whether or not they are being used correctly. Because of this, researchers do have to understand a bit about how statistical methods work.

You can summarize all of this by saying that you are studying **applied statistics**, the use of statistical methods to address research problems (Cohen). Throughout this course, we will emphasize the statistical knowledge needed to understand and produce research. When theory is introduced (we might think of theoretical statistics as the opposite of applied statistics), it will be included because it helps understanding of concepts we need as researchers.

Doing research is exciting and important because it’s our best tool for solving

big societal problems, discovering solutions, and separating fact from fiction. Because of this, many researchers are more fascinated by science and their field of study than they are about statistical methods. I have been teaching statistics for over ten years and can confirm that student attitudes about this topic vary widely. If you do not feel like you love statistics in this moment, that is a common feeling. At the same time, you should be aware that many professionals happily and confidently use statistics in their careers and/or daily lives without identifying as mathematicians. This is not to disparage the mathematical perspective, only to say research and mathematics are interesting in different ways. If you do not yet think studying statistics is useful or interesting, perhaps you will challenge your attitudes about math and statistics throughout this course. And if not, perhaps you will provide useful feedback to your instructor!

There are two broad situations where researchers need statistics. Researchers need statistics to:

1. Gather observations in a systematic way (**measurement**)
2. Summarize their observations (**descriptive statistics**)
3. Make conclusions about populations based on the observations (**inferential statistics**)

In the next sections, we will unpack these three functions.

1.2 Measurement

Text by David Schuster

Video: Measurement

The very beginning of statistics, and the most fundamental building block, is data. Data are what we get when we combine numbers and meaning. If I write down any number that comes to mind, I am generating numbers but not data (because there is no meaning). If I wonder about how many students attempt to cross the busy street outside of my office, I am starting to develop a research question (I would say there is meaning involved) but there are no numbers yet, so no data. If I go outside and count the number of pedestrians that cross the street in an hour, I am now gathering data.

Different kinds of data contain different kinds of information. I have already been simplifying my definitions by suggesting that data always involves numbers. If my research question is, “do students walk across the street?” and I go outside for an hour and observe them to do so, then I have gathered data. But, are these data very informative? What is the difference between observing that “some students cross the street per hour” and “40 students cross the street per hour”? They both say the same thing, but the second version provides more specific information. As another example, I could say that it is hot out today,

or I could say that it is 99 degrees Fahrenheit (37 degrees Celsius). I could use either of these labels to describe the same day.

The process of gathering data is called **measurement**. It will be useful for us to classify measures according to the kind of data they contain. We will classify measurement in three ways (from Stevens, 1946):

1. According to their *level of measurement*
2. Whether or not they are *continuous or discrete*
3. Whether they represent *qualitative or quantitative* data.

Once you understand these classifications, you should be able to classify a measure in these three ways.

1.2.1 Level of Measurement

A stair diagram is used because higher levels of measurement satisfy all the requirements of the levels below.

Ratio scale/ratio measurement. Examples: weight, length

Interval scale/interval measurement. Example: Fahrenheit temperature

Ordinal scale/ordinal measurement. Example: the order in which people finish a race

Nominal scale/Nominal measurement. Example: which is your favorite fruit?

Notice that these levels are stair steps. Each level has all the characteristics of the level below it. Interval scales meet all the requirements of ordinal and nominal scales as well (plus they meet the additional requirement for interval scales).

To determine the level of measurement, ask yourself these questions:

1. Can you rank/order the numbers? (if no, nominal scale. if yes, keep going)
example: kinds of fish. can you rank halibut and mullet? (no, nominal scale) example: Olympic medals, can you rank gold, silver, and bronze? (yes, keep going)
2. If you add/subtract the numbers, does the result have meaning? (if no, ordinal scale. if yes, keep going) example: 30 degrees F plus 10 degrees equals 40 degrees (yes, keep going) example: 1st place plus 2 equals 3rd place? (no, this does not make sense, ordinal scale)
3. Does the score have a value of 0 that means 'none' or 'nothing'? (if no, interval scale. if yes, ratio scale) example: counting people; 0 people means no people (yes, ratio scale) example: 0 degrees F means no heat? (no, interval scale)

That last property, having a zero meaning none/nothing/not any is called a **true zero**. Fahrenheit temperature does not have a true zero (it is just another temperature), but Kelvin does (zero degrees is absolute zero and indicates no heat energy).

I find making up values to be a helpful strategy, as I did when I asked Question 2, above. It does not matter what the values are, so you can invent ones to make the questions more concrete.

When students are confused about classifying measures, the most common pattern I see is that they abandon the stair-step-question method. I recommend not trying to skip answering the questions, even as you start to get comfortable with this concept. Start from question 1, and continue up the levels until the answer is no. It takes a few more seconds but is much more reliable. And, remember that each level has all the properties of the levels below it. In other words, ratio scales meet all the requirements of interval scales, ordinal scales, and nominal scales. For this reason, I find trying to match definitions to examples is more confusing than the stair-step method (which was taught to me by my graduate advisor, and I am still using it!).

The second common point of confusion happens when students focus on the data instead of the measurement scale. When we classify measures, we are classifying the measurement scale. The measurement scale includes all possible data that could ever be observed (even if only theoretically). I usually use an exercise question that asks students to classify the level of measurement of the age of a football stadium. Like all measures of duration, the best answer is ratio. Often students are uncomfortable picking that answer, because they do not see anyone observing a football stadium to be 0 years (or days, hours, or minutes) old. How could a football stadium have no age? Even though it is unlikely that a list of football stadium ages would ever observe this, there is an instant where a football stadium has been constructed or opened and is therefore 0 years (days, hours, and minutes) old. All this to say, do not get distracted by what values are the most common or realistic. Instead, when classifying measurement scales, focus on all possible values.

1.2.2 Continuous or Discrete

Separately, you can decide if your variable is continuous or discrete. If you can have an infinite number of fractions of a value, it's continuous. If you cannot, the measure is discrete. example: 5 yards, 5.0005 yards, 5.5 years, and 5.500001 yards are all valid measurements (continuous) example: Olympic medals; the measurement between gold and silver does not exist (discrete)

There may be instances where a grey area exists; at some level, all variables are discrete. For example, you could subdivide a measurement of length down to the molecule. At that point, you cannot have fractional values. Try to avoid over-thinking this issue. If you can reasonably talk about fractional values

(half seconds; twenty-five cents are a fraction of a dollar) then the measure is continuous. If you cannot (there is no such thing as half a dog or an eighth of an employee), then the measure is discrete.

1.2.3 Qualitative or Quantitative

Quantitative data is associated with a numerical value. Qualitative data is associated with labels that have no numerical value. Nominal and ordinal data are qualitative. Interval and ratio data are quantitative.

1.2.4 Distribution: A collection of our observations

When we make repeated, related observations and collect them together, we have data. When we represent data in numerical or categorical form, we form a **distribution**. When you see distribution, think of a collection of scores.

1.3 Descriptive Statistics: Summarizing our observations

Text by David Schuster

Video: Descriptive Statistics

The problem with distributions is that any collection of more than a couple observations quickly overwhelms our limited working memory and attention. We need a way to summarize distributions. Descriptive statistics does exactly that. A descriptive statistic summarizes a distribution (put another way, it measures a property of a distribution) using a single value.

Descriptive statistics lets us summarize two properties of distributions:

1. The value of the scores (central tendency)
2. How spread out the scores are from each other (variability)

Measures of central tendency are averages. There are multiple ways of expressing an average. Mean, median, and mode are different kinds of averages. That is about all we need for right now. Later, we will go into more detail on how these useful tools work.

Measures of variability put a number on how spread out the data are. Think about your workplace—Are some employees more content than others? Is everyone pretty much in agreement that your workplace is great (or awful)? If most people tend to agree, then we might say your workplace satisfaction has

low variability. If there was not so much agreement, we might say your workplace satisfaction has high variability. With measures of variability, we can do even better by quantifying variability. Variability is the concept—how different are scores in the distribution? Measures of variability turn this into a value. Measures of variability include range, sum of squares, variance, and standard deviation.

When there is no variability, we call the value a *constant*. For almost anything you can think to measure about people, there are no constants. We live in a world of complex variability. For me, this is one of the most fascinating and challenging aspects of psychology. You can easily manufacture a bolt to have the same property as another, but psychologists get a front row seat at the amazing diversity of human thought and behavior. Describing and making predictions about variability is also a linkage between statistics and the study of human diversity, which we will consider in more detail later in this course.

1.4 Inferential Statistics: Generalizing from our observations

Video: Inferential Statistics

Inferential statistics is the process of drawing conclusions about a group of interest (called a population) using a limited set of data (called a sample). Fundamentally, inferential statistics uses probability theory and logic that allow you to make conclusions about populations.

We will cover a number of inferential stat techniques in this course. These include the *t*-test, ANOVA, multiple regression, and others. Other terms associated with inferential statistics (we will define and discuss later, for now, just know they are part of inferential statistics) include null hypothesis significance testing (NHST) and Bayesian statistics.

As an example of a population we might want to study, imagine I am interested in studying middle school students' reading comprehension in the United States, and I want to see if it changes over time. To understand this population directly, I would have to measure the reading comprehension of every member. This is impossible. Instead, I take a random sample from the population by mailing surveys to 50 random middle school students with consent of their parents, I can use descriptive statistics to understand my sample data (50 scores) and inferential statistics to generalize the results to the population (millions of scores).

1.4.1 Populations and Samples: Who (or what) the research is about

A population is the entire group of interest. Examples: people, nursing home residents, repeat customers, etc. The population is the group we want to study.

Populations can be any group you want to draw conclusions about. The researcher defines the population, and this frames the entire research project. The findings of a study intending to measure college students may not apply to older adults. The population is the group to which you will generalize your findings.

The descriptive statistics we will cover can be applied to populations. If we can measure everybody and calculate the average, then we have calculated a **population parameter**. A **population distribution**, which is constructed by measuring every member of a population, is called a **census**. Most of the time, our populations of interest are very large, and it is impossible to measure everybody. How could you give a survey to every single college student in the United States? You would first need a list of every college student in the United States. What are some of the problems with conducting a study in this way? You might think of the ethical obstacles, meaning that it is all but guaranteed you would not get every college student to agree to participate. You might also think of the logistical obstacles, such as the time and cost associated with advertising and administering tens of millions of surveys (even digital ones). But even generating a list of the population would be impossible. Imagine these other concerns did not exist and there was such a list of every student in the United States. Would that list be accurate? Put another way, for how long would that list remain accurate? Every day, new students begin college and other students graduate or leave school. A list of all college students in the United States would only be accurate for an instant. In this seemingly-straightforward population example, we see that even the list of members is constantly changing. For all but the smallest populations of people, population-level research is not possible. Even a precise count of such populations are not possible. This hints at a point we will revisit later in the course—sampling and statistics can be useful regardless of the population size. For this reason, sampling is a powerful tool for understanding populations.

A **sample** is a smaller set from the population. The collection of scores from a sample is called a **sample distribution**. When statistics are computed from a sample distribution, they are called sample statistics, or just statistics.

There are many ways to measure a subset of a population; we call the strategy for obtaining a sample the **sampling method**. The best sampling method is **random sampling**. It has a precise definition: A random sample means that every member of the population has an equal chance of being selected. To do this properly, a researcher should generate a list of every member of the population and select from the list at random. To be a truly random sample, every individual selected would have to participate in your study. True random

samples meet this definition but there are other, more practical, sampling techniques that approximate random sampling; the closer to a random sample of the population, the more likely the sample will represent the population.

We can think of random sampling as one end of a spectrum with **convenience sampling** at the other. A researcher using a convenience sample asks whoever is available to participate in the study. The resulting sample is biased due to proximity, availability, and convenience. Put another way, convenience sampling is less systematic and more...well, convenient. The further away from a true random sample, the less likely it is that the sample collected will represent the population.

Inferential statistics are a collection of techniques to make conclusions about populations based on sample data. As you have seen, without this tool, we could never measure all the individuals we would wish to study. If you have not studied inferential statistics before, it may seem surprising, perhaps a bit unbelievable, that we could make conclusions from such little data. Inferential statistics is not magic; it does not guarantee perfect conclusions. We will see that researchers make certain assumptions when they use inferential statistics and they generate tentative conclusions. Often, the data suggest an answer rather than provide a definitive answer. Sometimes, the research results in more new questions than answers. These features suggest that science is challenging and takes skill to be done well. One of the goals of this course is to help you be a better researcher and a better evaluator of others' research.

Psychologists and others who study people often take for granted that the **units of analysis** are people. Often, this is the case. Through this lens, populations are groups of people and samples are made up of people. Nothing about statistics requires our observations to be about people, however. We could just as easily measure the number of miles a tire will last before it fails or the loudness of a lion's roar. We can also measure collections of people, such as the performance of a company, the frequency of communication of team members, or the outcomes of students in a school.

1.4.2 Constructs provide the context

The idea or concept represented by our data is called the **construct**. There is an important distinction between constructs and measures. A construct is a “concept, model, or schematic idea” (Shadish, Cook, & Campbell, 2002, p. 506). Constructs are the big ideas that researchers are interested in measuring: depression, patient outcomes, prevalence of cumulative trauma disorders, or even sales. For constructs in the social sciences, there is often disagreement and debate about how to define a construct. To do science, we must be able to quantify our observations (collect data) on the constructs. To go from a construct (the idea) to a measure requires an operational definition. An *operational definition* describes how a construct is measured.

Table 1.2: Admission figures for the six largest departments by gender

Department	Male Applicants	Male Percent Admitted	Female Applicants	Female Percent admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

1.5 The cautionary tale of Simpson's paradox

Text by Navarro (2018)

The following is a true story (I think...). In 1973, the University of California, Berkeley had some worries about the admissions of students into their post-graduate courses. Specifically, the thing that caused the problem was that the gender breakdown of their admissions, which looked like this...

	Number of applicants	Percent admitted
Males	8442	46%
Females	4321	35%

...and they were worried about being sued.¹ Given that there were nearly 13,000 applicants, a difference of 9% in admission rates between males and females is just way too big to be a coincidence. Pretty compelling data, right? And if I were to say to you that these data *actually* reflect a weak bias in favour of women (sort of!), you'd probably think that I was either crazy or sexist.

Oddly, it's actually sort of true ...when people started looking more carefully at the admissions data (Bickel et al., 1975) they told a rather different story. Specifically, when they looked at it on a department by department basis, it turned out that most of the departments actually had a slightly *higher* success rate for female applicants than for male applicants. Table 1.2 shows the admission figures for the six largest departments (with the names of the departments removed for privacy reasons):

Remarkably, most departments had a *higher* rate of admissions for females than for males! Yet the overall rate of admission across the university for females was *lower* than for males. How can this be? How can both of these statements be true at the same time?

¹Earlier versions of these notes incorrectly suggested that they actually were sued – apparently that's not true. There's a nice commentary on this here: <https://www.refsmmat.com/posts/2016-05-08-simpsons-paradox-berkeley.html>. A big thank you to Wilfried Van Hirtum for pointing this out to me!

Here's what's going on. Firstly, notice that the departments are *not* equal to one another in terms of their admission percentages: some departments (e.g., engineering, chemistry) tended to admit a high percentage of the qualified applicants, whereas others (e.g., English) tended to reject most of the candidates, even if they were high quality. So, among the six departments shown above, notice that department A is the most generous, followed by B, C, D, E and F in that order. Next, notice that males and females tended to apply to different departments. If we rank the departments in terms of the total number of male applicants, we get **A>B>D>C>F>E** (the “easy” departments are in bold). On the whole, males tended to apply to the departments that had high admission rates. Now compare this to how the female applicants distributed themselves. Ranking the departments in terms of the total number of female applicants produces a quite different ordering **C>E>D>F>A>B**. In other words, what these data seem to be suggesting is that the female applicants tended to apply to “harder” departments. And in fact, if we look at all Figure 1.1 we see that this trend is systematic, and quite striking. This effect is known as Simpson’s paradox. It’s not common, but it does happen in real life, and most people are very surprised by it when they first encounter it, and many people refuse to even believe that it’s real. It is very real. And while there are lots of very subtle statistical lessons buried in there, I want to use it to make a much more important point ...doing research is hard, and there are *lots* of subtle, counterintuitive traps lying in wait for the unwary. That’s reason #2 why scientists love statistics, and why we teach research methods. Because science is hard, and the truth is sometimes cunningly hidden in the nooks and crannies of complicated data.

Before leaving this topic entirely, I want to point out something else really critical that is often overlooked in a research methods class. Statistics only solves *part* of the problem. Remember that we started all this with the concern that Berkeley’s admissions processes might be unfairly biased against female applicants. When we looked at the “aggregated” data, it did seem like the university was discriminating against women, but when we “disaggregate” and looked at the individual behaviour of all the departments, it turned out that the actual departments were, if anything, slightly biased in favour of women. The gender bias in total admissions was caused by the fact that women tended to self-select for harder departments. From a legal perspective, that would probably put the university in the clear. Postgraduate admissions are determined at the level of the individual department (and there are good reasons to do that), and at the level of individual departments, the decisions are more or less unbiased (the weak bias in favour of females at that level is small, and not consistent across departments). Since the university can’t dictate which departments people choose to apply to, and the decision making takes place at the level of the department it can hardly be held accountable for any biases that those choices produce.

That was the basis for my somewhat glib remarks earlier, but that’s not exactly the whole story, is it? After all, if we’re interested in this from a more sociological and psychological perspective, we might want to ask *why* there are

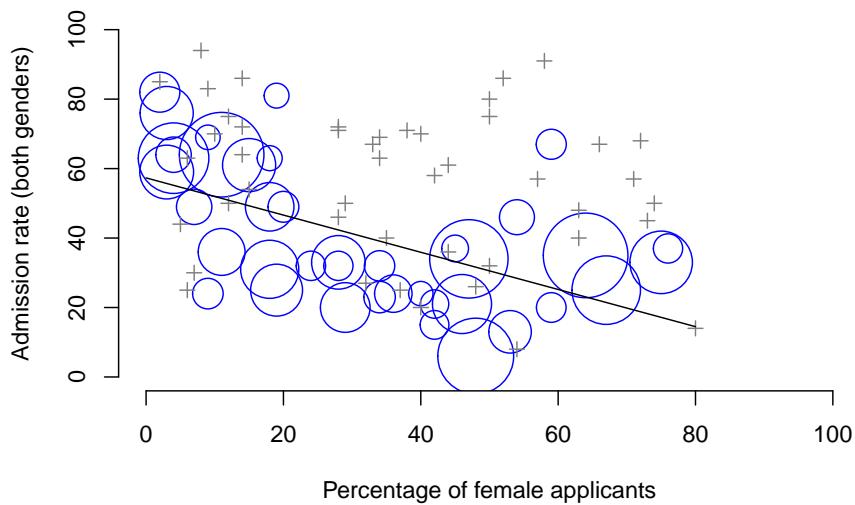


Figure 1.1: The Berkeley 1973 college admissions data. This figure plots the admission rate for the 85 departments that had at least one female applicant, as a function of the percentage of applicants that were female. The plot is a redrawing of Figure 1 from Bickel et al. (1975). Circles plot departments with more than 40 applicants; the area of the circle is proportional to the total number of applicants. The crosses plot department with fewer than 40 applicants.

such strong gender differences in applications. Why do males tend to apply to engineering more often than females, and why is this reversed for the English department? And why is it the case that the departments that tend to have a female-application bias tend to have lower overall admission rates than those departments that have a male-application bias? Might this not still reflect a gender bias, even though every single department is itself unbiased? It might. Suppose, hypothetically, that males preferred to apply to “hard sciences” and females prefer “humanities”. And suppose further that the reason for why the humanities departments have low admission rates is because the government doesn’t want to fund the humanities (Ph.D. places, for instance, are often tied to government funded research projects). Does that constitute a gender bias? Or just an unenlightened view of the value of the humanities? What if someone at a high level in the government cut the humanities funds because they felt that the humanities are “useless chick stuff”. That seems pretty *blatantly* gender biased. None of this falls within the purview of statistics, but it matters to the research project. If you’re interested in the overall structural effects of subtle gender biases, then you probably want to look at *both* the aggregated and disaggregated data. If you’re interested in the decision making process at Berkeley itself then you’re probably only interested in the disaggregated data.

In short there are a lot of critical questions that you can’t answer with statistics, but the answers to those questions will have a huge impact on how you analyse and interpret data. And this is the reason why you should always think of statistics as a *tool* to help you learn about your data, no more and no less. It’s a powerful tool to that end, but there’s no substitute for careful thought.

1.6 A brief introduction to research design

Text by Navarro (2018)

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

– Sir Ronald Fisher²

Note that this section is “special” in two ways. Firstly, it’s much more psychology-specific than the later chapters. Secondly, it focuses much more heavily on the scientific problem of research methodology, and much less on the statistical problem of data analysis. Nevertheless, the two problems are related to one another, so it’s traditional for stats textbooks to discuss the problem in a little detail. This chapter relies heavily on Campbell and Stanley (1963) for the discussion of study design, and Stevens (1946) for the discussion of scales of measurement. Later versions will attempt to be more precise in the citations.

²Presidential Address to the First Indian Statistical Congress, 1938. Source: http://en.wikiquote.org/wiki/Ronald_Fisher

1.6.1 Some thoughts about psychological measurement

Measurement itself is a subtle concept, but basically it comes down to finding some way of assigning numbers, or labels, or some other kind of well-defined descriptions to “stuff”. So, any of the following would count as a psychological measurement:

- My **age** is *33 years*.
- I *do not* like anchovies.
- My **chromosomal gender** is *male*.
- My **self-identified gender** is *male*.³

In the short list above, the **bolded part** is “the thing to be measured”, and the *italicised part* is “the measurement itself”. In fact, we can expand on this a little bit, by thinking about the set of possible measurements that could have arisen in each case:

- My **age** (in years) could have been *0, 1, 2, 3 ...*, etc. The upper bound on what my age could possibly be is a bit fuzzy, but in practice you’d be safe in saying that the largest possible age is *150*, since no human has ever lived that long.
- When asked if I **like anchovies**, I might have said that *I do*, or *I do not*, or *I have no opinion*, or *I sometimes do*.
- My **chromosomal gender** is almost certainly going to be *male (XY)* or *female (XX)*, but there are a few other possibilities. I could also have *Klinefelter's syndrome (XXY)*, which is more similar to male than to female. And I imagine there are other possibilities too.
- My **self-identified gender** is also very likely to be *male* or *female*, but it doesn’t have to agree with my chromosomal gender. I may also choose to identify with *neither*, or to explicitly call myself *transgender*.

As you can see, for some things (like age) it seems fairly obvious what the set of possible measurements should be, whereas for other things it gets a bit tricky.

³Well... now this is awkward, isn't it? This section is one of the oldest parts of the book, and it's outdated in a rather embarrassing way. I wrote this in 2010, at which point all of those facts *were* true. Revisiting this in 2018... well I'm not 33 any more, but that's not surprising I suppose. I can't imagine my chromosomes have changed, so I'm going to guess my karyotype was then and is now XY. The self-identified gender, on the other hand... ah. I suppose the fact that the title page now refers to me as Danielle rather than Daniel might possibly be a giveaway, but I don't typically identify as "male" on a gender questionnaire these days, and I prefer "she/her" pronouns as a default (it's a long story)! I did think a little about how I was going to handle this in the book, actually. The book has a somewhat distinct authorial voice to it, and I feel like it would be a rather different work if I went back and wrote everything as Danielle and updated all the pronouns in the work. Besides, it would be a lot of work, so I've left my name as "Dan" throughout the book, and in any case "Dan" is a perfectly good nickname for "Danielle", don't you think? In any case, it's not a big deal. I only wanted to mention it to make life a little easier for readers who aren't sure how to refer to me. I still don't like anchovies though :-)

But I want to point out that even in the case of someone's age, it's much more subtle than this. For instance, in the example above, I assumed that it was okay to measure age in years. But if you're a developmental psychologist, that's way too crude, and so you often measure age in *years and months* (if a child is 2 years and 11 months, this is usually written as "2;11"). If you're interested in newborns, you might want to measure age in *days since birth*, maybe even *hours since birth*. In other words, the way in which you specify the allowable measurement values is important.

Looking at this a bit more closely, you might also realise that the concept of "age" isn't actually all that precise. In general, when we say "age" we implicitly mean "the length of time since birth". But that's not always the right way to do it. Suppose you're interested in how newborn babies control their eye movements. If you're interested in kids that young, you might also start to worry that "birth" is not the only meaningful point in time to care about. If Baby Alice is born 3 weeks premature and Baby Bianca is born 1 week late, would it really make sense to say that they are the "same age" if we encountered them "2 hours after birth"? In one sense, yes: by social convention, we use birth as our reference point for talking about age in everyday life, since it defines the amount of time the person has been operating as an independent entity in the world, but from a scientific perspective that's not the only thing we care about. When we think about the biology of human beings, it's often useful to think of ourselves as organisms that have been growing and maturing since conception, and from that perspective Alice and Bianca aren't the same age at all. So you might want to define the concept of "age" in two different ways: the length of time since conception, and the length of time since birth. When dealing with adults, it won't make much difference, but when dealing with newborns it might.

Moving beyond these issues, there's the question of methodology. What specific "measurement method" are you going to use to find out someone's age? As before, there are lots of different possibilities:

- You could just ask people "how old are you?" The method of self-report is fast, cheap and easy, but it only works with people old enough to understand the question, and some people lie about their age.
- You could ask an authority (e.g., a parent) "how old is your child?" This method is fast, and when dealing with kids it's not all that hard since the parent is almost always around. It doesn't work as well if you want to know "age since conception", since a lot of parents can't say for sure when conception took place. For that, you might need a different authority (e.g., an obstetrician).
- You could look up official records, like birth certificates. This is time consuming and annoying, but it has its uses (e.g., if the person is now dead).

1.6.2 Operationalisation: defining your measurement

Video: Operationalization

All of the ideas discussed in the previous section all relate to the concept of *operationalisation*. To be a bit more precise about the idea, operationalisation is the process by which we take a meaningful but somewhat vague concept, and turn it into a precise measurement. The process of operationalisation can involve several different things:

- Being precise about what you are trying to measure. For instance, does “age” mean “time since birth” or “time since conception” in the context of your research?
- Determining what method you will use to measure it. Will you use self-report to measure age, ask a parent, or look up an official record? If you’re using self-report, how will you phrase the question?
- Defining the set of the allowable values that the measurement can take. Note that these values don’t always have to be numerical, though they often are. When measuring age, the values are numerical, but we still need to think carefully about what numbers are allowed. Do we want age in years, years and months, days, hours? Etc. For other types of measurements (e.g., gender), the values aren’t numerical. But, just as before, we need to think about what values are allowed. If we’re asking people to self-report their gender, what options do we allow them to choose between? Is it enough to allow only “male” or “female”? Do you need an “other” option? Or should we not give people any specific options, and let them answer in their own words? And if you open up the set of possible values to include all verbal responses, how will you interpret their answers?

Operationalisation is a tricky business, and there’s no “one, true way” to do it. The way in which you choose to operationalise the informal concept of “age” or “gender” into a formal measurement depends on what you need to use the measurement for. Often you’ll find that the community of scientists who work in your area have some fairly well-established ideas for how to go about it. In other words, operationalisation needs to be thought through on a case by case basis. Nevertheless, while there are a lot of issues that are specific to each individual research project, there are some aspects to it that are pretty general.

Before moving on, I want to take a moment to clear up our terminology, and in the process introduce one more term. Here are four different things that are closely related to each other:

- **A theoretical construct.** This is the thing that you’re trying to take a measurement of, like “age”, “gender” or an “opinion”. A theoretical construct can’t be directly observed, and often they’re actually a bit vague.

- **A measure.** The measure refers to the method or the tool that you use to make your observations. A question in a survey, a behavioural observation or a brain scan could all count as a measure.
- **An operationalisation.** The term “operationalisation” refers to the logical connection between the measure and the theoretical construct, or to the process by which we try to derive a measure from a theoretical construct.
- **A variable.** Finally, a new term. A variable is what we end up with when we apply our measure to something in the world. That is, variables are the actual “data” that we end up with in our data sets.

In practice, even scientists tend to blur the distinction between these things, but it’s very helpful to try to understand the differences.

1.6.3 The “role” of variables: predictors and outcomes

Okay, I’ve got one last piece of terminology that I need to explain to you before moving away from variables. Normally, when we do some research we end up with lots of different variables. Then, when we analyse our data we usually try to explain some of the variables in terms of some of the other variables. It’s important to keep the two roles “thing doing the explaining” and “thing being explained” distinct. So let’s be clear about this now. Firstly, we might as well get used to the idea of using mathematical symbols to describe variables, since it’s going to happen over and over again. Let’s denote the “to be explained” variable Y , and denote the variables “doing the explaining” as X_1 , X_2 , etc.

Now, when we doing an analysis, we have different names for X and Y , since they play different roles in the analysis. The classical names for these roles are ***independent variable*** (IV) and ***dependent variable*** (DV). The IV is the variable that you use to do the explaining (i.e., X) and the DV is the variable being explained (i.e., Y). The logic behind these names goes like this: if there really is a relationship between X and Y then we can say that Y depends on X , and if we have designed our study “properly” then X isn’t dependent on anything else. However, I personally find those names horrible: they’re hard to remember and they’re highly misleading, because (a) the IV is never actually “independent of everything else” and (b) if there’s no relationship, then the DV doesn’t actually depend on the IV. And in fact, because I’m not the only person who thinks that IV and DV are just awful names, there are a number of alternatives that I find more appealing. The terms that I’ll use in these notes are ***predictors*** and ***outcomes***. The idea here is that what you’re trying to do is use X (the predictors) to make guesses about Y (the outcomes).⁴ This is summarised in Table 1.3.

⁴Annoyingly, though, there’s a lot of different names used out there. I won’t list all of them – there would be no point in doing that – other than to note that R often uses “response variable” where I’ve used “outcome”, and a traditionalist would use “dependent variable”. Sigh. This sort of terminological confusion is very common, I’m afraid.

Table 1.3: The terminology used to distinguish between different roles that a variable can play when analysing a data set. Note that this book will tend to avoid the classical terminology in favour of the newer names.

role of the variable	classical name	modern name
to be explained	dependent variable (DV)	outcome
to do the explaining	independent variable (IV)	predictor

1.6.4 Experimental and non-experimental research

Video: Experimental, Quasi-, and Non-Experimental Research Designs

One of the big distinctions that you should be aware of is the distinction between “experimental research” and “non-experimental research”. When we make this distinction, what we’re really talking about is the degree of control that the researcher exercises over the people and events in the study.

1.6.4.1 Experimental research

The key features of **experimental research** is that the researcher controls all aspects of the study, especially what participants experience during the study. In particular, the researcher manipulates or varies the predictor variables (IVs), and then allows the outcome variable (DV) to vary naturally. The idea here is to deliberately vary the predictors (IVs) to see if they have any causal effects on the outcomes. Moreover, in order to ensure that there’s no chance that something other than the predictor variables is causing the outcomes, everything else is kept constant or is in some other way “balanced” to ensure that they have no effect on the results. In practice, it’s almost impossible to *think* of everything else that might have an influence on the outcome of an experiment, much less keep it constant. The standard solution to this is **randomisation**: that is, we randomly assign people to different groups, and then give each group a different treatment (i.e., assign them different values of the predictor variables). We’ll talk more about randomisation later in this course, but for now, it’s enough to say that what randomisation does is minimise (but not eliminate) the chances that there are any systematic difference between groups.

Let’s consider a very simple, completely unrealistic and grossly unethical example. Suppose you wanted to find out if smoking causes lung cancer. One way to do this would be to find people who smoke and people who don’t smoke, and look to see if smokers have a higher rate of lung cancer. This is *not* a proper experiment, since the researcher doesn’t have a lot of control over who is and isn’t a smoker. And this really matters: for instance, it might be that people who choose to smoke cigarettes also tend to have poor diets, or maybe they tend to work in asbestos mines, or whatever. The point here is that the groups

(smokers and non-smokers) actually differ on lots of things, not *just* smoking. So it might be that the higher incidence of lung cancer among smokers is caused by something else, not by smoking per se. In technical terms, these other things (e.g. diet) are called “confounds”, and we’ll talk about those in just a moment.

In the meantime, let’s now consider what a proper experiment might look like. Recall that our concern was that smokers and non-smokers might differ in lots of ways. The solution, as long as you have no ethics, is to *control* who smokes and who doesn’t. Specifically, if we randomly divide participants into two groups, and force half of them to become smokers, then it’s very unlikely that the groups will differ in any respect other than the fact that half of them smoke. That way, if our smoking group gets cancer at a higher rate than the non-smoking group, then we can feel pretty confident that (a) smoking does cause cancer and (b) we’re murderers.

1.7 Causality, Research, and Statistics

Text by David Schuster

Video: Causality

1.7.1 Experimental, Quasi-Experimental, and Non-Experimental Studies

In this section, I would like to add a bit more precision to the general concepts explained by Navarro (2018).

Research psychology is a process of identifying constructs and describing how they relate to other constructs. We can classify research designs as experiments, quasi-experiments, and non-experiments.

Experiments are the only kind of research that shows causal relationships (that is, that construct A causes a change in construct B). So an experiment could show if smoking causes lung cancer. To do this, experiments need two things (or they are not experiments)

All experiments have a manipulation. This means that the experimenter changes something within the environment of the experiment (called an independent variable) to see if it causes a change in the outcome (called a dependent variable). For our smoking example, a manipulation would be assigning one group of participants to a lifetime of smoking and another group of participants to a lifetime of no smoking.

Experiments require random assignment. The experimenter decides when to vary the levels of the manipulation (change the manipulation) based on random assignment. Random assignment means that every participant has the same

chance as being in one condition as another. For our smoking example, random assignment means each participant has a 50% chance of being in the smoking group.

As may be clear from the smoking example, we cannot always do experiments because of ethical (it would be wrong to assign people to smoke) or practical reasons (you cannot randomly assign people to genders, for example). The solution is a quasi- or non-experimental study.

In summary: experiments are powerful because they uniquely demonstrate causality (causal relationships). However, experiments require a manipulation and random assignment, which are not always possible.

In a quasi-experimental study, there is a manipulation but no random assignment. Whenever participants are assigned to levels of a manipulation non-randomly, the research is quasi-experimental. In a quasi-experimental smoking study, we could ask people if they had smoked before and assign them to smoking or non-smoking groups based on that answer.

In summary: quasi-experiments do not require random assignment, but they do not show causal relationships.

In a non-experimental study, no manipulation is done. If you want to look at the effects of gender on lung cancer, you would simply observe (collect data on) the genders of patients. By only observing, you would not be manipulating gender.

The differences between quasi- and non-experimental studies are sometimes slight (Pedhauzer & Schmelkin, 1991); if the researcher is manipulating an IV, then the work is quasi-experimental.

In summary: non-experimental studies are observational. Like quasi-experimental studies, they do not show causal relationships.

It's worth repeating that only experiments demonstrate causality. Quasi- and non-experiments can show that a relationship exists but do not say whether one variable causes the other. Any non-causal relationship has three possible explanations:

1. $A \rightarrow B$ one variable causes another; in an experiment, this is the only explanation
2. $B \leftarrow A$ the relationship is reversed; the first variable is actually the outcome
3. $C \rightarrow A; C \rightarrow B$ a third variable exists that was not measured in the study; the third variable causes a change in both A and B. There are many 'C' variables, potentially.

In a non-experimental smoking study, you could not say whether smoking causes lung cancer or people who are predisposed to lung cancer are more likely to smoke. A third possibility is that a separate, third variable causes both lung cancer and a desire to smoke.

1.7.2 Demonstrating Causality

In the 19th century, John Stewart Mill said that we could be satisfied that a relationship is causal if the following three things could be demonstrated:

1. The cause preceded the effect
2. The cause was related to the effect
3. We can find no plausible alternative explanation for the effect other than the cause

Experiments aim to identify causal relationships by manipulating something, observing the outcome, seeing a relationship, and using various methods to reduce other explanations.

1.7.3 Statistics and Causality

Statistics are an important tool for establishing causality, but it's important to know that the choice of statistical technique does not affect the level of causal evidence; demonstrating causality is the job of the research design, not the statistics.

A common misconception arises from the term correlational research design, which people use as a label for quasi-experimental and non-experimental research. It is easy to confuse this term with correlation which is a statistical technique.

Recall that statistics has two branches: Descriptive stats provides tools to summarize variability. Inferential stats provides tools for generalizing samples to populations.

To demonstrate causality, we need to satisfy Mill's second requirement. Inferential statistics can help us do that. Two techniques are particularly useful: correlation (and its statistic r) and the t -test (and its statistic, t). Next, we will see how these techniques work.

1.7.4 Validity and Reliability

Text by David Schuster

1.7.4.1 Define validity and reliability

Reliability and validity are fundamental to critiquing psychological research and to developing your own high-quality research. There are different types of validity and reliability that are relevant to us, which sometimes confuses people.

Because of this, introductory textbooks often present convoluted definitions of these concepts. Fortunately, the real definitions are simple:

Reliability means consistency. Something is reliable if it is consistent. The more consistency, the more reliability.

Validity means truth. Something is valid if it is true. Truth is either-or; there is no such thing as “more true” or “less true.”

In other words, good psychological science requires certain types of consistency and for some of the claims we make to be true. Next, we will look at the specific kinds of reliability and validity that are important for scientists.

1.7.4.2 Types of consistency = Types of Reliability

Reliability

Here are arguably the three most important types of reliability:

Type	Situation of Re- lia- bil- ity	Definition	How to assess
Test-retest	You administer a measure to a participant, then wait some period of time, and give them the test again. The participant’s true score on the measure has not changed (e.g., IQ, personality).	The extent to which a measure is consistent across different administrations	Look for a correlation between the two administrations
Interrater	A measure involves two or more raters who record subjective observations (e.g., counting the number of times a participant has a tic, counting the number of times a married couple shows affection)	The extent to which two observers are consistent in their ratings	Look for a correlation between the two raters
Internal consistency	You are measuring a construct using several items (e.g., five items all rating your enjoyment of a course)	The extent to which items on a measure are consistent with each other; expected if the items measure the same construct	Cronbach’s alpha (.7 is acceptable, .8 is good, and .9 is excellent)

1.7.4.3 Validity is a property of inferences

Video: Validity & Threats

Validity is a specific kind of truth. Validity is the truth of an inference, or a claim. In other words, validity is a property of inferences. An inference (a claim) is valid if it is true.

For example, I could claim that the earth is round. Hopefully, it is a claim that you accept as being true. If you agree, then you could label my claim as valid.

Validity in research is frequently misunderstood, which leads to bizarre and confusing definitions of validity. There is no such thing as “a valid study.” Only claims about the study are valid or not. There is also no such thing as “a valid researcher.” A researcher can make claims. Only the researcher’s claims are valid or not. There is also no such thing as “more valid” or “increasing validity.” Validity is truth of a claim. Either a claim is true, or it is not.

For better or for worse, we usually don’t know with 100% certainty if a claim is true or false (if we did, we wouldn’t need the research). Therefore, research methods get very interesting when we listen to other researcher’s claims and then debate if we agree with them or not. When we do this, we are evaluating the validity of claims made about the study. Next, let’s look at different types of claims (inferences) that are made in research.

1.7.4.4 Types of inferences in a study = Types of validity

Here are some of the most important types of validity.

Type of Validity	Type of Claim	Definition	Example claim
Construct validity	The study operations represent the constructs of interest	The truth of claims that study operations match study constructs	“The Stanford-Binet was used to measure IQ”
Internal validity	The study IV caused a change in the study DV	The truth of claims that the IV causes changes in the DV	“The control group reported lower levels of stress than the experimental group, suggesting that the manipulation raised stress.”

Type of Validity	Type of Claim	Definition	Example claim
External validity	The study results apply to situation X	The truth of claims that the findings will apply as participants/units/variables/contexts change.	“Although data were collected from college students, a similar effect would be working adults.”
Statistical conclusion validity	The statistical analysis was significant or not significant	The truth of claims about the size and direction of the relationship between the IV and the DV. Or, that the statistical results are correct.	“ $p < .05$, indicating a significant difference”

Finally, you might encounter these other types of validity, but they are less clearly defined and evaluated:

- Content validity: The truth of claims that a measure adequately samples (includes the important elements of) the domain of interest. For example, if IQ includes both verbal and math ability, an IQ test would need to have both verbal and math items.
- Face validity: The truth of claims that a study operation “seems like” the construct. For example, a study about distractions from mobile devices might not support claims of “seeming real” if the phone in the study is a paper mockup.
- Criterion validity: The truth of claims that a measure can predict or correlate with some outcome of interest. A personality test as part of a job application would have criterion validity if it predicted applicants’ success in the job.

1.7.4.5 Threats to validity

Threats to validity are specific reasons why an inference about a study is wrong. They can help us anticipate problems in the design of our own research. The best way to address threats to validity is to change the design of our research. Understanding threats to validity also helps you critique research done by others.

Chapter 2

Getting started with R

Text by Navarro (2018)

2.1 Videos

Video: RStudio for the Total Beginner

Video: Jump Start Guide to R

Video: Robot Metaphor for R

2.2 Introduction

Robots are nice to work with.

–Roger Zelazny¹

In this chapter I'll discuss how to get started in R. I'll briefly talk about how to download and install R, but most of the chapter will be focused on getting you started typing R commands. Our goal in this chapter is not to learn any statistical concepts: we're just trying to learn the basics of how R works and get comfortable interacting with the system. To do this, we'll spend a bit of time using R as a simple calculator, since that's the easiest thing to do with R. In doing so, you'll get a bit of a feel for what it's like to work in R. From there I'll introduce some very basic programming ideas: in particular, I'll talk about the idea of defining *variables* to store information, and a few things that you can do with these variables.

¹Source: *Dismal Light* (1968).

However, before going into any of the specifics, it's worth talking a little about why you might want to use R at all. Given that you're reading this, you've probably got your own reasons. However, if those reasons are "because that's what my stats class uses", it might be worth explaining a little why your lecturer has chosen to use R for the class. Of course, I don't really know why *other* people choose R, so I'm really talking about why I use it.

- It's sort of obvious, but worth saying anyway: doing your statistics on a computer is faster, easier and more powerful than doing statistics by hand. Computers excel at mindless repetitive tasks, and a lot of statistical calculations are both mindless and repetitive. For most people, the only reason to ever do statistical calculations with pencil and paper is for learning purposes. In my class I do occasionally suggest doing some calculations that way, but the only real value to it is pedagogical. It does help you to get a "feel" for statistics to do some calculations yourself, so it's worth doing it once. But only once!
- Doing statistics in a spreadsheet (e.g., Microsoft Excel) is generally a bad idea in the long run. Although many people are likely feel more familiar with them, spreadsheets are very limited in terms of what analyses they allow you do. If you get into the habit of trying to do your real life data analysis using spreadsheets, then you've dug yourself into a very deep hole.
- Avoiding proprietary software is a very good idea. There are a lot of commercial packages out there that you can buy, some of which I like and some of which I don't. They're usually very glossy in their appearance, and generally very powerful (much more powerful than spreadsheets). However, they're also very expensive: usually, the company sells "student versions" (limited versions of the real thing) very cheaply; they sell full powered "educational versions" at a price that makes me wince; and they sell commercial licences with a staggeringly high price tag. The business model here is to suck you in during your student days, and then leave you dependent on their tools when you go out into the real world. It's hard to blame them for trying, but personally I'm not in favour of shelling out thousands of dollars if I can avoid it. And you can avoid it: if you make use of packages like R that are open source and free, you never get trapped having to pay exorbitant licensing fees.
- Something that you might not appreciate now, but will love later on if you do anything involving data analysis, is the fact that R is highly extensible. When you download and install R, you get all the basic "packages", and those are very powerful on their own. However, because R is so open and so widely used, it's become something of a standard tool in statistics, and so lots of people write their own packages that extend the system. And these are freely available too. One of the consequences of this, I've noticed, is that if you open up an advanced textbook (a recent one, that is) rather than introductory textbooks, is that a *lot* of them use R. In other words, if you learn how to do your basic statistics in R, then you're a lot closer to being able to use the state of the art methods than you would be if

you'd started out with a "simpler" system: so if you want to become a genuine expert in psychological data analysis, learning R is a very good use of your time.

- Related to the previous point: R is a real programming language. As you get better at using R for data analysis, you're also learning to program. To some people this might seem like a bad thing, but in truth, programming is a core research skill across a lot of the social and behavioural sciences. Think about how many surveys and experiments are done online, or presented on computers. Think about all those online social environments which you might be interested in studying; and maybe collecting data from in an automated fashion. Think about artificial intelligence systems, computer vision and speech recognition. If any of these are things that you think you might want to be involved in – as someone "doing research in psychology", that is – you'll need to know a bit of programming. And if you don't already know how to program, then learning how to do statistics using R is a nice way to start.

Those are the main reasons I use R. It's not without its flaws: it's not easy to learn, and it has a few very annoying quirks to it that we're all pretty much stuck with, but on the whole I think the strengths outweigh the weakness; more so than any other option I've encountered so far.

2.3 Installing R

Okay, enough with the sales pitch. Let's get started. Just as with any piece of software, R needs to be installed on a "computer", which is a magical box that does cool things and delivers free ponies. Or something along those lines: I may be confusing computers with the iPad marketing campaigns. Anyway, R is freely distributed online, and you can download it from the R homepage, which is:

<http://cran.r-project.org/>

At the top of the page – under the heading "Download and Install R" – you'll see separate links for Windows users, Mac users, and Linux users. If you follow the relevant link, you'll see that the online instructions are pretty self-explanatory, but I'll walk you through the installation anyway. As of this writing, the current version of R is 3.0.2 ("Frisbee Sailing"), but they usually issue updates every six months, so you'll probably have a newer version.²

²Although R is updated frequently, it doesn't usually make much of a difference for the sort of work we'll do in this book. In fact, during the writing of the book I upgraded several times, and didn't have to change much except these sections describing the downloading.

2.3.1 Installing R on a Windows computer

The CRAN homepage changes from time to time, and it's not particularly pretty, or all that well-designed quite frankly. But it's not difficult to find what you're after. In general you'll find a link at the top of the page with the text "Download R for Windows". If you click on that, it will take you to a page that offers you a few options. Again, at the very top of the page you'll be told to click on a link that says to click here if you're installing R for the first time. That's probably what you want. This will take you to a page that has a prominent link at the top called "Download R 3.0.2 for Windows". That's the one you want. Click on that and your browser should start downloading a file called `R-3.0.2-win.exe`, or whatever the equivalent version number is by the time you read this. The file for version 3.0.2 is about 54MB in size, so it may take some time depending on how fast your internet connection is. Once you've downloaded the file, double click to install it. As with any software you download online, Windows will ask you some questions about whether you trust the file and so on. After you click through those, it'll ask you where you want to install it, and what components you want to install. The default values should be fine for most people, so again, just click through. Once all that is done, you should have R installed on your system. You can access it from the Start menu, or from the desktop if you asked it to add a shortcut there. You can now open up R in the usual way if you want to, but what I'm going to suggest is that instead of doing that you should now install RStudio (see Section 2.3.4 for instructions).

2.3.2 Installing R on a Mac

When you click on the Mac OS X link, you should find yourself on a page with the title "R for Mac OS X". The vast majority of Mac users will have a fairly recent version of the operating system: as long as you're running Mac OS X 10.6 (Snow Leopard) or higher, then you'll be fine.³ There's a fairly prominent link on the page called "`R-3.0.2.pkg`", which is the one you want. Click on that link and you'll start downloading the installer file, which is (not surprisingly) called `R-3.0.2.pkg`. It's about 61MB in size, so the download can take a while on slower internet connections.

Once you've downloaded `R-3.0.2.pkg`, all you need to do is open it by double clicking on the package file. The installation should go smoothly from there: just follow all the instructions just like you usually do when you install something. Once it's finished, you'll find a file called `R.app` in the Applications folder. You can now open up R in the usual way⁴ if you want to, but what I'm going to

³If you're running an older version of the Mac OS, then you need to follow the link to the "old" page (<http://cran.r-project.org/bin/macosx/old/>). You should be able to find the installer file that you need at the bottom of the page.

⁴Tip for advanced Mac users. You can run R from the terminal if you want to. The com-

suggest is that instead of doing that you should now install RStudio (see Section 2.3.4 for instructions).

2.3.3 Installing R on a Linux computer

If you’re successfully managing to run a Linux box, regardless of what distribution, then you should find the instructions on the website easy enough. You can compile R from source yourself if you want, or install it through your package management system, which will probably have R in it. Alternatively, the CRAN site has precompiled binaries for Debian, Red Hat, Suse and Ubuntu and has separate instructions for each. Once you’ve got R installed, you can run it from the command line just by typing `R`. However, if you’re feeling envious of Windows and Mac users for their fancy GUIs, you can download RStudio too (see Section 2.3.4 for instructions).

2.3.4 Downloading and installing RStudio

Okay, so regardless of what operating system you’re using, the last thing that I told you to do is to download RStudio. To understand why I’ve suggested this, you need to understand a little bit more about R itself. The term R doesn’t really refer to a specific application on your computer. Rather, it refers to the underlying statistical language. You can use this language through lots of different applications. When you install R initially, it comes with one application that lets you do this: it’s the `R.exe` application on a Windows machine, and the `R.app` application on a Mac. But that’s not the only way to do it. There are lots of different applications that you can use that will let you interact with R. One of those is called RStudio, and it’s the one I’m going to suggest that you use. RStudio provides a clean, professional interface to R that I find much nicer to work with than either the Windows or Mac defaults. Like R itself, RStudio is free software: you can find all the details on their webpage. In the meantime, you can download it here:

<http://www.RStudio.org/>

When you visit the RStudio website, you’ll probably be struck by how much cleaner and simpler it is than the CRAN website,⁵ and how obvious it is what you need to do: click the big green button that says “Download”.

mand is just “R”. It behaves like the normal desktop version, except that help documentation behaves like a “man” page instead of opening in a new window.

⁵This is probably no coincidence: the people who design and distribute the core R language itself are focused on technical stuff. And sometimes they almost seem to forget that there’s an actual human user at the end. The people who design and distribute RStudio are focused on user interface. They want to make R as usable as possible. The two websites reflect that difference.

When you click on the download button on the homepage it will ask you to choose whether you want the desktop version or the server version. You want the desktop version. After choosing the desktop version it will take you to a page <http://www.RStudio.org/download/desktop>) that shows several possible downloads: there's a different one for each operating system. However, the nice people at RStudio have designed the webpage so that it automatically recommends the download that is most appropriate for your computer. Click on the appropriate link, and the RStudio installer file will start downloading.

Once it's finished downloading, open the installer file in the usual way to install RStudio. After it's finished installing, you can start R by opening RStudio. You don't need to open R.app or R.exe in order to access R. RStudio will take care of that for you. To illustrate what RStudio looks like, Figure 2.1 shows a screenshot of an R session in progress. In this screenshot, you can see that it's running on a Mac, but it looks almost identical no matter what operating system you have. The Windows version looks more like a Windows application (e.g., the menus are attached to the application window and the colour scheme is slightly different), but it's more or less identical. There are a few minor differences in where things are located in the menus (I'll point them out as we go along) and in the shortcut keys, because RStudio is trying to "feel" like a proper Mac application or a proper Windows application, and this means that it has to change its behaviour a little bit depending on what computer it's running on. Even so, these differences are very small: I started out using the Mac version of RStudio and then started using the Windows version as well in order to write these notes.

The only "shortcoming" I've found with RStudio is that – as of this writing – it's still a work in progress. The "problem" is that they keep improving it. New features keep turning up the more recent releases, so there's a good chance that by the time you read this book there will be a version out that has some really neat things that weren't in the version that I'm using now.

2.3.5 Starting up R

One way or another, regardless of what operating system you're using and regardless of whether you're using RStudio, or the default GUI, or even the command line, it's time to open R and get started. When you do that, the first thing you'll see (assuming that you're looking at the **R console**, that is) is a whole lot of text that doesn't make much sense. It should look something like this:

```
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"  
Copyright (C) 2013 The R Foundation for Statistical Computing  
Platform: x86_64-apple-darwin10.8.0 (64-bit)
```

R is free software and comes with ABSOLUTELY NO WARRANTY.

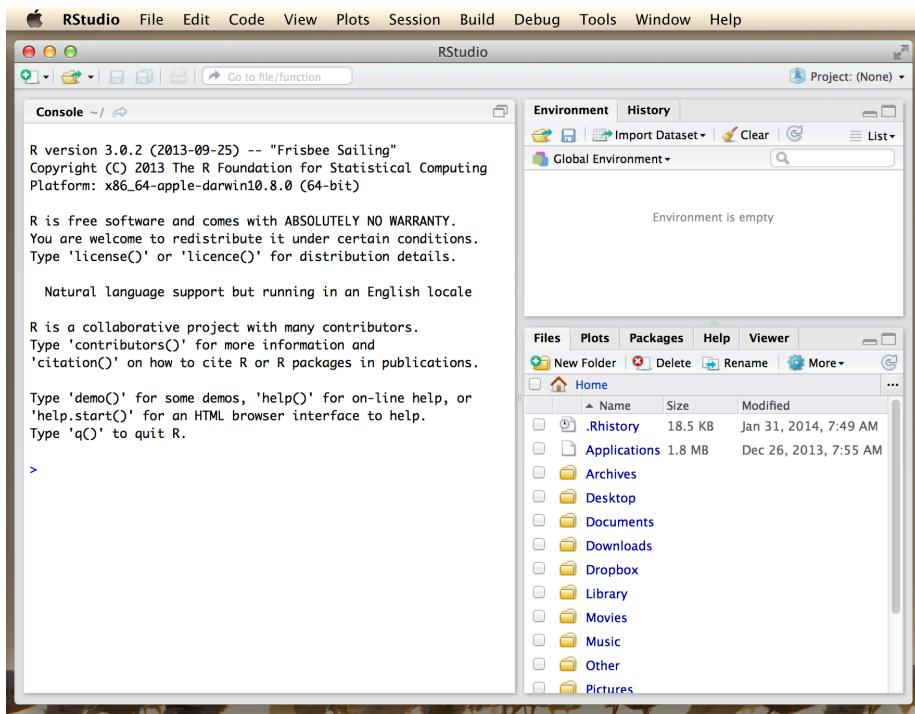


Figure 2.1: An R session in progress running through RStudio. The picture shows RStudio running on a Mac, but the Windows interface is almost identical.

You are welcome to redistribute it under certain conditions.
 Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
 Type 'contributors()' for more information and
 'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
 'help.start()' for an HTML browser interface to help.
 Type 'q()' to quit R.

>

Most of this text is pretty uninteresting, and when doing real data analysis you'll never really pay much attention to it. The important part of it is this...

>

... which has a flashing cursor next to it. That's the ***command prompt***. When you see this, it means that R is waiting patiently for you to do something!

2.4 Typing commands at the R console

See also: video links at the start of this chapter

One of the easiest things you can do with R is use it as a simple calculator, so it's a good place to start. For instance, try typing `10 + 20`, and hitting enter.⁶ When you do this, you've entered a ***command***, and R will "execute" that command. What you see on screen now will be this:

```
> 10 + 20
[1] 30
```

Not a lot of surprises in this extract. But there's a few things worth talking about, even with such a simple example. Firstly, it's important that you understand how to read the extract. In this example, what *I* typed was the `10 + 20` part. I didn't type the `>` symbol: that's just the R command prompt and isn't

⁶Seriously. If you're in a position to do so, open up R and start typing. The simple act of typing it rather than "just reading" makes a big difference. It makes the concepts more concrete, and it ties the abstract ideas (programming and statistics) to the actual context in which you need to use them. Statistics is something you *do*, not just something you read about in a textbook.

part of the actual command. And neither did I type the [1] 30 part. That's what R printed out in response to my command.

Secondly, it's important to understand how the output is formatted. Obviously, the correct answer to the sum 10 + 20 is 30, and not surprisingly R has printed that out as part of its response. But it's also printed out this [1] part, which probably doesn't make a lot of sense to you right now. You're going to see that a lot. I'll talk about what this means in a bit more detail later on, but for now you can think of [1] 30 as if R were saying "the answer to the 1st question you asked is 30". That's not quite the truth, but it's close enough for now. And in any case it's not really very interesting at the moment: we only asked R to calculate one thing, so obviously there's only one answer printed on the screen. Later on this will change, and the [1] part will start to make a bit more sense. For now, I just don't want you to get confused or concerned by it.

2.4.1 An important digression about formatting

Now that I've taught you these rules I'm going to change them pretty much immediately. That is because I want you to be able to copy code from the book directly into R if you want to test things or conduct your own analyses. However, if you copy this kind of code (that shows the command prompt and the results) directly into R you will get an error

```
> 10 + 20  
[1] 30
```

```
## Error: <text>:1:1: unexpected '>'  
## 1: >  
##      ^
```

So instead, I'm going to provide code in a slightly different format so that it looks like this...

```
10 + 20  
  
## [1] 30
```

There are two main differences.

- In your console, you type after the >, but from now on I won't show the command prompt in the book.
- In the book, output is commented out with ##, in your console it appears directly after your code.

These two differences mean that if you’re working with an electronic version of the book, you can easily copy code out of the book and into the console.

So for example if you copied the two lines of code from the book you’d get this

```
10 + 20
```

```
## [1] 30
```

```
## [1] 30
```

2.4.2 Be very careful to avoid typos

Before we go on to talk about other types of calculations that we can do with R, there’s a few other things I want to point out. The first thing is that, while R is good software, it’s still software. It’s pretty stupid, and because it’s stupid it can’t handle typos. It takes it on faith that you meant to type *exactly* what you did type. For example, suppose that you forgot to hit the shift key when trying to type `+`, and as a result your command ended up being `10 = 20` rather than `10 + 20`. Here’s what happens:

```
10 = 20
```

```
## Error in 10 = 20: invalid (do_set) left-hand side to assignment
```

What’s happened here is that R has attempted to interpret `10 = 20` as a command, and spits out an error message because the command doesn’t make any sense to it. When a *human* looks at this, and then looks down at his or her keyboard and sees that `+` and `=` are on the same key, it’s pretty obvious that the command was a typo. But R doesn’t know this, so it gets upset. And, if you look at it from its perspective, this makes sense. All that R “knows” is that 10 is a legitimate number, 20 is a legitimate number, and `=` is a legitimate part of the language too. In other words, from its perspective this really does look like the user meant to type `10 = 20`, since all the individual parts of that statement are legitimate and it’s too stupid to realise that this is probably a typo. Therefore, R takes it on faith that this is exactly what you meant... it only “discovers” that the command is nonsense when it tries to follow your instructions, typo and all. And then it whinges, and spits out an error.

Even more subtle is the fact that some typos won’t produce errors at all, because they happen to correspond to “well-formed” R commands. For instance, suppose that not only did I forget to hit the shift key when trying to type `10 + 20`, I also managed to press the key next to one I meant do. The resulting typo would produce the command `10 - 20`. Clearly, R has no way of knowing that you meant to *add* 20 to 10, not *subtract* 20 from 10, so what happens this time is this:

```
10 - 20
```

```
## [1] -10
```

In this case, R produces the right answer, but to the wrong question.

To some extent, I'm stating the obvious here, but it's important. The people who wrote R are smart. You, the user, are smart. But R itself is dumb. And because it's dumb, it has to be mindlessly obedient. It does *exactly* what you ask it to do. There is no equivalent to "autocorrect" in R, and for good reason. When doing advanced stuff – and even the simplest of statistics is pretty advanced in a lot of ways – it's dangerous to let a mindless automaton like R try to overrule the human user. But because of this, it's your responsibility to be careful. Always make sure you type *exactly what you mean*. When dealing with computers, it's not enough to type "approximately" the right thing. In general, you absolutely *must* be precise in what you say to R ... like all machines it is too stupid to be anything other than absurdly literal in its interpretation.

2.4.3 R is (a bit) flexible with spacing

Of course, now that I've been so uptight about the importance of always being precise, I should point out that there are some exceptions. Or, more accurately, there are some situations in which R does show a bit more flexibility than my previous description suggests. The first thing R is smart enough to do is ignore redundant spacing. What I mean by this is that, when I typed `10 + 20` before, I could equally have done this

```
10 + 20
```

```
## [1] 30
```

or this

```
10+20
```

```
## [1] 30
```

and I would get exactly the same answer. However, that doesn't mean that you can insert spaces in any old place. When we looked at the startup documentation in Section 2.3.5 it suggested that you could type `citation()` to get some information about how to cite R. If I do so...

```
citation()

##
## To cite R in publications use:
##
## R Core Team (2020). R: A language and environment for statistical
## computing. R Foundation for Statistical Computing, Vienna, Austria.
## URL https://www.R-project.org/.
##
## A BibTeX entry for LaTeX users is
##
## @Manual{,
##   title = {R: A Language and Environment for Statistical Computing},
##   author = {{R Core Team}},
##   organization = {R Foundation for Statistical Computing},
##   address = {Vienna, Austria},
##   year = {2020},
##   url = {https://www.R-project.org/},
## }
##
## We have invested a lot of time and effort in creating R, please cite it
## when using it for data analysis. See also 'citation("pkgname")' for
## citing R packages.
```

... it tells me to cite the R manual (R Core Team, 2013). Let's see what happens when I try changing the spacing. If I insert spaces in between the word and the parentheses, or inside the parentheses themselves, then all is well. That is, either of these two commands

```
citation ()
cita tion( )
```

will produce exactly the same response. However, what I can't do is insert spaces in the middle of the word. If I try to do this, R gets upset:

```
citat ion()

## Error: <text>:1:7: unexpected symbol
## 1: citat ion
##           ^
```

Throughout this book I'll vary the way I use spacing a little bit, just to give you a feel for the different ways in which spacing can be used. I'll try not to do it too much though, since it's generally considered to be good practice to be consistent in how you format your commands.

2.4.4 R can sometimes tell that you're not finished yet (but not often)

One more thing I should point out. If you hit enter in a situation where it's "obvious" to R that you haven't actually finished typing the command, R is just smart enough to keep waiting. For example, if you type 10 + and then press enter, even R is smart enough to realise that you probably wanted to type in another number. So here's what happens (for illustrative purposes I'm breaking my own code formatting rules in this section):

```
> 10+
+
```

and there's a blinking cursor next to the plus sign. What this means is that R is still waiting for you to finish. It "thinks" you're still typing your command, so it hasn't tried to execute it yet. In other words, this plus sign is actually another command prompt. It's different from the usual one (i.e., the > symbol) to remind you that R is going to "add" whatever you type now to what you typed last time. For example, if I then go on to type 3 and hit enter, what I get is this:

```
> 10 +
+ 20
[1] 30
```

And as far as R is concerned, this is *exactly* the same as if you had typed 10 + 20. Similarly, consider the `citation()` command that we talked about in the previous section. Suppose you hit enter after typing `citation()`. Once again, R is smart enough to realise that there must be more coming – since you need to add the) character – so it waits. I can even hit enter several times and it will keep waiting:

```
> citation(
+
+
+ )
```

I'll make use of this a lot in this book. A lot of the commands that we'll have to type are pretty long, and they're visually a bit easier to read if I break it up over several lines. If you start doing this yourself, you'll eventually get yourself in trouble (it happens to us all). Maybe you start typing a command, and then you realise you've screwed up. For example,

```
> citblation(
+
+
```

You'd probably prefer R not to try running this command, right? If you want to get out of this situation, just hit the 'escape' key.⁷ R will return you to the normal command prompt (i.e. >) *without* attempting to execute the botched command.

That being said, it's not often the case that R is smart enough to tell that there's more coming. For instance, in the same way that I can't add a space in the middle of a word, I can't hit enter in the middle of a word either. If I hit enter after typing `citat` I get an error, because R thinks I'm interested in an "object" called `citat` and can't find it:

```
> citat
Error: object 'citat' not found
```

What about if I typed `citation` and hit enter? In this case we get something very odd, something that we definitely *don't* want, at least at this stage. Here's what happens:

```
citation
## function (package = "base", lib.loc = NULL, auto = NULL)
## {
##   dir <- system.file(package = package, lib.loc = lib.loc)
##   if (dir == "") 
##     stop(gettextf("package '%s' not found", package), domain = NA)

BLAH BLAH BLAH
```

where the BLAH BLAH BLAH goes on for rather a long time, and you don't know enough R yet to understand what all this gibberish actually means (of course, it doesn't actually say BLAH BLAH BLAH - it says some other things we don't understand or need to know that I've edited for length) This incomprehensible output can be quite intimidating to novice users, and unfortunately it's very easy to forget to type the parentheses; so almost certainly you'll do this by accident. Do not panic when this happens. Simply ignore the gibberish. As you become more experienced this gibberish will start to make sense, and you'll find it quite handy to print this stuff out.⁸ But for now just try to remember to add the parentheses when typing your commands.

⁷If you're running R from the terminal rather than from RStudio, escape doesn't work: use CTRL-C instead.

⁸For advanced users: yes, as you've probably guessed, R is printing out the source code for the function.

Table 2.1: Basic arithmetic operations in R. These five operators are used very frequently throughout the text, so it's important to be familiar with them at the outset.

operation	operator	example input	example output
addition	'+'	10 + 2	12
subtraction	'-'	9 - 3	6
multiplication	'*'	5 * 5	25
division	'/'	10 / 3	3
power	'^'	5 ^ 2	25

2.5 Doing simple calculations with R

Okay, now that we've discussed some of the tedious details associated with typing R commands, let's get back to learning how to use the most powerful piece of statistical software in the world as a \$2 calculator. So far, all we know how to do is addition. Clearly, a calculator that only did addition would be a bit stupid, so I should tell you about how to perform other simple calculations using R. But first, some more terminology. Addition is an example of an "operation" that you can perform (specifically, an arithmetic operation), and the **operator** that performs it is `+`. To people with a programming or mathematics background, this terminology probably feels pretty natural, but to other people it might feel like I'm trying to make something very simple (addition) sound more complicated than it is (by calling it an arithmetic operation). To some extent, that's true: if addition was the only operation that we were interested in, it'd be a bit silly to introduce all this extra terminology. However, as we go along, we'll start using more and more different kinds of operations, so it's probably a good idea to get the language straight now, while we're still talking about very familiar concepts like addition!

2.5.1 Adding, subtracting, multiplying and dividing

So, now that we have the terminology, let's learn how to perform some arithmetic operations in R. To that end, Table 2.1 lists the operators that correspond to the basic arithmetic we learned in primary school: addition, subtraction, multiplication and division.

As you can see, R uses fairly standard symbols to denote each of the different operations you might want to perform: addition is done using the `+` operator, subtraction is performed by the `-` operator, and so on. So if I wanted to find out what 57 times 61 is (and who wouldn't?), I can use R instead of a calculator, like so:

```
57 * 61
```

```
## [1] 3477
```

So that's handy.

2.5.2 Taking powers

The first four operations listed in Table 2.1 are things we all learned in primary school, but they aren't the only arithmetic operations built into R. There are three other arithmetic operations that I should probably mention: taking powers, doing integer division, and calculating a modulus. Of the three, the only one that is of any real importance for the purposes of this book is taking powers, so I'll discuss that one here: the other two are discussed in Chapter ??.

For those of you who can still remember your high school maths, this should be familiar. But for some people high school maths was a long time ago, and others of us didn't listen very hard in high school. It's not complicated. As I'm sure everyone will probably remember the moment they read this, the act of multiplying a number x by itself n times is called "raising x to the n -th power". Mathematically, this is written as x^n . Some values of n have special names: in particular x^2 is called x -squared, and x^3 is called x -cubed. So, the 4th power of 5 is calculated like this:

$$5^4 = 5 \times 5 \times 5 \times 5$$

One way that we could calculate 5^4 in R would be to type in the complete multiplication as it is shown in the equation above. That is, we could do this

```
5 * 5 * 5 * 5
```

```
## [1] 625
```

but it does seem a bit tedious. It would be very annoying indeed if you wanted to calculate 5^{15} , since the command would end up being quite long. Therefore, to make our lives easier, we use the power operator instead. When we do that, our command to calculate 5^4 goes like this:

```
5 ^ 4
```

```
## [1] 625
```

Much easier.

2.5.3 Doing calculations in the right order

Okay. At this point, you know how to take one of the most powerful pieces of statistical software in the world, and use it as a \$2 calculator. And as a bonus, you've learned a few very basic programming concepts. That's not nothing (you could argue that you've just saved yourself \$2) but on the other hand, it's not very much either. In order to use R more effectively, we need to introduce more programming concepts.

In most situations where you would want to use a calculator, you might want to do multiple calculations. R lets you do this, just by typing in longer commands.

⁹ In fact, we've already seen an example of this earlier, when I typed in `5 * 5 * 5 * 5`. However, let's try a slightly different example:

```
1 + 2 * 4
```

```
## [1] 9
```

Clearly, this isn't a problem for R either. However, it's worth stopping for a second, and thinking about what R just did. Clearly, since it gave us an answer of 9 it must have multiplied `2 * 4` (to get an interim answer of 8) and then added 1 to that. But, suppose it had decided to just go from left to right: if R had decided instead to add `1+2` (to get an interim answer of 3) and then multiplied by 4, it would have come up with an answer of 12.

To answer this, you need to know the *order of operations* that R uses. If you remember back to your high school maths classes, it's actually the same order that you got taught when you were at school: the “**BEDMAS**” order.¹⁰ That is, first calculate things inside **B**rackets (), then calculate **E**xponents ^, then **D**ivision / and **M**ultiplication *, then **A**ddition + and **S**ubtraction -. So, to continue the example above, if we want to force R to calculate the `1+2` part before the multiplication, all we would have to do is enclose it in brackets:

```
(1 + 2) * 4
```

```
## [1] 12
```

This is a fairly useful thing to be able to do. The only other thing I should point out about order of operations is what to expect when you have two operations

⁹If you're reading this with R open, a good learning trick is to try typing in a few different variations on what I've done here. If you experiment with your commands, you'll quickly learn what works and what doesn't

¹⁰For advanced users: if you want a table showing the complete order of operator precedence in R, type `?Syntax`. I haven't included it in this book since there are quite a few different operators, and we don't need that much detail. Besides, in practice most people seem to figure it out from seeing examples: until writing this book I never looked at the formal statement of operator precedence for any language I ever coded in, and never ran into any difficulties.

that have the same priority: that is, how does R resolve ties? For instance, multiplication and division are actually the same priority, but what should we expect when we give R a problem like $4 / 2 * 3$ to solve? If it evaluates the multiplication first and then the division, it would calculate a value of two-thirds. But if it evaluates the division first it calculates a value of 6. The answer, in this case, is that R goes from *left to right*, so in this case the division step would come first:

```
4 / 2 * 3
```

```
## [1] 6
```

All of the above being said, it's helpful to remember that *brackets always come first*. So, if you're ever unsure about what order R will do things in, an easy solution is to enclose the thing *you* want it to do first in brackets. There's nothing stopping you from typing $(4 / 2) * 3$. By enclosing the division in brackets we make it clear which thing is supposed to happen first. In this instance you wouldn't have needed to, since R would have done the division first anyway, but when you're first starting out it's better to make sure R does what you want!

2.6 Storing a number as a variable

One of the most important things to be able to do in R (or any programming language, for that matter) is to store information in *variables*. Variables in R aren't exactly the same thing as the variables we talked about in the last chapter on research methods, but they are similar. At a conceptual level you can think of a variable as *label* for a certain piece of information, or even several different pieces of information. When doing statistical analysis in R all of your data (the variables you measured in your study) will be stored as variables in R, but as well see later in the book you'll find that you end up creating variables for other things too. However, before we delve into all the messy details of data sets and statistical analysis, let's look at the very basics for how we create variables and work with them.

2.6.1 Variable assignment using `<-` and `->`

Since we've been working with numbers so far, let's start by creating variables to store our numbers. And since most people like concrete examples, let's invent one. Suppose I'm trying to calculate how much money I'm going to make from this book. There's several different numbers I might want to store. Firstly, I need to figure out how many copies I'll sell. This isn't exactly *Harry Potter*, so let's assume I'm only going to sell one copy per student in my class. That's

350 sales, so let's create a variable called `sales`. What I want to do is assign a *value* to my variable `sales`, and that value should be 350. We do this by using the ***assignment operator***, which is `<-`. Here's how we do it:

```
sales <- 350
```

When you hit enter, R doesn't print out any output.¹¹ It just gives you another command prompt. However, behind the scenes R has created a variable called `sales` and given it a value of 350. You can check that this has happened by asking R to print the variable on screen. And the simplest way to do *that* is to type the name of the variable and hit enter¹².

```
sales
```

```
## [1] 350
```

So that's nice to know. Anytime you can't remember what R has got stored in a particular variable, you can just type the name of the variable and hit enter.

Okay, so now we know how to assign variables. Actually, there's a bit more you should know. Firstly, one of the curious features of R is that there are several different ways of making assignments. In addition to the `<-` operator, we can also use `->` and `=`, and it's pretty important to understand the differences between them.¹³ Let's start by considering `->`, since that's the easy one (we'll discuss the use of `=` in Section 2.7.1). As you might expect from just looking at the symbol, it's almost identical to `<-`. It's just that the arrow (i.e., the assignment) goes from left to right. So if I wanted to define my `sales` variable using `->`, I would write it like this:

```
350 -> sales
```

This has the same effect: and it *still* means that I'm only going to sell 350 copies. Sigh. Apart from this superficial difference, `<-` and `->` are identical. In fact, as far as R is concerned, they're actually the same operator, just in a "left form" and a "right form".¹⁴

¹¹If you are using RStudio, and the "environment" panel (formerly known as the "workspace" panel) is visible when you typed the command, then you probably saw something happening there. That's to be expected, and is quite helpful. However, there's two things to note here (1) I haven't yet explained what that panel does, so for now just ignore it, and (2) this is one of the helpful things RStudio does, not a part of R itself.

¹²As we'll discuss later, by doing this we are implicitly using the `print()` function

¹³Actually, in keeping with the R tradition of providing you with a billion different screw-drivers (even when you're actually looking for a hammer) these aren't the only options. There's also `theassign()` function, and the `<<-` and `->>` operators. However, we won't be using these at all in this book.

¹⁴A quick reminder: when using operators like `<-` and `->` that span multiple characters, you can't insert spaces in the middle. That is, if you type `- >` or `< -`, R will interpret your command the wrong way. And I will cry.

2.6.2 Doing calculations using variables

Okay, let's get back to my original story. In my quest to become rich, I've written this textbook. To figure out how good a strategy is, I've started creating some variables in R. In addition to defining a `sales` variable that counts the number of copies I'm going to sell, I can also create a variable called `royalty`, indicating how much money I get per copy. Let's say that my royalties are about \$7 per book:

```
sales <- 350
royalty <- 7
```

The nice thing about variables (in fact, the whole point of having variables) is that we can do anything with a variable that we ought to be able to do with the information that it stores. That is, since R allows me to multiply 350 by 7

```
350 * 7
```

```
## [1] 2450
```

it also allows me to multiply `sales` by `royalty`

```
sales * royalty
```

```
## [1] 2450
```

As far as R is concerned, the `sales * royalty` command is the same as the `350 * 7` command. Not surprisingly, I can assign the output of this calculation to a new variable, which I'll call `revenue`. And when we do this, the new variable `revenue` gets the value 2450. So let's do that, and then get R to print out the value of `revenue` so that we can verify that it's done what we asked:

```
revenue <- sales * royalty
revenue
```

```
## [1] 2450
```

That's fairly straightforward. A slightly more subtle thing we can do is reassign the value of my variable, based on its current value. For instance, suppose that one of my students (no doubt under the influence of psychotropic drugs) loves the book so much that he or she donates me an extra \$550. The simplest way to capture this is by a command like this:

```
revenue <- revenue + 550
revenue
## [1] 3000
```

In this calculation, R has taken the old value of `revenue` (i.e., 2450) and added 550 to that value, producing a value of 3000. This new value is assigned to the `revenue` variable, overwriting its previous value. In any case, we now know that I'm expecting to make \$3000 off this. Pretty sweet, I thinks to myself. Or at least, that's what I thinks until I do a few more calculation and work out what the implied hourly wage I'm making off this looks like.

2.6.3 Rules and conventions for naming variables

In the examples that we've seen so far, my variable names (`sales` and `revenue`) have just been English-language words written using lowercase letters. However, R allows a lot more flexibility when it comes to naming your variables, as the following list of rules¹⁵ illustrates:

- Variable names can only use the upper case alphabetic characters A-Z as well as the lower case characters a-z. You can also include numeric characters 0-9 in the variable name, as well as the period . or underscore _ character. In other words, you can use `SaL.e_s` as a variable name (though I can't think why you would want to), but you can't use `Sales?`.
- Variable names cannot include spaces: therefore `my sales` is not a valid name, but `my.sales` is.
- Variable names are case sensitive: that is, `Sales` and `sales` are *different* variable names.
- Variable names must start with a letter or a period. You can't use something like `_sales` or `1sales` as a variable name. You can use `.sales` as a variable name if you want, but it's not usually a good idea. By convention, variables starting with a . are used for special purposes, so you should avoid doing so.
- Variable names cannot be one of the reserved keywords. These are special names that R needs to keep "safe" from us mere users, so you can't use them as the names of variables. The keywords are: `if`, `else`, `repeat`, `while`, `function`, `for`, `in`, `next`, `break`, `TRUE`, `FALSE`, `NULL`, `Inf`, `NaN`, `NA`, `NA_integer_`, `NA_real_`, `NA_complex_`, and finally, `NA_character_`. Don't feel especially obliged to memorise these: if you make a mistake and try to use one of the keywords as a variable name, R will complain about it like the whiny little automaton it is.

¹⁵Actually, you can override any of these rules if you want to, and quite easily. All you have to do is add quote marks or backticks around your non-standard variable name. For instance ``my sales` <- 350` would work just fine, but it's almost never a good idea to do this.

In addition to those rules that R enforces, there are some informal conventions that people tend to follow when naming variables. One of them you've already seen: i.e., don't use variables that start with a period. But there are several others. You aren't obliged to follow these conventions, and there are many situations in which it's advisable to ignore them, but it's generally a good idea to follow them when you can:

- Use informative variable names. As a general rule, using meaningful names like `sales` and `revenue` is preferred over arbitrary ones like `variable1` and `variable2`. Otherwise it's very hard to remember what the contents of different variables are, and it becomes hard to understand what your commands actually do.
- Use short variable names. Typing is a pain and no-one likes doing it. So we much prefer to use a name like `sales` over a name like `sales.for.this.book.that.you.are.reading`. Obviously there's a bit of a tension between using informative names (which tend to be long) and using short names (which tend to be meaningless), so use a bit of common sense when trading off these two conventions.
- Use one of the conventional naming styles for multi-word variable names. Suppose I want to name a variable that stores "my new salary". Obviously I can't include spaces in the variable name, so how should I do this? There are three different conventions that you sometimes see R users employing. Firstly, you can separate the words using periods, which would give you `my.new.salary` as the variable name. Alternatively, you could separate words using underscores, as in `my_new_salary`. Finally, you could use capital letters at the beginning of each word (except the first one), which gives you `myNewSalary` as the variable name. I don't think there's any strong reason to prefer one over the other,¹⁶ but it's important to be consistent.

2.7 Using functions to do calculations

The symbols `+`, `-`, `*` and so on are examples of operators. As we've seen, you can do quite a lot of calculations just by using these operators. However, in order to do more advanced calculations (and later on, to do actual statistics), you're going to need to start using *functions*.¹⁷ I'll talk in more detail about functions and how they work in Section ??, but for now let's just dive in and

¹⁶For very advanced users: there is one exception to this. If you're naming a function, don't use `.` in the name unless you are intending to make use of the S3 object oriented programming system in R. If you don't know what S3 is, then you definitely don't want to be using it! For function naming, there's been a trend among R users to prefer `myFunctionName`.

¹⁷A side note for students with a programming background. Technically speaking, operators *are* functions in R: the addition operator `+` is actually a convenient way of calling the addition function `+()`. Thus `10+20` is equivalent to the function call `+(20, 30)`. Not surprisingly, no-one ever uses this version. Because that would be stupid.

use a few. To get started, suppose I wanted to take the square root of 225. The square root, in case your high school maths is a bit rusty, is just the opposite of squaring a number. So, for instance, since “5 squared is 25” I can say that “5 is the square root of 25”. The usual notation for this is

$$\sqrt{25} = 5$$

though sometimes you’ll also see it written like this $25^{0.5} = 5$. This second way of writing it is kind of useful to “remind” you of the mathematical fact that “square root of x ” is actually the same as “raising x to the power of 0.5”. Personally, I’ve never found this to be terribly meaningful psychologically, though I have to admit it’s quite convenient mathematically. Anyway, it’s not important. What is important is that you remember what a square root is, since we’re going to need it later on.

To calculate the square root of 25, I can do it in my head pretty easily, since I memorised my multiplication tables when I was a kid. It gets harder when the numbers get bigger, and pretty much impossible if they’re not whole numbers. This is where something like R comes in very handy. Let’s say I wanted to calculate $\sqrt{225}$, the square root of 225. There’s two ways I could do this using R. Firstly, since the square root of 255 is the same thing as raising 255 to the power of 0.5, I could use the power operator $^$, just like we did earlier:

```
225 ^ 0.5
```

```
## [1] 15
```

However, there’s a second way that we can do this, since R also provides a ***square root function***, `sqrt()`. To calculate the square root of 255 using this function, what I do is insert the number 255 in the parentheses. That is, the command I type is this:

```
sqrt( 225 )
```

```
## [1] 15
```

and as you might expect from our previous discussion, the spaces in between the parentheses are purely cosmetic. I could have typed `sqrt(225)` or `sqrt(225)` and gotten the same result. When we use a function to do something, we generally refer to this as ***calling*** the function, and the values that we type into the function (there can be more than one) are referred to as the ***arguments*** of that function.

Obviously, the `sqrt()` function doesn’t really give us any new functionality, since we already knew how to do square root calculations by using the power

operator \wedge , though I do think it looks nicer when we use `sqrt()`. However, there are lots of other functions in R: in fact, almost everything of interest that I'll talk about in this book is an R function of some kind. For example, one function that we will need to use in this book is the *absolute value function*. Compared to the square root function, it's extremely simple: it just converts negative numbers to positive numbers, and leaves positive numbers alone. Mathematically, the absolute value of x is written $|x|$ or sometimes $\text{abs}(x)$. Calculating absolute values in R is pretty easy, since R provides the `abs()` function that you can use for this purpose. When you feed it a positive number...

```
abs( 21 )
```

```
## [1] 21
```

the absolute value function does nothing to it at all. But when you feed it a negative number, it spits out the positive version of the same number, like this:

```
abs( -13 )
```

```
## [1] 13
```

In all honesty, there's nothing that the absolute value function does that you couldn't do just by looking at the number and erasing the minus sign if there is one. However, there's a few places later in the book where we have to use absolute values, so I thought it might be a good idea to explain the meaning of the term early on.

Before moving on, it's worth noting that – in the same way that R allows us to put multiple operations together into a longer command, like `1 + 2*4` for instance – it also lets us put functions together and even combine functions with operators if we so desire. For example, the following is a perfectly legitimate command:

```
sqrt( 1 + abs(-8) )
```

```
## [1] 3
```

When R executes this command, starts out by calculating the value of `abs(-8)`, which produces an intermediate value of 8. Having done so, the command simplifies to `sqrt(1 + 8)`. To solve the square root¹⁸ it first needs to add 1 + 8 to get 9, at which point it evaluates `sqrt(9)`, and so it finally outputs a value of 3.

¹⁸A note for the mathematically inclined: R does support complex numbers, but unless you explicitly specify that you want them it assumes all calculations must be real valued. By default, the square root of a negative number is treated as undefined: `sqrt(-9)` will produce `NaN` (not a number) as its output. To get complex numbers, you would type `sqrt(-9+0i)` and R would now return `0+3i`. However, since we won't have any need for complex numbers in this book, I won't refer to them again.

2.7.1 Function arguments, their names and their defaults

There's two more fairly important things that you need to understand about how functions work in R, and that's the use of "named" arguments, and default values" for arguments. Not surprisingly, that's not to say that this is the last we'll hear about how functions work, but they are the last things we desperately need to discuss in order to get you started. To understand what these two concepts are all about, I'll introduce another function. The `round()` function can be used to round some value to the nearest whole number. For example, I could type this:

```
round( 3.1415 )
```

```
## [1] 3
```

Pretty straightforward, really. However, suppose I only wanted to round it to two decimal places: that is, I want to get 3.14 as the output. The `round()` function supports this, by allowing you to input a second argument to the function that specifies the number of decimal places that you want to round the number to. In other words, I could do this:

```
round( 3.14165, 2 )
```

```
## [1] 3.14
```

What's happening here is that I've specified *two* arguments: the first argument is the number that needs to be rounded (i.e., 3.1415), the second argument is the number of decimal places that it should be rounded to (i.e., 2), and the two arguments are separated by a comma. In this simple example, it's quite easy to remember which one argument comes first and which one comes second, but for more complicated functions this is not easy. Fortunately, most R functions make use of **argument names**. For the `round()` function, for example the number that needs to be rounded is specified using the `x` argument, and the number of decimal points that you want it rounded to is specified using the `digits` argument. Because we have these names available to us, we can specify the arguments to the function by name. We do so like this:

```
round( x = 3.1415, digits = 2 )
```

```
## [1] 3.14
```

Notice that this is kind of similar in spirit to variable assignment (Section 2.6), except that I used `=` here, rather than `<-`. In both cases we're specifying specific

values to be associated with a label. However, there are some differences between what I was doing earlier on when creating variables, and what I'm doing here when specifying arguments, and so as a consequence it's important that you use `=` in this context.

As you can see, specifying the arguments by name involves a lot more typing, but it's also a lot easier to read. Because of this, the commands in this book will usually specify arguments by name,¹⁹ since that makes it clearer to you what I'm doing. However, one important thing to note is that when specifying the arguments using their names, it doesn't matter what order you type them in. But if you don't use the argument names, then you have to input the arguments in the correct order. In other words, these three commands all produce the same output...

```
round( 3.14165, 2 )
## [1] 3.14

round( x = 3.1415, digits = 2 )
## [1] 3.14

round( digits = 2, x = 3.1415 )
## [1] 3.14
```

but this one does not...

```
round( 2, 3.14165 )
## [1] 2
```

How do you find out what the correct order is? There's a few different ways, but the easiest one is to look at the help documentation for the function (see Section 2.27). However, if you're ever unsure, it's probably best to actually type in the argument name.

¹⁹The two functions discussed previously, `sqrt()` and `abs()`, both only have a single argument, `x`. So I could have typed something like `sqrt(x = 225)` or `abs(x = -13)` earlier. The fact that all these functions use `x` as the name of the argument that corresponds the “main” variable that you’re working with is no coincidence. That’s a fairly widely used convention. Quite often, the writers of R functions will try to use conventional names like this to make your life easier. Or at least that’s the theory. In practice it doesn’t always work as well as you’d hope.

Okay, so that's the first thing I said you'd need to know: argument names. The second thing you need to know about is default values. Notice that the first time I called the `round()` function I didn't actually specify the `digits` argument at all, and yet R somehow knew that this meant it should round to the nearest whole number. How did that happen? The answer is that the `digits` argument has a *default value* of 0, meaning that if you decide not to specify a value for `digits` then R will act as if you had typed `digits = 0`. This is quite handy: the vast majority of the time when you want to round a number you want to round it to the nearest whole number, and it would be pretty annoying to have to specify the `digits` argument every single time. On the other hand, sometimes you actually do want to round to something other than the nearest whole number, and it would be even more annoying if R didn't allow this! Thus, by having `digits = 0` as the default value, we get the best of both worlds.

2.8 Letting RStudio help you with your commands

Time for a bit of a digression. At this stage you know how to type in basic commands, including how to use R functions. And it's probably beginning to dawn on you that there are a *lot* of R functions, all of which have their own arguments. You're probably also worried that you're going to have to remember all of them! Thankfully, it's not that bad. In fact, very few data analysts bother to try to remember all the commands. What they really do is use tricks to make their lives easier. The first (and arguably most important one) is to use the internet. If you don't know how a particular R function works, Google it. Second, you can look up the R help documentation. I'll talk more about these two tricks in Section 2.27. But right now I want to call your attention to a couple of simple tricks that RStudio makes available to you.

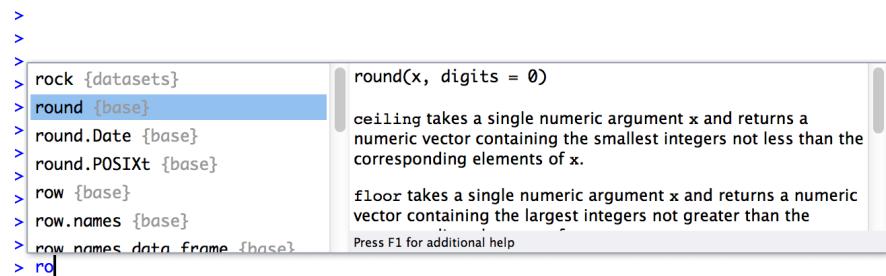
2.8.1 Autocomplete using “tab”

The first thing I want to call your attention to is the *autocomplete* ability in RStudio.²⁰

Let's stick to our example above and assume that what you want to do is to round a number. This time around, start typing the name of the function that you want, and then hit the “tab” key. RStudio will then display a little window like the one shown in Figure 2.2. In this figure, I've typed the letters `ro` at the command line, and then hit tab. The window has two panels. On the

²⁰For advanced users: obviously, this isn't just an RStudio thing. If you're running R in a terminal window, tab autocomplete still works, and does so in exactly the way you'd expect. It's not as visually pretty as the RStudio version, of course, and lacks some of the cooler features that RStudio provides. I don't bother to document that here: my assumption is that if you are running R in the terminal then you're already familiar with using tab autocomplete.

left, there's a list of variables and functions that start with the letters that I've typed shown in black text, and some grey text that tells you where that variable/function is stored. Ignore the grey text for now: it won't make much sense to you until we've talked about packages in Section 2.17. In Figure 2.2 you can see that there's quite a few things that start with the letters `ro`: there's something called `rock`, something called `round`, something called `round.Date` and so on. The one we want is `round`, but if you're typing this yourself you'll notice that when you hit the tab key the window pops up with the top entry (i.e., `rock`) highlighted. You can use the up and down arrow keys to select the one that you want. Or, if none of the options look right to you, you can hit the escape key ("esc") or the left arrow key to make the window go away.



The screenshot shows an RStudio console window. A user has typed the letters "ro" and pressed the tab key. A dropdown menu appears, listing several items:

- `rock {datasets}`
- `round {base}` (highlighted in blue)
- `round.Date {base}`
- `round.POSIXt {base}`
- `row {base}`
- `row.names {base}`
- `row.names.data.frame {base}`
- `ro`

To the right of the dropdown, a tooltip provides information about the `round` function:

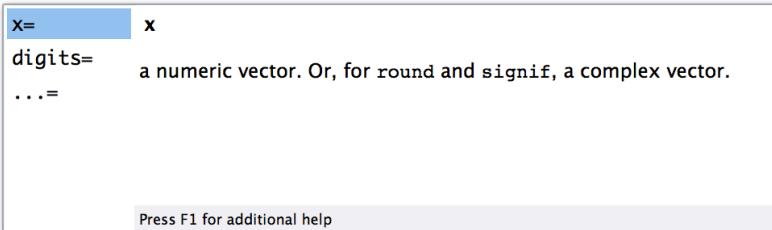
`round(x, digits = 0)`
`ceiling` takes a single numeric argument `x` and returns a numeric vector containing the smallest integers not less than the corresponding elements of `x`.
`floor` takes a single numeric argument `x` and returns a numeric vector containing the largest integers not greater than the
Press F1 for additional help

Figure 2.2: Start typing the name of a function or a variable, and hit the “tab” key. RStudio brings up a little dialog box like this one that lets you select the one you want, and even prints out a little information about it.

In our case, the thing we want is the `round` option, so we'll select that. When you do this, you'll see that the panel on the right changes. Previously, it had been telling us something about the `rock` data set (i.e., “Measurements on 48 rock samples...”) that is distributed as part of R. But when we select `round`, it displays information about the `round()` function, exactly as it is shown in Figure 2.2. This display is really handy. The very first thing it says is `round(x, digits = 0)`: what this is telling you is that the `round()` function has two arguments. The first argument is called `x`, and it doesn't have a default value. The second argument is `digits`, and it has a default value of 0. In a lot of situations, that's all the information you need. But RStudio goes a bit further, and provides some additional information about the function underneath. Sometimes that additional information is very helpful, sometimes it's not: RStudio pulls that text from the R help documentation, and my experience is that the helpfulness of that documentation varies wildly. Anyway, if you've decided that `round()` is the function that you want to use, you can hit the right arrow or the enter key, and RStudio will finish typing the rest of the function name for you.

The RStudio autocomplete tool works slightly differently if you've already got the name of the function typed and you're now trying to type the arguments. For instance, suppose I've typed `round(` into the console, and *then* I hit tab. RStudio is smart enough to recognise that I already know the name of the

function that I want, because I've already typed it! Instead, it figures that what I'm interested in is the *arguments* to that function. So that's what pops up in the little window. You can see this in Figure 2.3. Again, the window has two panels, and you can interact with this window in exactly the same way that you did with the window shown in Figure 2.2. On the left hand panel, you can see a list of the argument names. On the right hand side, it displays some information about what the selected argument does.



The screenshot shows the RStudio Argument Completion Window. A tooltip-like window is open over the console area. The window has a title bar with the text 'x=' and 'x'. Inside, there is a list of arguments: 'digits=' followed by the description 'a numeric vector. Or, for round and signif, a complex vector.' and '...=' below it. At the bottom of the window, there is a message 'Press F1 for additional help'.

```
>
>
> x= x
> digits= a numeric vector. Or, for round and signif, a complex vector.
> ...=
>
>
>
>
>
> round(
```

Figure 2.3: If you've typed the name of a function already along with the left parenthesis and then hit the “tab” key, RStudio brings up a different window to the one shown above. This one lists all the arguments to the function on the left, and information about each argument on the right.

2.8.2 Browsing your command history

One thing that R does automatically is keep track of your “command history”. That is, it remembers all the commands that you've previously typed. You can access this history in a few different ways. The simplest way is to use the up and down arrow keys. If you hit the up key, the R console will show you the most recent command that you've typed. Hit it again, and it will show you the command before that. If you want the text on the screen to go away, hit escape²¹ Using the up and down keys can be really handy if you've typed a long command that had one typo in it. Rather than having to type it all again from scratch, you can use the up key to bring up the command and fix it.

The second way to get access to your command history is to look at the history panel in RStudio. On the upper right hand side of the RStudio window you'll see a tab labelled “History”. Click on that, and you'll see a list of all your recent commands displayed in that panel: it should look something like Figure 2.4. If you double click on one of the commands, it will be copied to the R console. (You can achieve the same result by selecting the command you want with the mouse and then clicking the “To Console” button).²²

²¹Incidentally, that always works: if you've started typing a command and you want to clear it and start again, hit escape.

²²Another method is to start typing some text and then hit the Control key and the up

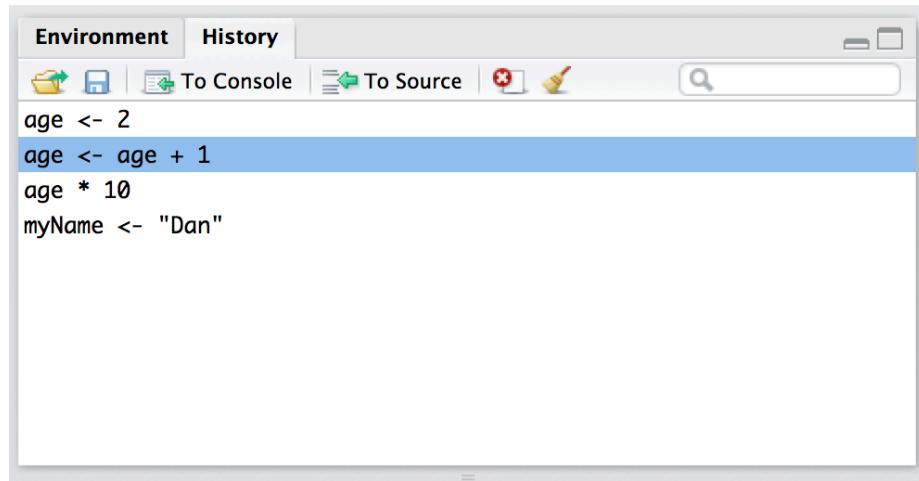


Figure 2.4: The history panel is located in the top right hand side of the RStudio window. Click on the word “History” and it displays this panel.

2.9 Storing many numbers as a vector

At this point we’ve covered functions in enough detail to get us safely through the next couple of chapters (with one small exception: see Section 2.26, so let’s return to our discussion of variables. When I introduced variables in Section 2.6 I showed you how we can use variables to store a single number. In this section, we’ll extend this idea and look at how to store multiple numbers within the one variable. In R, the name for a variable that can store multiple values is a *vector*. So let’s create one.

2.9.1 Creating a vector

Let’s stick to my silly “get rich quick by textbook writing” example. Suppose the textbook company (if I actually had one, that is) sends me sales data on a monthly basis. Since my class start in late February, we might expect most of the sales to occur towards the start of the year. Let’s suppose that I have 100 sales in February, 200 sales in March and 50 sales in April, and no other sales for the rest of the year. What I would like to do is have a variable – let’s call it `sales.by.month` – that stores all this sales data. The first number stored should be 0 since I had no sales in January, the second should be 100, and so on. The simplest way to do this in R is to use the *combine* function, `c()`. To

arrow together (on Windows or Linux) or the Command key and the up arrow together (on a Mac). This will bring up a window showing all your recent commands that started with the same text as what you’ve currently typed. That can come in quite handy sometimes.

do so, all we have to do is type all the numbers you want to store in a comma separated list, like this:²³

```
sales.by.month <- c(0, 100, 200, 50, 0, 0, 0, 0, 0, 0, 0, 0)
sales.by.month
```

```
## [1] 0 100 200 50 0 0 0 0 0 0 0 0
```

To use the correct terminology here, we have a single variable here called `sales.by.month`: this variable is a vector that consists of 12 *elements*.

2.9.2 A handy digression

Now that we've learned how to put information into a vector, the next thing to understand is how to pull that information back out again. However, before I do so it's worth taking a slight detour. If you've been following along, typing all the commands into R yourself, it's possible that the output that you saw when we printed out the `sales.by.month` vector was slightly different to what I showed above. This would have happened if the window (or the RStudio panel) that contains the R console is really, really narrow. If that were the case, you might have seen output that looks something like this:

```
sales.by.month
```

```
## [1] 0 100 200 50
## [5] 0 0 0 0
## [9] 0 0 0 0
```

Because there wasn't much room on the screen, R has printed out the results over three lines. But that's not the important thing to notice. The important point is that the first line has a `[1]` in front of it, whereas the second line starts with `[5]` and the third with `[9]`. It's pretty clear what's happening here. For the first row, R has printed out the 1st element through to the 4th element, so it starts that row with a `[1]`. For the second row, R has printed out the 5th element of the vector through to the 8th one, and so it begins that row with a `[5]` so that you can tell where it's up to at a glance. It might seem a bit odd to you that R does this, but in some ways it's a kindness, especially when dealing with larger data sets!

²³Notice that I didn't specify any argument names here. The `c()` function is one of those cases where we don't use names. We just type all the numbers, and R just dumps them all in a single variable.

2.9.3 Getting information out of vectors

To get back to the main story, let's consider the problem of how to get information out of a vector. At this point, you might have a sneaking suspicion that the answer has something to do with the [1] and [9] things that R has been printing out. And of course you are correct. Suppose I want to pull out the February sales data only. February is the second month of the year, so let's try this:

```
sales.by.month[2]
```

```
## [1] 100
```

Yep, that's the February sales all right. But there's a subtle detail to be aware of here: notice that R outputs [1] 100, *not* [2] 100. This is because R is being extremely literal. When we typed in `sales.by.month[2]`, we asked R to find exactly *one* thing, and that one thing happens to be the second element of our `sales.by.month` vector. So, when it outputs [1] 100 what R is saying is that the first number *that we just asked for* is 100. This behaviour makes more sense when you realise that we can use this trick to create new variables. For example, I could create a `february.sales` variable like this:

```
february.sales <- sales.by.month[2]
february.sales
```

```
## [1] 100
```

Obviously, the new variable `february.sales` should only have one element and so when I print it out this new variable, the R output begins with a [1] because 100 is the value of the first (and only) element of `february.sales`. The fact that this also happens to be the value of the second element of `sales.by.month` is irrelevant. We'll pick this topic up again shortly (Section 2.12).

2.9.4 Altering the elements of a vector

Sometimes you'll want to change the values stored in a vector. Imagine my surprise when the publisher rings me up to tell me that the sales data for May are wrong. There were actually an additional 25 books sold in May, but there was an error or something so they hadn't told me about it. How can I fix my `sales.by.month` variable? One possibility would be to assign the whole vector again from the beginning, using `c()`. But that's a lot of typing. Also, it's a little wasteful: why should R have to redefine the sales figures for all 12 months, when only the 5th one is wrong? Fortunately, we can tell R to change only the 5th element, using this trick:

```

sales.by.month[5] <- 25
sales.by.month

## [1] 0 100 200 50 25 0 0 0 0 0 0

```

Another way to edit variables is to use the `edit()` or `fix()` functions. I won't discuss them in detail right now, but you can check them out on your own.

2.9.5 Useful things to know about vectors

Before moving on, I want to mention a couple of other things about vectors. Firstly, you often find yourself wanting to know how many elements there are in a vector (usually because you've forgotten). You can use the `length()` function to do this. It's quite straightforward:

```

length( x = sales.by.month )

## [1] 12

```

Secondly, you often want to alter all of the elements of a vector at once. For instance, suppose I wanted to figure out how much money I made in each month. Since I'm earning an exciting \$7 per book (no seriously, that's actually pretty close to what authors get on the very expensive textbooks that you're expected to purchase), what I want to do is multiply each element in the `sales.by.month` vector by 7. R makes this pretty easy, as the following example shows:

```

sales.by.month * 7

## [1] 0 700 1400 350 175 0 0 0 0 0 0

```

In other words, when you multiply a vector by a single number, all elements in the vector get multiplied. The same is true for addition, subtraction, division and taking powers. So that's neat. On the other hand, suppose I wanted to know how much money I was making per day, rather than per month. Since not every month has the same number of days, I need to do something slightly different. Firstly, I'll create two new vectors:

```

days.per.month <- c(31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31)
profit <- sales.by.month * 7

```

Obviously, the `profit` variable is the same one we created earlier, and the `days.per.month` variable is pretty straightforward. What I want to do is divide every element of `profit` by the *corresponding* element of `days.per.month`. Again, R makes this pretty easy:

```
profit / days.per.month
```

```
## [1] 0.000000 25.000000 45.161290 11.666667 5.645161 0.000000 0.000000
## [8] 0.000000 0.000000 0.000000 0.000000 0.000000
```

I still don't like all those zeros, but that's not what matters here. Notice that the second element of the output is 25, because R has divided the second element of `profit` (i.e. 700) by the second element of `days.per.month` (i.e. 28). Similarly, the third element of the output is equal to 1400 divided by 31, and so on. We'll talk more about calculations involving vectors later on (and in particular a thing called the "recycling rule"; Section ??), but that's enough detail for now.

2.10 Storing text data

A lot of the time your data will be numeric in nature, but not always. Sometimes your data really needs to be described using text, not using numbers. To address this, we need to consider the situation where our variables store text. To create a variable that stores the word "hello", we can type this:

```
greeting <- "hello"
greeting
```

```
## [1] "hello"
```

When interpreting this, it's important to recognise that the quote marks here *aren't* part of the string itself. They're just something that we use to make sure that R knows to treat the characters that they enclose as a piece of text data, known as a *character string*. In other words, R treats "`hello`" as a string containing the word "hello"; but if I had typed `hello` instead, R would go looking for a variable by that name! You can also use '`hello`' to specify a character string.

Okay, so that's how we store the text. Next, it's important to recognise that when we do this, R stores the entire word "`hello`" as a *single* element: our `greeting` variable is *not* a vector of five different letters. Rather, it has only the one element, and that element corresponds to the entire character string "`hello`". To illustrate this, if I actually ask R to find the first element of `greeting`, it prints the whole string:

```
greeting[1]
```

```
## [1] "hello"
```

Of course, there's no reason why I can't create a vector of character strings. For instance, if we were to continue with the example of my attempts to look at the monthly sales data for my book, one variable I might want would include the names of all 12 months.²⁴ To do so, I could type in a command like this

```
months <- c("January", "February", "March", "April", "May", "June",
          "July", "August", "September", "October", "November",
          "December")
```

This is a ***character vector*** containing 12 elements, each of which is the name of a month. So if I wanted R to tell me the name of the fourth month, all I would do is this:

```
months[4]
```

```
## [1] "April"
```

2.10.1 Working with text

Working with text data is somewhat more complicated than working with numeric data, and I discuss some of the basic ideas in Section ??, but for purposes of the current chapter we only need this bare bones sketch. The only other thing I want to do before moving on is show you an example of a function that can be applied to text data. So far, most of the functions that we have seen (i.e., `sqrt()`, `abs()` and `round()`) only make sense when applied to numeric data (e.g., you can't calculate the square root of "hello"), and we've seen one function that can be applied to pretty much any variable or vector (i.e., `length()`). So it might be nice to see an example of a function that can be applied to text.

The function I'm going to introduce you to is called `nchar()`, and what it does is count the number of individual characters that make up a string. Recall earlier that when we tried to calculate the `length()` of our `greeting` variable it returned a value of 1: the `greeting` variable contains only the one string, which happens to be "hello". But what if I want to know how many letters there are in the word? Sure, I could *count* them, but that's boring, and more to the point it's a terrible strategy if what I wanted to know was the number of letters in *War and Peace*. That's where the `nchar()` function is helpful:

²⁴Though actually there's no real need to do this, since R has an inbuilt variable called `month.name` that you can use for this purpose.

```
nchar( x = greeting )
```

```
## [1] 5
```

That makes sense, since there are in fact 5 letters in the string "hello". Better yet, you can apply `nchar()` to whole vectors. So, for instance, if I want R to tell me how many letters there are in the names of each of the 12 months, I can do this:

```
nchar( x = months )
```

```
## [1] 7 8 5 5 3 4 4 6 9 7 8 8
```

So that's nice to know. The `nchar()` function can do a bit more than this, and there's a lot of other functions that you can do to extract more information from text or do all sorts of fancy things. However, the goal here is not to teach any of that! The goal right now is just to see an example of a function that actually does work when applied to text.

2.11 Storing “true or false” data

Time to move onto a third kind of data. A key concept in that a lot of R relies on is the idea of a *logical value*. A logical value is an assertion about whether something is true or false. This is implemented in R in a pretty straightforward way. There are two logical values, namely `TRUE` and `FALSE`. Despite the simplicity, a logical values are very useful things. Let's see how they work.

2.11.1 Assessing mathematical truths

In George Orwell's classic book *1984*, one of the slogans used by the totalitarian Party was “two plus two equals five”, the idea being that the political domination of human freedom becomes complete when it is possible to subvert even the most basic of truths. It's a terrifying thought, especially when the protagonist Winston Smith finally breaks down under torture and agrees to the proposition. “Man is infinitely malleable”, the book says. I'm pretty sure that this isn't true of humans²⁵ but it's definitely not true of R. R is not infinitely malleable. It has rather firm opinions on the topic of what is and isn't true, at least as regards basic mathematics. If I ask it to calculate $2 + 2$, it always gives the same answer, and it's not bloody 5:

²⁵I offer up my teenage attempts to be “cool” as evidence that some things just can't be done.

```
2 + 2
```

```
## [1] 4
```

Of course, so far R is just doing the calculations. I haven’t asked it to explicitly assert that $2 + 2 = 4$ is a true statement. If I want R to make an explicit judgement, I can use a command like this:

```
2 + 2 == 4
```

```
## [1] TRUE
```

What I’ve done here is use the *equality operator*, `==`, to force R to make a “true or false” judgement.²⁶ Okay, let’s see what R thinks of the Party slogan:

```
2+2 == 5
```

```
## [1] FALSE
```

Booyah! Freedom and ponies for all! Or something like that. Anyway, it’s worth having a look at what happens if I try to *force* R to believe that two plus two is five by making an assignment statement like $2 + 2 = 5$ or $2 + 2 <- 5$. When I do this, here’s what happens:

```
2 + 2 = 5
```

```
## Error in 2 + 2 = 5: target of assignment expands to non-language object
```

R doesn’t like this very much. It recognises that $2 + 2$ is *not* a variable (that’s what the “non-language object” part is saying), and it won’t let you try to “reassign” it. While R is pretty flexible, and actually does let you do some quite remarkable things to redefine parts of R itself, there are just some basic, primitive truths that it refuses to give up. It won’t change the laws of addition, and it won’t change the definition of the number 2.

That’s probably for the best.

²⁶Note that this is a very different operator to the assignment operator `=` that I talked about in Section 2.6. A common typo that people make when trying to write logical commands in R (or other languages, since the “`=` versus `==`” distinction is important in most programming languages) is to accidentally type `=` when you really mean `==`. Be especially cautious with this – I’ve been programming in various languages since I was a teenager, and I *still* screw this up a lot. Hm. I think I see why I wasn’t cool as a teenager. And why I’m still not cool.

Table 2.2: Some logical operators. Technically I should be calling these "binary relational operators", but quite frankly I don't want to. It's my book so no-one can make me.

operation	operator	example input	answer
less than	<	$2 < 3$	'TRUE'
less than or equal to	\leq	$2 \leq 2$	'TRUE'
greater than	>	$2 > 3$	'FALSE'
greater than or equal to	\geq	$2 \geq 2$	'TRUE'
equal to	\equiv	$2 \equiv 3$	'FALSE'
not equal to	\neq	$2 \neq 3$	'TRUE'

2.11.2 Logical operations

So now we've seen logical operations at work, but so far we've only seen the simplest possible example. You probably won't be surprised to discover that we can combine logical operations with other operations and functions in a more complicated way, like this:

```
3*3 + 4*4 == 5*5
```

```
## [1] TRUE
```

or this

```
sqrt( 25 ) == 5
```

```
## [1] TRUE
```

Not only that, but as Table 2.2 illustrates, there are several other logical operators that you can use, corresponding to some basic mathematical concepts.

Hopefully these are all pretty self-explanatory: for example, the *less than* operator $<$ checks to see if the number on the left is less than the number on the right. If it's less, then R returns an answer of TRUE:

```
99 < 100
```

```
## [1] TRUE
```

but if the two numbers are equal, or if the one on the right is larger, then R returns an answer of FALSE, as the following two examples illustrate:

```
100 < 100
## [1] FALSE

100 < 99
## [1] FALSE
```

In contrast, the *less than or equal to* operator `<=` will do exactly what it says. It returns a value of TRUE if the number of the left hand side is less than or equal to the number on the right hand side. So if we repeat the previous two examples using `<=`, here’s what we get:

```
100 <= 100
## [1] TRUE

100 <= 99
## [1] FALSE
```

And at this point I hope it’s pretty obvious what the *greater than* operator `>` and the *greater than or equal to* operator `>=` do! Next on the list of logical operators is the *not equal to* operator `!=` which – as with all the others – does what it says it does. It returns a value of TRUE when things on either side are not identical to each other. Therefore, since $2 + 2$ isn’t equal to 5, we get:

```
2 + 2 != 5
## [1] TRUE
```

We’re not quite done yet. There are three more logical operations that are worth knowing about, listed in Table 2.3.

These are the *not* operator `!`, the *and* operator `&`, and the *or* operator `|`. Like the other logical operators, their behaviour is more or less exactly what you’d expect given their names. For instance, if I ask you to assess the claim that “either $2 + 2 = 4$ or $2 + 2 = 5$ ” you’d say that it’s true. Since it’s an “either-or” statement, all we need is for one of the two parts to be true. That’s what the `|` operator does:

Table 2.3: Some more logical operators.

operation	operator	example input	answer
not	!	<code>!(1==1)</code>	‘FALSE’
or		<code>(1==1) (2==3)</code>	‘TRUE’
and	&	<code>(1==1) & (2==3)</code>	‘FALSE’

```
(2+2 == 4) | (2+2 == 5)
```

```
## [1] TRUE
```

On the other hand, if I ask you to assess the claim that “both $2 + 2 = 4$ and $2 + 2 = 5$ ” you’d say that it’s false. Since this is an *and* statement we need both parts to be true. And that’s what the `&` operator does:

```
(2+2 == 4) & (2+2 == 5)
```

```
## [1] FALSE
```

Finally, there’s the *not* operator, which is simple but annoying to describe in English. If I ask you to assess my claim that “it is not true that $2 + 2 = 5$ ” then you would say that my claim is true; because my claim is that “ $2 + 2 = 5$ is false”. And I’m right. If we write this as an R command we get this:

```
! (2+2 == 5)
```

```
## [1] TRUE
```

In other words, since `2+2 == 5` is a FALSE statement, it must be the case that `!(2+2 == 5)` is a TRUE one. Essentially, what we’ve really done is claim that “not false” is the same thing as “true”. Obviously, this isn’t really quite right in real life. But R lives in a much more black or white world: for R everything is either true or false. No shades of gray are allowed. We can actually see this much more explicitly, like this:

```
! FALSE
```

```
## [1] TRUE
```

Of course, in our $2 + 2 = 5$ example, we didn’t really need to use “not” `!` and “equals to” `==` as two separate operators. We could have just used the “not equals to” operator `!=` like this:

```
2+2 != 5
```

```
## [1] TRUE
```

But there are many situations where you really do need to use the `!` operator. We’ll see some later on.²⁷

2.11.3 Storing and using logical data

Up to this point, I’ve introduced *numeric data* (in Sections 2.6 and 2.9) and *character data* (in Section 2.10). So you might not be surprised to discover that these `TRUE` and `FALSE` values that R has been producing are actually a third kind of data, called *logical data*. That is, when I asked R if $2 + 2 == 5$ and it said `[1] FALSE` in reply, it was actually producing information that we can store in variables. For instance, I could create a variable called `is.the.Party.correct`, which would store R’s opinion:

```
is.the.Party.correct <- 2 + 2 == 5
is.the.Party.correct
```

```
## [1] FALSE
```

Alternatively, you can assign the value directly, by typing `TRUE` or `FALSE` in your command. Like this:

```
is.the.Party.correct <- FALSE
is.the.Party.correct
```

```
## [1] FALSE
```

²⁷A note for those of you who have taken a computer science class: yes, R does have a function for exclusive-or, namely `xor()`. Also worth noting is the fact that R makes the distinction between element-wise operators `&` and `|` and operators that look only at the first element of the vector, namely `&&` and `||`. To see the distinction, compare the behaviour of a command like `c(FALSE, TRUE) & c(TRUE, TRUE)` to the behaviour of something like `c(FALSE, TRUE) && c(TRUE, TRUE)`. If this doesn’t mean anything to you, ignore this footnote entirely. It’s not important for the content of this book.

Better yet, because it's kind of tedious to type TRUE or FALSE over and over again, R provides you with a shortcut: you can use T and F instead (but it's case sensitive: t and f won't work).²⁸ So this works:

```
is.the.Party.correct <- F
is.the.Party.correct
```

```
## [1] FALSE
```

but this doesn't:

```
is.the.Party.correct <- f
```

```
## Error in eval(expr, envir, enclos): object 'f' not found
```

2.11.4 Vectors of logicals

The next thing to mention is that you can store vectors of logical values in exactly the same way that you can store vectors of numbers (Section 2.9) and vectors of text data (Section 2.10). Again, we can define them directly via the `c()` function, like this:

```
x <- c(TRUE, TRUE, FALSE)
x
```

```
## [1] TRUE TRUE FALSE
```

or you can produce a vector of logicals by applying a logical operator to a vector. This might not make a lot of sense to you, so let's unpack it slowly. First, let's suppose we have a vector of numbers (i.e., a "non-logical vector"). For instance, we could use the `sales.by.month` vector that we were using in Section 2.9. Suppose I wanted R to tell me, for each month of the year, whether I actually sold a book in that month. I can do that by typing this:

```
sales.by.month > 0
```

```
## [1] FALSE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE
```

²⁸Warning! `TRUE` and `FALSE` are reserved keywords in R, so you can trust that they always mean what they say they do. Unfortunately, the shortcut versions `T` and `F` do not have this property. It's even possible to create variables that set up the reverse meanings, by typing commands like `T <- FALSE` and `F <- TRUE`. This is kind of insane, and something that is generally thought to be a design flaw in R. Anyway, the long and short of it is that it's safer to use `TRUE` and `FALSE`.

and again, I can store this in a vector if I want, as the example below illustrates:

```
any.sales.this.month <- sales.by.month > 0
any.sales.this.month

## [1] FALSE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE
```

In other words, `any.sales.this.month` is a logical vector whose elements are TRUE only if the corresponding element of `sales.by.month` is greater than zero. For instance, since I sold zero books in January, the first element is FALSE.

2.11.5 Applying logical operation to text

In a moment (Section 2.12) I'll show you why these logical operations and logical vectors are so handy, but before I do so I want to very briefly point out that you can apply them to text as well as to logical data. It's just that we need to be a bit more careful in understanding how R interprets the different operations. In this section I'll talk about how the equal to operator `==` applies to text, since this is the most important one. Obviously, the not equal to operator `!=` gives the exact opposite answers to `==` so I'm implicitly talking about that one too, but I won't give specific commands showing the use of `!=`. As for the other operators, I'll defer a more detailed discussion of this topic to Section `??`.

Okay, let's see how it works. In one sense, it's very simple. For instance, I can ask R if the word "cat" is the same as the word "dog", like this:

```
"cat" == "dog"
```

```
## [1] FALSE
```

That's pretty obvious, and it's good to know that even R can figure that out. Similarly, R does recognise that a "cat" is a "cat":

```
"cat" == "cat"
```

```
## [1] TRUE
```

Again, that's exactly what we'd expect. However, what you need to keep in mind is that R is not at all tolerant when it comes to grammar and spacing. If two strings differ in any way whatsoever, R will say that they're not equal to each other, as the following examples indicate:

```
" cat" == "cat"
## [1] FALSE

"cat" == "CAT"
## [1] FALSE

"cat" == "c a t"
## [1] FALSE
```

2.12 Indexing vectors

One last thing to add before finishing up this chapter. So far, whenever I've had to get information out of a vector, all I've done is typed something like `months[4]`; and when I do this R prints out the fourth element of the `months` vector. In this section, I'll show you two additional tricks for getting information out of the vector.

2.12.1 Extracting multiple elements

One very useful thing we can do is pull out more than one element at a time. In the previous example, we only used a single number (i.e., 2) to indicate which element we wanted. Alternatively, we can use a vector. So, suppose I wanted the data for February, March and April. What I could do is use the vector `c(2,3,4)` to indicate which elements I want R to pull out. That is, I'd type this:

```
sales.by.month[ c(2,3,4) ]
```

```
## [1] 100 200 50
```

Notice that the order matters here. If I asked for the data in the reverse order (i.e., April first, then March, then February) by using the vector `c(4,3,2)`, then R outputs the data in the reverse order:

```
sales.by.month[ c(4,3,2) ]
```

```
## [1] 50 200 100
```

A second thing to be aware of is that R provides you with handy shortcuts for very common situations. For instance, suppose that I wanted to extract everything from the 2nd month through to the 8th month. One way to do this is to do the same thing I did above, and use the vector `c(2,3,4,5,6,7,8)` to indicate the elements that I want. That works just fine

```
sales.by.month[ c(2,3,4,5,6,7,8) ]
```

```
## [1] 100 200 50 25 0 0 0
```

but it's kind of a lot of typing. To help make this easier, R lets you use `2:8` as shorthand for `c(2,3,4,5,6,7,8)`, which makes things a lot simpler. First, let's just check that this is true:

```
2:8
```

```
## [1] 2 3 4 5 6 7 8
```

Next, let's check that we can use the `2:8` shorthand as a way to pull out the 2nd through 8th elements of `sales.by.months`:

```
sales.by.month[2:8]
```

```
## [1] 100 200 50 25 0 0 0
```

So that's kind of neat.

2.12.2 Logical indexing

At this point, I can introduce an extremely useful tool called *logical indexing*. In the last section, I created a logical vector `any.sales.this.month`, whose elements are TRUE for any month in which I sold at least one book, and FALSE for all the others. However, that big long list of TRUES and FALSEs is a little bit hard to read, so what I'd like to do is to have R select the names of the months for which I sold any books. Earlier on, I created a vector `months` that contains the names of each of the months. This is where logical indexing is handy. What I need to do is this:

```
months[ sales.by.month > 0 ]
```

```
## [1] "February" "March"    "April"     "May"
```

To understand what's happening here, it's helpful to notice that `sales.by.month > 0` is the same logical expression that we used to create the `any.sales.this.month` vector in the last section. In fact, I could have just done this:

```
months[ any.sales.this.month ]
## [1] "February" "March"      "April"       "May"
```

and gotten exactly the same result. In order to figure out which elements of `months` to include in the output, what R does is look to see if the corresponding element in `any.sales.this.month` is TRUE. Thus, since element 1 of `any.sales.this.month` is FALSE, R does not include "January" as part of the output; but since element 2 of `any.sales.this.month` is TRUE, R does include "February" in the output. Note that there's no reason why I can't use the same trick to find the actual sales numbers for those months. The command to do that would just be this:

```
sales.by.month [ sales.by.month > 0 ]
## [1] 100 200 50 25
```

In fact, we can do the same thing with text. Here's an example. Suppose that – to continue the saga of the textbook sales – I later find out that the bookshop only had sufficient stocks for a few months of the year. They tell me that early in the year they had "high" stocks, which then dropped to "low" levels, and in fact for one month they were "out" of copies of the book for a while before they were able to replenish them. Thus I might have a variable called `stock.levels` which looks like this:

```
stock.levels<-c("high", "high", "low", "out", "out", "high",
                 "high", "high", "high", "high", "high")
stock.levels
## [1] "high" "high" "low"   "out"   "out"   "high" "high" "high" "high" "high"
## [11] "high" "high"
```

Thus, if I want to know the months for which the bookshop was out of my book, I could apply the logical indexing trick, but with the character vector `stock.levels`, like this:

```
months[stock.levels == "out"]
```

```
## [1] "April" "May"
```

Alternatively, if I want to know when the bookshop was either low on copies or out of copies, I could do this:

```
months[stock.levels == "out" | stock.levels == "low"]
```

```
## [1] "March" "April" "May"
```

or this

```
months[stock.levels != "high" ]
```

```
## [1] "March" "April" "May"
```

Either way, I get the answer I want.

At this point, I hope you can see why logical indexing is such a useful thing. It's a very basic, yet very powerful way to manipulate data. We'll talk a lot more about how to manipulate data in Chapter ??, since it's a critical skill for real world research that is often overlooked in introductory research methods classes (or at least, that's been my experience). It does take a bit of practice to become completely comfortable using logical indexing, so it's a good idea to play around with these sorts of commands. Try creating a few different variables of your own, and then ask yourself questions like "how do I get R to spit out all the elements that are [blah]". Practice makes perfect, and it's only by practicing logical indexing that you'll perfect the art of yelling frustrated insults at your computer.²⁹

2.13 Quitting R

There's one last thing I should cover in this chapter: how to quit R. When I say this, I'm not trying to imply that R is some kind of pathological addition and that you need to call the R QuitLine or wear patches to control the cravings (although you certainly might argue that there's something seriously pathological about being addicted to R). I just mean how to exit the program. Assuming you're running R in the usual way (i.e., through RStudio or the default GUI on

²⁹Well, I say that... but in my personal experience it wasn't until I started learning "regular expressions" that my loathing of computers reached its peak.

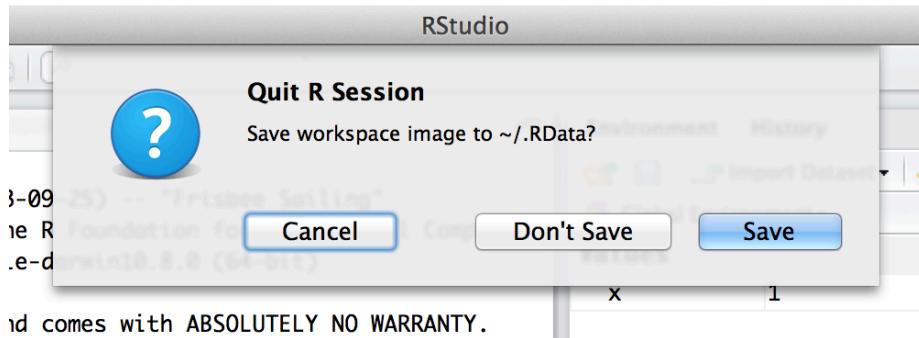


Figure 2.5: The dialog box that shows up when you try to close RStudio.

a Windows or Mac computer), then you can just shut down the application in the normal way. However, R also has a function, called `q()` that you can use to quit, which is pretty handy if you’re running R in a terminal window.

Regardless of what method you use to quit R, when you do so for the first time R will probably ask you if you want to save the “workspace image”. We’ll talk a lot more about loading and saving data in Section 2.20, but I figured we’d better quickly cover this now otherwise you’re going to get annoyed when you close R at the end of the chapter. If you’re using RStudio, you’ll see a dialog box that looks like the one shown in Figure 2.5. If you’re using a text based interface you’ll see this:

```
q()
## Save workspace image? [y/n/c]:
```

The `y/n/c` part here is short for “yes / no / cancel”. Type `y` if you want to save, `n` if you don’t, and `c` if you’ve changed your mind and you don’t want to quit after all.

What does this actually *mean*? What’s going on is that R wants to know if you want to save all those variables that you’ve been creating, so that you can use them later. This sounds like a great idea, so it’s really tempting to type `y` or click the “Save” button. To be honest though, I very rarely do this, and it kind of annoys me a little bit... what R is *really* asking is if you want it to store these variables in a “default” data file, which it will automatically reload for you next time you open R. And quite frankly, if I’d wanted to save the variables, then I’d have already saved them before trying to quit. Not only that, I’d have saved them to a location of *my* choice, so that I can find it again later. So I personally never bother with this.

In fact, every time I install R on a new machine one of the first things I do is change the settings so that it never asks me again. You can do this in RStudio

really easily: use the menu system to find the RStudio option; the dialog box that comes up will give you an option to tell R never to whine about this again (see Figure 2.6. On a Mac, you can open this window by going to the “RStudio” menu and selecting “Preferences”. On a Windows machine you go to the “Tools” menu and select “Global Options”. Under the “General” tab you’ll see an option that reads “Save workspace to .Rdata on exit”. By default this is set to “ask”. If you want R to stop asking, change it to “never”.

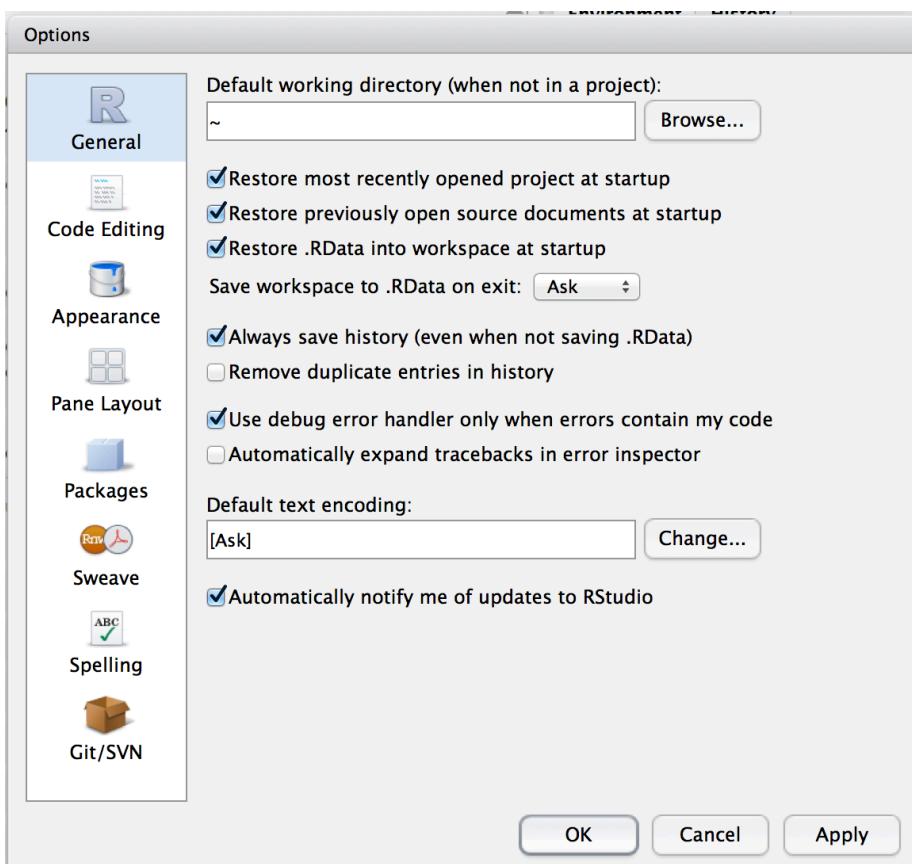


Figure 2.6: The options window in RStudio. On a Mac, you can open this window by going to the “RStudio” menu and selecting “Preferences”. On a Windows machine you go to the “Tools” menu and select “Global Options”

2.14 Summary

Every book that tries to introduce basic programming ideas to novices has to cover roughly the same topics, and in roughly the same order. Mine is no

exception, and so in the grand tradition of doing it just the same way everyone else did it, this chapter covered the following topics:

- Getting started. We downloaded and installed R and RStudio
- Basic commands. We talked a bit about the logic of how R works and in particular how to type commands into the R console (Section @ref(#firstcommand), and in doing so learned how to perform basic calculations using the arithmetic operators `+`, `-`, `*`, `/` and `^`.
- Introduction to functions. We saw several different functions, three that are used to perform numeric calculations (`sqrt()`, `abs()`, `round()`, one that applies to text (`nchar()`; Section 2.10.1), and one that works on any variable (`length()`; Section 2.9.5). In doing so, we talked a bit about how argument names work, and learned about default values for arguments. (Section 2.7.1)
- Introduction to variables. We learned the basic idea behind variables, and how to assign values to variables using the assignment operator `<-` (Section 2.6). We also learned how to create vectors using the combine function `c()` (Section 2.9).
- Data types. Learned the distinction between numeric, character and logical data; including the basics of how to enter and use each of them. (Sections 2.6 to 2.11)
- Logical operations. Learned how to use the logical operators `==`, `!=`, `<`, `>`, `<=`, `=>`, `!`, `&` and `|`. And learned how to use logical indexing. (Section 2.12)

We still haven't arrived at anything that resembles a "data set", of course. Maybe the next Chapter will get us a bit closer...

2.15 Additional R concepts

Form follows function

– Louis Sullivan

So far, our main goal was to get started in R. As we go through the book we'll run into a lot of new R concepts, which I'll explain alongside the relevant data analysis concepts. However, there's still quite a few things that I need to talk about now, otherwise we'll run into problems when we start trying to work with data and do statistics. So that's the goal in this section: to build on the introductory content from the last section, to get you to the point that we can start using R for statistics. Broadly speaking, the section comes in two parts. The first half of the section is devoted to the "mechanics" of R: installing and loading packages, managing the workspace, navigating the file system, and loading and saving data. In the second half, I'll talk more about what kinds

of variables exist in R, and introduce three new kinds of variables: factors, data frames and formulas. I'll finish up by talking a little bit about the help documentation in R as well as some other avenues for finding assistance. In general, I'm not trying to be comprehensive in this chapter, I'm trying to make sure that you've got the basic foundations needed to tackle the content that comes later in the book. However, a lot of the topics are revisited in more detail later, especially in Chapters ?? and ??.

2.16 Using comments

Before discussing any of the more complicated stuff, I want to introduce the ***comment*** character, **#**. It has a simple meaning: it tells R to ignore everything else you've written on this line. You won't have much need of the **#** character immediately, but it's very useful later on when writing scripts (see Chapter ??). However, while you don't need to use it, I want to be able to include comments in my R extracts. For instance, if you read this:³⁰

```
seeker <- 3.1415      # create the first variable
lover <- 2.7183       # create the second variable
keeper <- seeker * lover # now multiply them to create a third one
print( keeper )        # print out the value of 'keeper'

## [1] 8.539539
```

it's a lot easier to understand what I'm doing than if I just write this:

```
seeker <- 3.1415
lover <- 2.7183
keeper <- seeker * lover
print( keeper )
```

```
## [1] 8.539539
```

You might have already noticed that the code extracts in Chapter 2 included the **#** character, but from now on, you'll start seeing **#** characters appearing in the extracts, with some human-readable explanatory remarks next to them. These are still perfectly legitimate commands, since R knows that it should ignore the **#** character and everything after it. But hopefully they'll help make things a little easier to understand.

³⁰Notice that I used `print(keeper)` rather than just typing `keeper`. Later on in the text I'll sometimes use the `print()` function to display things because I think it helps make clear what I'm doing, but in practice people rarely do this.

2.17 Installing and loading packages

In this section I discuss R *packages*, since almost all of the functions you might want to use in R come in packages. A package is basically just a big collection of functions, data sets and other R objects that are all grouped together under a common name. Some packages are already installed when you put R on your computer, but the vast majority of them of R packages are out there on the internet, waiting for you to download, install and use them.

When I first started writing this book, RStudio didn't really exist as a viable option for using R, and as a consequence I wrote a very lengthy section that explained how to do package management using raw R commands. It's not actually terribly hard to work with packages that way, but it's clunky and unpleasant. Fortunately, we don't have to do things that way anymore. In this section, I'll describe how to work with packages using the RStudio tools, because they're so much simpler. Along the way, you'll see that whenever you get RStudio to do something (e.g., install a package), you'll actually see the R commands that get created. I'll explain them as we go, because I think that helps you understand what's going on.

However, before we get started, there's a critical distinction that you need to understand, which is the difference between having a package *installed* on your computer, and having a package *loaded* in R. As of this writing, there are just over 5000 R packages freely available "out there" on the internet.³¹ When you install R on your computer, you don't get all of them: only about 30 or so come bundled with the basic R installation. So right now there are about 30 packages "installed" on your computer, and another 5000 or so that are not installed. So that's what installed means: it means "it's on your computer somewhere". The critical thing to remember is that just because something is on your computer doesn't mean R can use it. In order for R to be able to *use* one of your 30 or so installed packages, that package must also be "loaded". Generally, when you open up R, only a few of these packages (about 7 or 8) are actually loaded. Basically what it boils down to is this:

A package must be installed before it can be loaded.

A package must be loaded before it can be used.

This two step process might seem a little odd at first, but the designers of R had very good reasons to do it this way,³² and you get the hang of it pretty quickly.

³¹More precisely, there are 5000 or so packages on CRAN, the Comprehensive R Archive Network.

³²Basically, the reason is that there are 5000 packages, and probably about 4000 authors of packages, and no-one really knows what all of them do. Keeping the installation separate from the loading minimizes the chances that two packages will interact with each other in a nasty way.

2.17.1 The package panel in RStudio

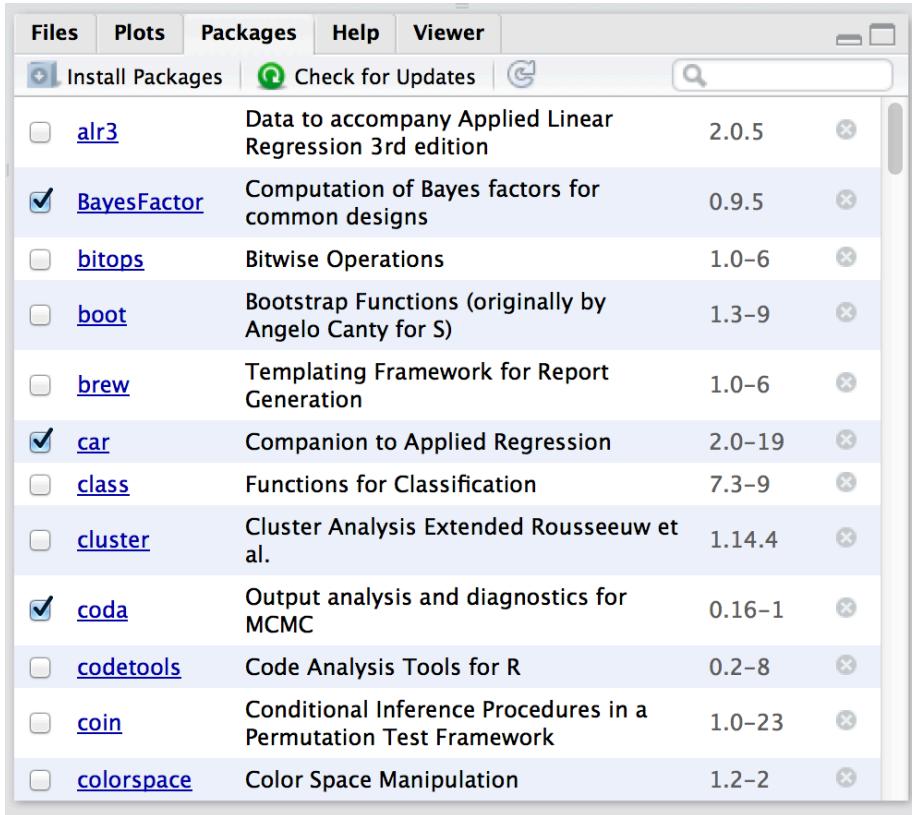


Figure 2.7: The packages panel.

Right, lets get started. The first thing you need to do is look in the lower right hand panel in RStudio. You'll see a tab labelled "Packages". Click on the tab, and you'll see a list of packages that looks something like Figure 2.7. Every row in the panel corresponds to a different package, and every column is a useful piece of information about that package.³³ Going from left to right, here's what each column is telling you:

- The check box on the far left column indicates whether or not the package is loaded.
- The one word of text immediately to the right of the check box is the name of the package.
- The short passage of text next to the name is a brief description of the package.

³³If you're using the command line, you can get the same information by typing `library()` at the command line.

- The number next to the description tells you what version of the package you have installed.
- The little x-mark next to the version number is a button that you can push to uninstall the package from your computer (you almost never need this).

2.17.2 Loading a package

That seems straightforward enough, so let's try loading and unloading packages. For this example, I'll use the `foreign` package. The `foreign` package is a collection of tools that are very handy when R needs to interact with files that are produced by other software packages (e.g., SPSS). It comes bundled with R, so it's one of the ones that you have installed already, but it won't be one of the ones loaded. Inside the `foreign` package is a function called `read.spss()`. It's a handy little function that you can use to import an SPSS data file into R, so let's pretend we want to use it. Currently, the `foreign` package isn't loaded, so if I ask R to tell me if it knows about a function called `read.spss()` it tells me that there's no such thing...

```
exists( "read.spss" )  
  
## [1] FALSE
```

Now let's load the package. In RStudio, the process is dead simple: go to the package tab, find the entry for the `foreign` package, and check the box on the left hand side. The moment that you do this, you'll see a command like this appear in the R console:

```
library("foreign", lib.loc="/Library/Frameworks/R.framework/Versions/3.0/Resources/lib")
```

The `lib.loc` bit will look slightly different on Macs versus on Windows, because that part of the command is just RStudio telling R where to look to find the installed packages. What I've shown you above is the Mac version. On a Windows machine, you'll probably see something that looks like this:

```
library("foreign", lib.loc="C:/Program Files/R/R-3.0.2/library")
```

But actually it doesn't matter much. The `lib.loc` bit is almost always unnecessary. Unless you've taken to installing packages in idiosyncratic places (which is something that you can do if you really want) R already knows where to look. So in the vast majority of cases, the command to load the `foreign` package is just this:

```
library("foreign")
```

Throughout this book, you'll often see me typing in `library()` commands. You don't actually have to type them in yourself: you can use the RStudio package panel to do all your package loading for you. The only reason I include the `library()` commands sometimes is as a reminder to you to make sure that you have the relevant package loaded. Oh, and I suppose we should check to see if our attempt to load the package actually worked. Let's see if R now knows about the existence of the `read.spss()` function...

```
exists( "read.spss" )
```

```
## [1] TRUE
```

Yep. All good.

2.17.3 Unloading a package

Sometimes, especially after a long session of working with R, you find yourself wanting to get rid of some of those packages that you've loaded. The RStudio package panel makes this exactly as easy as loading the package in the first place. Find the entry corresponding to the package you want to unload, and uncheck the box. When you do that for the `foreign` package, you'll see this command appear on screen:

```
detach("package:foreign", unload=TRUE)
```

And the package is unloaded. We can verify this by seeing if the `read.spss()` function still `exists()`:

```
exists( "read.spss" )
```

```
## [1] FALSE
```

Nope. Definitely gone.

2.17.4 A few extra comments

Sections 2.17.2 and 2.17.3 cover the main things you need to know about loading and unloading packages. However, there's a couple of other details that I want to draw your attention to. A concrete example is the best way to illustrate. One of the other packages that you already have installed on your computer is the `Matrix` package, so let's load that one and see what happens:

```
library( Matrix )
## Loading required package: lattice
```

This is slightly more complex than the output that we got last time, but it's not too complicated. The `Matrix` package makes use of some of the tools in the `lattice` package, and R has kept track of this dependency. So when you try to load the `Matrix` package, R recognises that you're also going to need to have the `lattice` package loaded too. As a consequence, *both* packages get loaded, and R prints out a helpful little note on screen to tell you that it's done so.

R is pretty aggressive about enforcing these dependencies. Suppose, for example, I try to unload the `lattice` package while the `Matrix` package is still loaded. This is easy enough to try: all I have to do is uncheck the box next to "lattice" in the packages panel. But if I try this, here's what happens:

```
detach("package:lattice", unload=TRUE)
## Error: package `lattice' is required by `Matrix' so will not be detached
```

R refuses to do it. This can be quite useful, since it stops you from accidentally removing something that you still need. So, if I want to remove both `Matrix` and `lattice`, I need to do it in the correct order

Something else you should be aware of. Sometimes you'll attempt to load a package, and R will print out a message on screen telling you that something or other has been "masked". This will be confusing to you if I don't explain it now, and it actually ties very closely to the whole reason why R forces you to load packages separately from installing them. Here's an example. Two of the package that I'll refer to a lot in this book are called `car` and `psych`. The `car` package is short for "Companion to Applied Regression" (which is a really great book, I'll add), and it has a lot of tools that I'm quite fond of. The `car` package was written by a guy called John Fox, who has written a lot of great statistical tools for social science applications. The `psych` package was written by William Revelle, and it has a lot of functions that are very useful for psychologists in particular, especially in regards to psychometric techniques. For the most part, `car` and `psych` are quite unrelated to each other. They do different things, so not surprisingly almost all of the function names are different. But... there's one exception to that. The `car` package and the `psych` package *both* contain a function called `logit()`.³⁴ This creates a naming conflict. If I load both packages into R, an ambiguity is created. If the user types in `logit(100)`, should R use the `logit()` function in the `car` package, or the one in the `psych` package? The answer is: R uses whichever package you loaded most recently,

³⁴The logit function a simple mathematical function that happens not to have been included in the basic R distribution.

and it tells you this very explicitly. Here's what happens when I load the `car` package, and then afterwards load the `psych` package:

```
library(car)

## Loading required package: carData

library(psych)

##
## Attaching package: 'psych'

## The following object is masked from 'package:car':
##
##     logit
```

The output here is telling you that the `logit` object (i.e., function) in the `car` package is no longer accessible to you. It's been hidden (or "masked") from you by the one in the `psych` package.³⁵

2.17.5 Downloading new packages

One of the main selling points for R is that there are thousands of packages that have been written for it, and these are all available online. So whereabouts online are these packages to be found, and how do we download and install them? There is a big repository of packages called the "Comprehensive R Archive Network" (CRAN), and the easiest way of getting and installing a new package is from one of the many CRAN mirror sites. Conveniently for us, R provides a function called `install.packages()` that you can use to do this. Even *more* conveniently, the RStudio team runs its own CRAN mirror and RStudio has a clean interface that lets you install packages without having to learn how to use the `install.packages()` command³⁶

Using the RStudio tools is, again, dead simple. In the top left hand corner of the packages panel (Figure 2.7) you'll see a button called "Install Packages". If you click on that, it will bring up a window like the one shown in Figure 2.8.

There are a few different buttons and boxes you can play with. Ignore most of them. Just go to the line that says "Packages" and start typing the name of the package that you want. As you type, you'll see a dropdown menu appear (Figure 2.9), listing names of packages that start with the letters that you've typed so far.

³⁵Tip for advanced users. You can get R to use the one from the `car` package by using `car::logit()` as your command rather than `logit()`, since the `car::` part tells R explicitly which package to use. See also `:::` if you're especially keen to force R to use functions it otherwise wouldn't, but take care, since `:::` can be dangerous.

³⁶It is not very difficult.

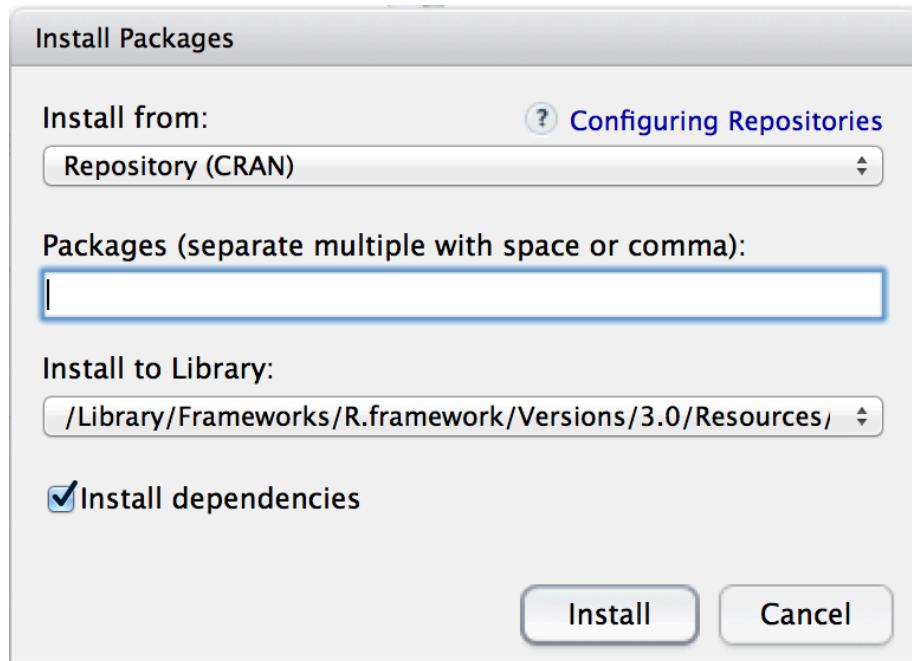


Figure 2.8: The package installation dialog box in RStudio

You can select from this list, or just keep typing. Either way, once you've got the package name that you want, click on the install button at the bottom of the window. When you do, you'll see the following command appear in the R console:

```
install.packages("psych")
```

This is the R command that does all the work. R then goes off to the internet, has a conversation with CRAN, downloads some stuff, and installs it on your computer. You probably don't care about all the details of R's little adventure on the web, but the `install.packages()` function is rather chatty, so it reports a bunch of gibberish that you really aren't all that interested in:

```
trying URL 'http://cran.rstudio.com/bin/macosx/contrib/3.0/psych_1.4.1.tgz'
Content type 'application/x-gzip' length 2737873 bytes (2.6 Mb)
opened URL
=====
downloaded 2.6 Mb
```

The downloaded binary packages are in

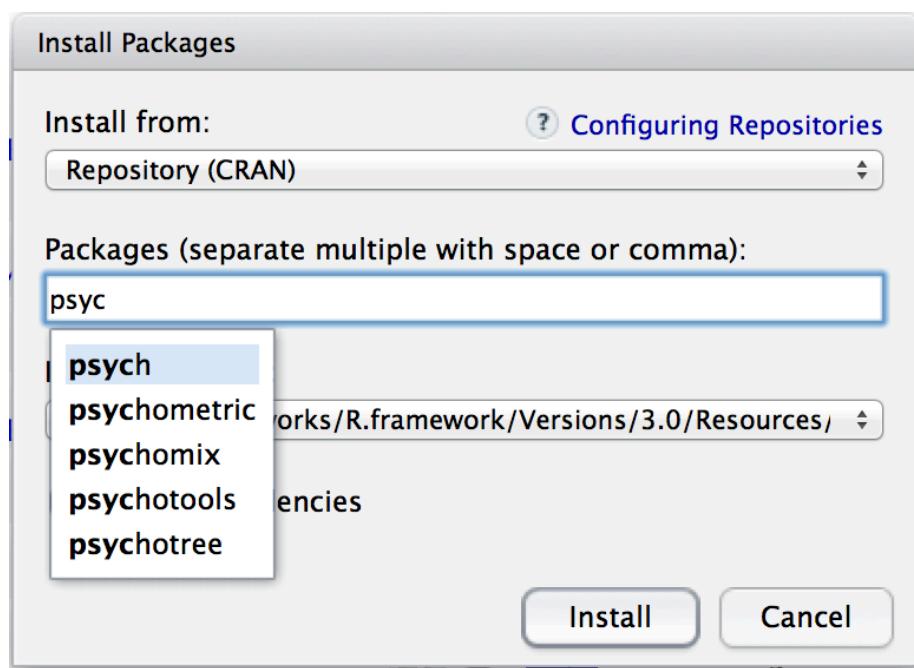


Figure 2.9: When you start typing, you'll see a dropdown menu suggest a list of possible packages that you might want to install

```
/var/folders/cl/thhsyrz53g73q0w1kb5z31_80000gn/T//RtmpmQ9VT3/downloaded_packages
```

Despite the long and tedious response, all that really means is “I’ve installed the psych package”. I find it best to humour the talkative little automaton. I don’t actually read any of this garbage, I just politely say “thanks” and go back to whatever I was doing.

2.17.6 Updating R and R packages

Every now and then the authors of packages release updated versions. The updated versions often add new functionality, fix bugs, and so on. It’s generally a good idea to update your packages periodically. There’s an `update.packages()` function that you can use to do this, but it’s probably easier to stick with the RStudio tool. In the packages panel, click on the “Update Packages” button. This will bring up a window that looks like the one shown in Figure 2.10. In this window, each row refers to a package that needs to be updated. You can tell R which updates you want to install by checking the boxes on the left. If you’re feeling lazy and just want to update everything, click the “Select All” button, and then click the “Install Updates” button. R then prints out a *lot* of garbage on the screen, individually downloading and installing all the new packages. This might take a while to complete depending on how good your internet connection is. Go make a cup of coffee. Come back, and all will be well.

About every six months or so, a new version of R is released. You can’t update R from within RStudio (not to my knowledge, at least): to get the new version you can go to the CRAN website and download the most recent version of R, and install it in the same way you did when you originally installed R on your computer. This used to be a slightly frustrating event, because whenever you downloaded the new version of R, you would lose all the packages that you’d downloaded and installed, and would have to repeat the process of re-installing them. This was pretty annoying, and there were some neat tricks you could use to get around this. However, newer versions of R don’t have this problem so I no longer bother explaining the workarounds for that issue.

2.17.7 What packages does this book use?

There are several packages that I make use of in this book. The most prominent ones are:

- `lsr`. This is the *Learning Statistics with R* package that accompanies this book. It doesn’t have a lot of interesting high-powered tools: it’s just a small collection of handy little things that I think can be useful to novice users. As you get more comfortable with R this package should start to feel pretty useless to you.

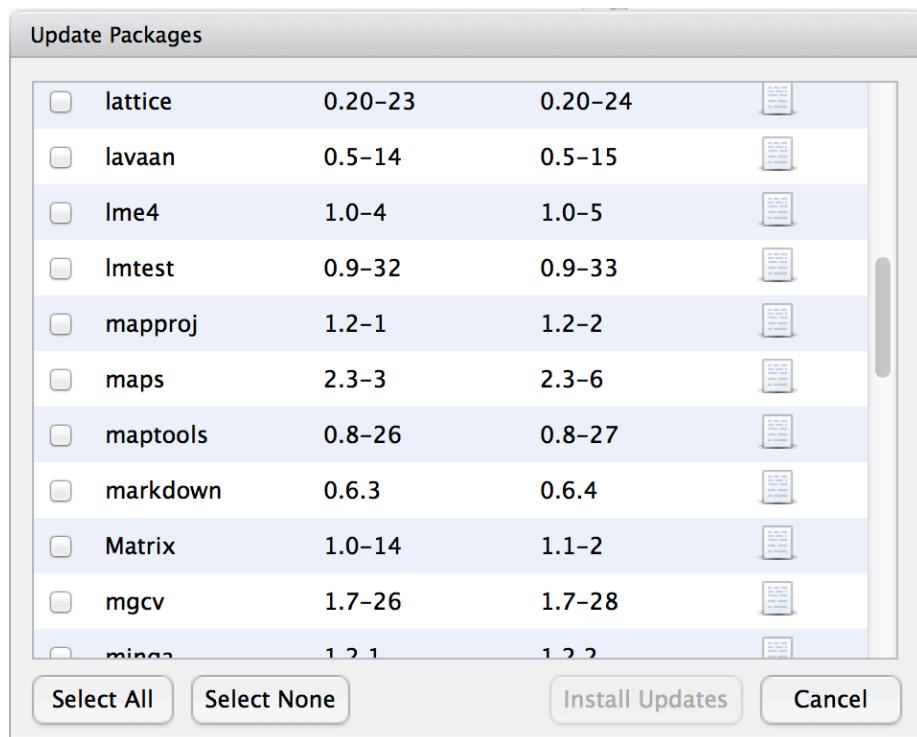


Figure 2.10: The RStudio dialog box for updating packages

- **psych.** This package, written by William Revelle, includes a lot of tools that are of particular use to psychologists. In particular, there's several functions that are particularly convenient for producing analyses or summaries that are very common in psych, but less common in other disciplines.
- **car.** This is the *Companion to Applied Regression* package, which accompanies the excellent book of the same name by (Fox and Weisberg, 2011). It provides a lot of very powerful tools, only some of which we'll touch in this book.

Besides these three, there are a number of packages that I use in a more limited fashion: **gplots**, **sciplot**, **foreign**, **effects**, **R.matlab**, **gdata**, **lmtest**, and probably one or two others that I've missed. There are also a number of packages that I refer to but don't actually use in this book, such as **reshape**, **compute.es**, **HistData** and **multcomp** among others. Finally, there are a number of packages that provide more advanced tools that I hope to talk about in future versions of the book, such as **sem**, **ez**, **nlme** and **lme4**. In any case, whenever I'm using a function that isn't in the core packages, I'll make sure to note this in the text.

2.18 Managing the workspace

Let's suppose that you're reading through this book, and what you're doing is sitting down with it once a week and working through a whole chapter in each sitting. Not only that, you've been following my advice and typing in all these commands into R. So far during this chapter, you'd have typed quite a few commands, although the only ones that actually involved creating variables were the ones you typed during Section 2.16. As a result, you currently have three variables; **seeker**, **lover**, and **keeper**. These three variables are the contents of your **workspace**, also referred to as the **global environment**. The workspace is a key concept in R, so in this section we'll talk a lot about what it is and how to manage its contents.

2.18.1 Listing the contents of the workspace

The first thing that you need to know how to do is examine the contents of the workspace. If you're using RStudio, you will probably find that the easiest way to do this is to use the "Environment" panel in the top right hand corner. Click on that, and you'll see a list that looks very much like the one shown in Figures 2.11 and 2.12. If you're using the command line, then the **objects()** function may come in handy:

```
objects()
```

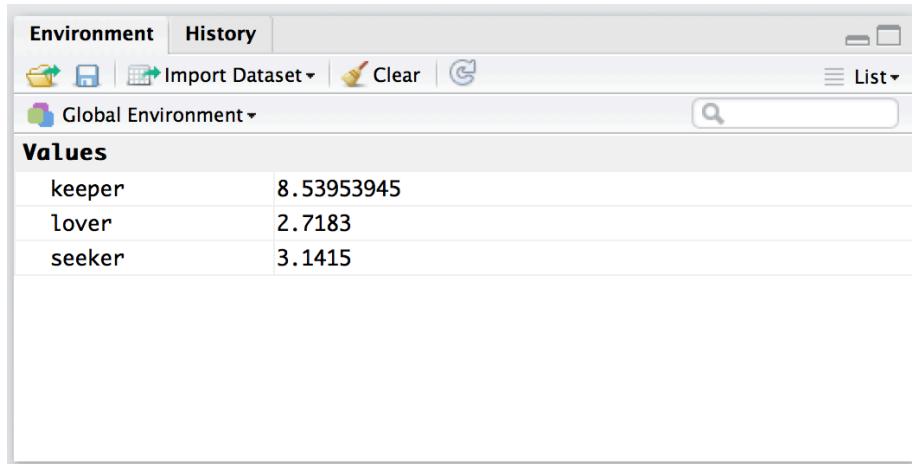


Figure 2.11: The RStudio Environment panel shows you the contents of the workspace. The view shown above is the list view. To switch to the grid view, click on the menu item on the top right that currently reads list. Select grid from the dropdown menu, and then it will switch to a view like the one shown in the other workspace figure

A screenshot of the RStudio Environment panel in Grid view. The top navigation bar and toolbar are identical to Figure 2.11. The "Global Environment" dropdown is open. The main area displays a table with columns: Name, Type, Length, Size, and Value. The data rows are: keeper (numeric, 1, 48 B, 8.53953945), lover (numeric, 1, 48 B, 2.7183), and seeker (numeric, 1, 48 B, 3.1415).

	Name	Type	Length	Size	Value
<input type="checkbox"/>	keeper	numeric	1	48 B	8.53953945
<input type="checkbox"/>	lover	numeric	1	48 B	2.7183
<input type="checkbox"/>	seeker	numeric	1	48 B	3.1415

Figure 2.12: The RStudio “Environment” panel shows you the contents of the workspace. Compare this “grid” view to the “list” earlier

```
## [1] "any.sales.this.month" "berkeley"           "berkeley.small"
## [4] "coef"                  "days.per.month"      "february.sales"
## [7] "greeting"              "is.the.Party.correct" "keeper"
## [10] "lover"                 "months"             "profit"
## [13] "projecthome"           "revenue"            "royalty"
## [16] "sales"                 "sales.by.month"     "seeker"
## [19] "simpson"               "stock.levels"       "x"
## [22] "xlu"
```

Of course, in the true R tradition, the `objects()` function has a lot of fancy capabilities that I'm glossing over in this example. Moreover there are also several other functions that you can use, including `ls()` which is pretty much identical to `objects()`, and `ls.str()` which you can use to get a fairly detailed description of all the variables in the workspace. In fact, the `lsr` package actually includes its own function that you can use for this purpose, called `who()`. The reason for using the `who()` function is pretty straightforward: in my everyday work I find that the output produced by the `objects()` command isn't *quite* informative enough, because the only thing it prints out is the name of each variable; but the `ls.str()` function is *too* informative, because it prints out a lot of additional information that I really don't like to look at. The `who()` function is a compromise between the two. First, now that we've got the `lsr` package installed, we need to load it:

```
library(lsrr)
```

and now we can use the `who()` function:

```
who()
```

##	-- Name --	-- Class --	-- Size --
##	any.sales.this.month	logical	12
##	berkeley	data.frame	39 x 3
##	berkeley.small	data.frame	46 x 2
##	coef	numeric	2
##	days.per.month	numeric	12
##	february.sales	numeric	1
##	greeting	character	1
##	is.the.Party.correct	logical	1
##	keeper	numeric	1
##	lover	numeric	1
##	months	character	12
##	profit	numeric	12
##	projecthome	character	1
##	revenue	numeric	1
##	royalty	numeric	1

```
##   sales           numeric    1
## sales.by.month  numeric    12
## seeker          numeric    1
## simpson         matrix    6 x 5
## stock.levels   character 12
## x               logical    3
## xlu              numeric    1
```

As you can see, the `who()` function lists all the variables and provides some basic information about what kind of variable each one is and how many elements it contains. Personally, I find this output much easier more useful than the very compact output of the `objects()` function, but less overwhelming than the extremely verbose `ls.str()` function. Throughout this book you'll see me using the `who()` function a lot. You don't have to use it yourself: in fact, I suspect you'll find it easier to look at the RStudio environment panel. But for the purposes of writing a textbook I found it handy to have a nice text based description: otherwise there would be about another 100 or so screenshots added to the book.³⁷

2.18.2 Removing variables from the workspace

Looking over that list of variables, it occurs to me that I really don't need them any more. I created them originally just to make a point, but they don't serve any useful purpose anymore, and now I want to get rid of them. I'll show you how to do this, but first I want to warn you – there's no “undo” option for variable removal. Once a variable is removed, it's gone forever unless you save it to disk. I'll show you how to do *that* in Section 2.20, but quite clearly we have no need for these variables at all, so we can safely get rid of them.

In RStudio, the easiest way to remove variables is to use the environment panel. Assuming that you're in grid view (i.e., Figure 2.12), check the boxes next to the variables that you want to delete, then click on the “Clear” button at the top of the panel. When you do this, RStudio will show a dialog box asking you to confirm that you really do want to delete the variables. It's always worth checking that you really do, because as RStudio is at pains to point out, you can't undo this. Once a variable is deleted, it's gone.³⁸ In any case, if you click “yes”, that variable will disappear from the workspace: it will no longer appear in the environment panel, and it won't show up when you use the `who()` command.

Suppose you don't access to RStudio, and you still want to remove variables. This is where the **remove** function `rm()` comes in handy. The simplest way to

³⁷This would be especially annoying if you're reading an electronic copy of the book because the text displayed by the `who()` function is searchable, whereas text shown in a screen shot isn't!

³⁸Mind you, all that means is that it's been removed from the workspace. If you've got the data saved to file somewhere, then that *file* is perfectly safe.

use `rm()` is just to type in a (comma separated) list of all the variables you want to remove. Let's say I want to get rid of `seeker` and `lover`, but I would like to keep `keeper`. To do this, all I have to do is type:

```
rm( seeker, lover )
```

There's no visible output, but if I now inspect the workspace

```
who()
```

##	-- Name --	-- Class --	-- Size --
##	any.sales.this.month	logical	12
##	berkeley	data.frame	39 x 3
##	berkeley.small	data.frame	46 x 2
##	coef	numeric	2
##	days.per.month	numeric	12
##	february.sales	numeric	1
##	greeting	character	1
##	is.the.Party.correct	logical	1
##	keeper	numeric	1
##	months	character	12
##	profit	numeric	12
##	projecthome	character	1
##	revenue	numeric	1
##	royalty	numeric	1
##	sales	numeric	1
##	sales.by.month	numeric	12
##	simpson	matrix	6 x 5
##	stock.levels	character	12
##	x	logical	3
##	xlu	numeric	1

I see that there's only the `keeper` variable left. As you can see, `rm()` can be very handy for keeping the workspace tidy.

2.19 Navigating the file system

In this section I talk a little about how R interacts with the file system on your computer. It's not a terribly interesting topic, but it's useful. As background to this discussion, I'll talk a bit about how file system locations work in Section 2.19.1. Once upon a time *everyone* who used computers could safely be assumed to understand how the file system worked, because it was impossible to successfully use a computer if you didn't! However, modern operating

systems are much more “user friendly”, and as a consequence of this they go to great lengths to hide the file system from users. So these days it’s not at all uncommon for people to have used computers most of their life and not be familiar with the way that computers organise files. If you already know this stuff, skip straight to Section 2.19.2. Otherwise, read on. I’ll try to give a brief introduction that will be useful for those of you who have never been forced to learn how to navigate around a computer using a DOS or UNIX shell.

2.19.1 The file system itself

In this section I describe the basic idea behind file locations and file paths. Regardless of whether you’re using Window, Mac OS or Linux, every file on the computer is assigned a (fairly) human readable address, and every address has the same basic structure: it describes a *path* that starts from a *root* location , through as series of *folders* (or if you’re an old-school computer user, *directories*), and finally ends up at the file.

On a Windows computer the root is the physical drive³⁹ on which the file is stored, and for most home computers the name of the hard drive that stores all your files is C: and therefore most file names on Windows begin with C:. After that comes the folders, and on Windows the folder names are separated by a \ symbol. So, the complete path to this book on my Windows computer might be something like this:

C:\Users\danRbook\LSR.pdf

and what that *means* is that the book is called LSR.pdf, and it’s in a folder called book which itself is in a folder called dan which itself is ... well, you get the idea. On Linux, Unix and Mac OS systems, the addresses look a little different, but they’re more or less identical in spirit. Instead of using the backslash, folders are separated using a forward slash, and unlike Windows, they don’t treat the physical drive as being the root of the file system. So, the path to this book on my Mac might be something like this:

/Users/dan/Rbook/LSR.pdf

So that’s what we mean by the “path” to a file. The next concept to grasp is the idea of a *working directory* and how to change it. For those of you who have used command line interfaces previously, this should be obvious already. But if not, here’s what I mean. The working directory is just “whatever folder I’m currently looking at”. Suppose that I’m currently looking for files in Explorer (if you’re using Windows) or using Finder (on a Mac). The folder I currently

³⁹Well, the partition, technically.

Table 2.4: Basic arithmetic operations in R. These five operators are used very frequently throughout the text, so it’s important to be familiar with them at the outset.

absolute path (i.e., from root)	relative path (i.e. from C:\Users\danRbook)
C:\\\\Users\\\\dan	..
C:\\\\Users	..\\\\.. \\\\
C:\\\\Users\\\\danRbook\\\\source	.\\\\source
C:\\\\Users\\\\dan\\\\nerdstuff	..\\\\nerdstuff

have open is my user directory (i.e., C:\\Users\\dan or /Users/dan). That’s my current working directory.

The fact that we can imagine that the program is “in” a particular directory means that we can talk about moving *from* our current location *to* a new one. What that means is that we might want to specify a new location in relation to our current location. To do so, we need to introduce two new conventions. Regardless of what operating system you’re using, we use . to refer to the current working directory, and .. to refer to the directory above it. This allows us to specify a path to a new location in relation to our current location, as the following examples illustrate. Let’s assume that I’m using my Windows computer, and my working directory is C:\\Users\\danRbook). The table below shows several addresses in relation to my current one:

There’s one last thing I want to call attention to: the ~ directory. I normally wouldn’t bother, but R makes reference to this concept sometimes. It’s quite common on computers that have multiple users to define ~ to be the user’s home directory. On my Mac, for instance, the home directory ~ for the “dan” user is \\Users\\dan\\. And so, not surprisingly, it is possible to define other directories in terms of their relationship to the home directory. For example, an alternative way to describe the location of the LSR.pdf file on my Mac would be

~Rbook\\LSR.pdf

That’s about all you really need to know about file paths. And since this section already feels too long, it’s time to look at how to navigate the file system in R.

2.19.2 Navigating the file system using the R console

In this section I’ll talk about how to navigate this file system from within R itself. It’s not particularly user friendly, and so you’ll probably be happy to know that RStudio provides you with an easier method, and I will describe it in Section 2.19.4. So in practice, you won’t *really* need to use the commands

that I babble on about in this section, but I do think it helps to see them in operation at least once before forgetting about them forever.

Okay, let's get started. When you want to load or save a file in R it's important to know what the working directory is. You can find out by using the `getwd()` command. For the moment, let's assume that I'm using Mac OS or Linux, since there's some subtleties to Windows. Here's what happens:

```
getwd()
## [1] "/Users/dan"
```

We can change the working directory quite easily using `setwd()`. The `setwd()` function has only the one argument, `dir`, is a character string specifying a path to a directory, or a path relative to the working directory. Since I'm currently located at `/Users/dan`, the following two are equivalent:

```
setwd("/Users/dan/Rbook/data")
setwd("./Rbook/data")
```

Now that we're here, we can type `list.files()` command to get a listing of all the files in that directory. Since this is the directory in which I store all of the data files that we'll use in this book, here's what we get as the result:

```
list.files()
## [1] "afl24.Rdata"           "aflsmall.Rdata"        "aflsmall12.Rdata"
## [4] "agpp.Rdata"            "all.zip"              "annoying.Rdata"
## [7] "anscombesquartet.Rdata" "awesome.Rdata"         "awesome2.Rdata"
## [10] "booksales.csv"          "booksales.Rdata"       "booksales2.csv"
## [13] "cakes.Rdata"            "cards.Rdata"           "chapek9.Rdata"
## [16] "chico.Rdata"           "clinicaltrial_old.Rdata" "clinicaltrial.Rdata"
## [19] "coffee.Rdata"           "drugs.wmc.rt.Rdata"    "dwr_all.Rdata"
## [22] "effort.Rdata"           "happy.Rdata"           "harpo.Rdata"
## [25] "harpo2.Rdata"          "likert.Rdata"          "nightgarden.Rdata"
## [28] "nightgarden2.Rdata"     "parenthood.Rdata"      "parenthood2.Rdata"
## [31] "randomness.Rdata"       "repeated.Rdata"        "rtfm.Rdata"
## [34] "salem.Rdata"            "zeppo.Rdata"
```

Not terribly exciting, I'll admit, but it's useful to know about. In any case, there's only one more thing I want to make a note of, which is that R also makes use of the home directory. You can find out what it is by using the `path.expand()` function, like this:

```
path.expand("~")
## [1] "/Users/dan"
```

You can change the user directory if you want, but we’re not going to make use of it very much so there’s no reason to. The only reason I’m even bothering to mention it at all is that when you use RStudio to open a file, you’ll see output on screen that defines the path to the file relative to the ~ directory. I’d prefer you not to be confused when you see it.⁴⁰

2.19.3 Why do the Windows paths use the wrong slash?

Let’s suppose I’m on Windows. As before, I can find out what my current working directory is like this:

```
getwd()
## [1] "C:/Users/dan/"
```

This seems about right, but you might be wondering why R is displaying a Windows path using the wrong type of slash. The answer is slightly complicated, and has to do with the fact that R treats the \ character as “special” (see Section ??). If you’re deeply wedded to the idea of specifying a path using the Windows style slashes, then what you need to do is to type / whenever you mean \. In other words, if you want to specify the working directory on a Windows computer, you need to use one of the following commands:

```
setwd( "C:/Users/dan" )
setwd( "C:\\Users\\dan" )
```

It’s kind of annoying to have to do it this way, but as you’ll see later on in Section ?? it’s a necessary evil. Fortunately, as we’ll see in the next section, RStudio provides a much simpler way of changing directories...

2.19.4 Navigating the file system using the RStudio file panel

Although I think it’s important to understand how all this command line stuff works, in many (maybe even most) situations there’s an easier way. For our purposes, the easiest way to navigate the file system is to make use of RStudio’s built in tools. The “file” panel – the lower right hand area in Figure 2.13 – is actually a pretty decent file browser. Not only can you just point and click on the names to move around the file system, you can also use it to set the working directory, and even load files.

⁴⁰One additional thing worth calling your attention to is the `file.choose()` function. Suppose you want to load a file and you don’t quite remember where it is, but would like to browse for it. Typing `file.choose()` at the command line will open a window in which you can browse to find the file; when you click on the file you want, R will print out the full path to that file. This is kind of handy.

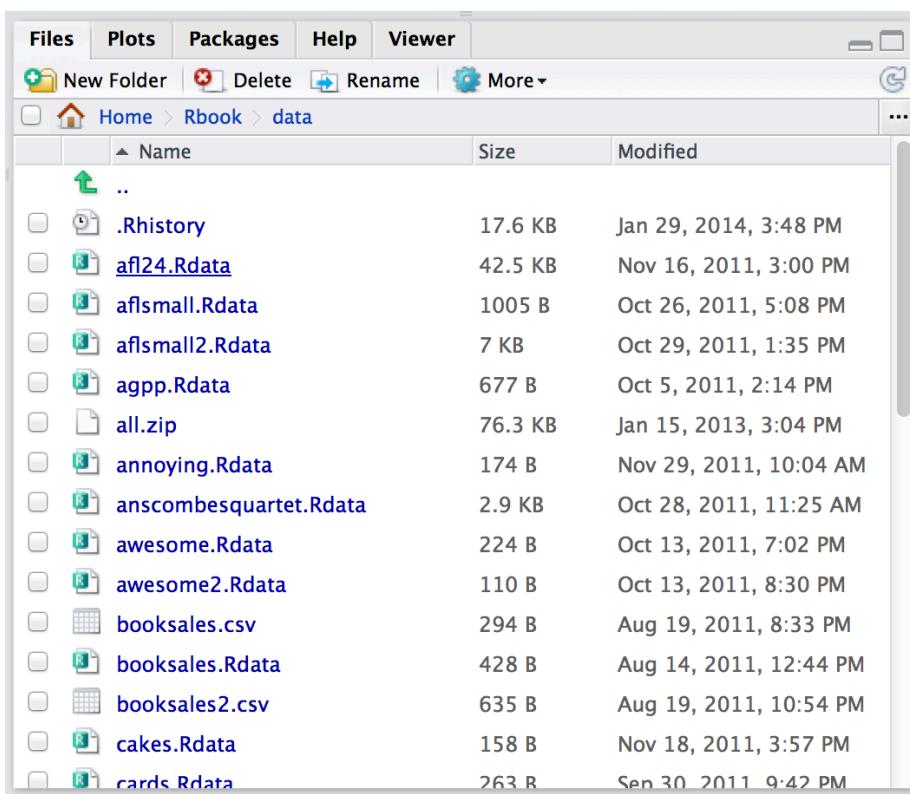


Figure 2.13: The “file panel” is the area shown in the lower right hand corner. It provides a very easy way to browse and navigate your computer using R. See main text for details.

Here's what you need to do to change the working directory using the file panel. Let's say I'm looking at the actual screen shown in Figure 2.13. At the top of the file panel you see some text that says "Home > Rbook > data". What that means is that it's *displaying* the files that are stored in the

```
/Users/dan/Rbook/data
```

directory on my computer. It does *not* mean that this is the R working directory. If you want to change the R working directory, using the file panel, you need to click on the button that reads "More". This will bring up a little menu, and one of the options will be "Set as Working Directory". If you select that option, then R really will change the working directory. You can tell that it has done so because this command appears in the console:

```
setwd("~/Rbook/data")
```

In other words, RStudio sends a command to the R console, exactly as if you'd typed it yourself. The file panel can be used to do other things too. If you want to move "up" to the parent folder (e.g., from `/Users/dan/Rbook/data` to `/Users/dan/Rbook` click on the ".." link in the file panel. To move to a subfolder, click on the name of the folder that you want to open. You can open some types of file by clicking on them. You can delete files from your computer using the "delete" button, rename them with the "rename" button, and so on.

As you can tell, the file panel is a very handy little tool for navigating the file system. But it can do more than just navigate. As we'll see later, it can be used to open files. And if you look at the buttons and menu options that it presents, you can even use it to rename, delete, copy or move files, and create new folders. However, since most of that functionality isn't critical to the basic goals of this book, I'll let you discover those on your own.

2.20 Loading and saving data

There are several different types of files that are likely to be relevant to us when doing data analysis. There are three in particular that are especially important from the perspective of this book:

- *Workspace files* are those with a .Rdata file extension. This is the standard kind of file that R uses to store data and variables. They're called "workspace files" because you can use them to save your whole workspace.
- *Comma separated value (CSV) files* are those with a .csv file extension. These are just regular old text files, and they can be opened with almost any software. It's quite typical for people to store data in CSV files, precisely because they're so simple.

- *Script files* are those with a .R file extension. These aren't data files at all; rather, they're used to save a collection of commands that you want R to execute later. They're just text files, but we won't make use of them until Chapter ??.

There are also several other types of file that R makes use of,⁴¹ but they're not really all that central to our interests. There are also several other kinds of data file that you might want to import into R. For instance, you might want to open Microsoft Excel spreadsheets (.xlsx files), or data files that have been saved in the native file formats for other statistics software, such as SPSS, SAS, Minitab, Stata or Systat. Finally, you might have to handle databases. R tries hard to play nicely with other software, so it has tools that let you open and work with any of these and many others. I'll discuss some of these other possibilities elsewhere in this book (Section ??), but for now I want to focus primarily on the two kinds of data file that you're most likely to need: .Rdata files and .csv files. In this section I'll talk about how to load a workspace file, how to import data from a CSV file, and how to save your workspace to a workspace file. Throughout this section I'll first describe the (sometimes awkward) R commands that do all the work, and then I'll show you the (much easier) way to do it using RStudio.

2.20.1 Loading workspace files using R

When I used the `list.files()` command to list the contents of the `/Users/dan/Rbook/data` directory (in Section 2.19.2), the output referred to a file called `booksales.Rdata`. Let's say I want to load the data from this file into my workspace. The way I do this is with the `load()` function. There are two arguments to this function, but the only one we're interested in is

- `file`. This should be a character string that specifies a path to the file that needs to be loaded. You can use an absolute path or a relative path to do so.

Using the absolute file path, the command would look like this:

```
load( file = "/Users/dan/Rbook/data/booksales.Rdata" )
```

but this is pretty lengthy. Given that the working directory (remember, we changed the directory at the end of Section 2.19.4) is `/Users/dan/Rbook/data`, I could use a relative file path, like so:

```
load( file = "../data/booksales.Rdata" )
```

⁴¹Notably those with .rda, .Rd, .Rhistory, .rd and .rdx extensions

However, my preference is usually to change the working directory first, and *then* load the file. What that would look like is this:

```
setwd( "../data" )           # move to the data directory
load( "booksales.Rdata" )    # load the data
```

If I were then to type `who()` I'd see that there are several new variables in my workspace now. Throughout this book, whenever you see me loading a file, I will assume that the file is actually stored in the working directory, or that you've changed the working directory so that R is pointing at the directory that contains the file. Obviously, *you* don't need type that command yourself: you can use the RStudio file panel to do the work.

2.20.2 Loading workspace files using RStudio

Okay, so how do we open an `.Rdata` file using the RStudio file panel? It's terribly simple. First, use the file panel to find the folder that contains the file you want to load. If you look at Figure 2.13, you can see that there are several `.Rdata` files listed. Let's say I want to load the `booksales.Rdata` file. All I have to do is click on the file name. RStudio brings up a little dialog box asking me to confirm that I do want to load this file. I click yes. The following command then turns up in the console,

```
load("~/Rbook/data/booksales.Rdata")
```

and the new variables will appear in the workspace (you'll see them in the Environment panel in RStudio, or if you type `who()`). So easy it barely warrants having its own section.

2.20.3 Importing data from CSV files using `loadingcsv`

One quite commonly used data format is the humble “comma separated value” file, also called a CSV file, and usually bearing the file extension `.csv`. CSV files are just plain old-fashioned text files, and what they store is basically just a table of data. This is illustrated in Figure 2.14, which shows a file called `booksales.csv` that I've created. As you can see, each row corresponds to a variable, and each row represents the book sales data for one month. The first row doesn't contain actual data though: it has the names of the variables.

If RStudio were not available to you, the easiest way to open this file would be to use the `read.csv()` function.⁴² This function is pretty flexible, and I'll talk

⁴²In a lot of books you'll see the `read.table()` function used for this purpose instead of `read.csv()`. They're more or less identical functions, with the same arguments and everything. They differ only in the default values.

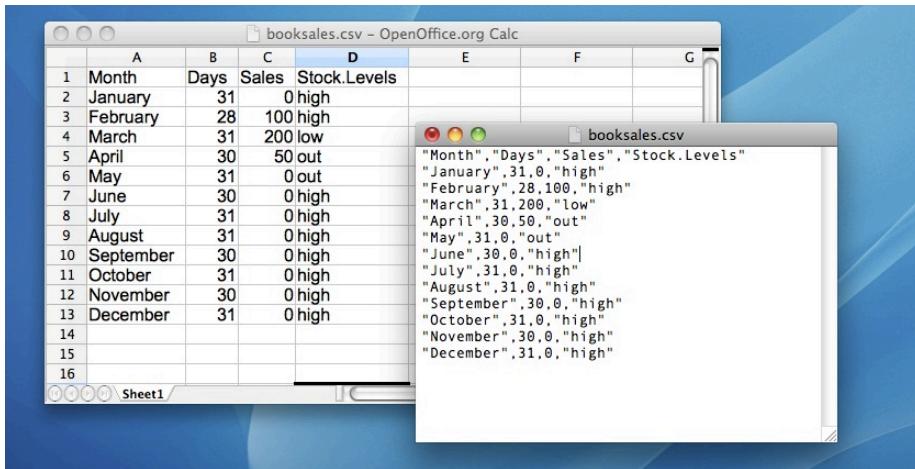


Figure 2.14: The booksales.csv data file. On the left, I've opened the file in using a spreadsheet program (OpenOffice), which shows that the file is basically a table. On the right, the same file is open in a standard text editor (theTextEdit program on a Mac), which shows how the file is formatted. The entries in the table are wrapped in quote marks and separated by commas.

a lot more about it's capabilities in Section ?? for more details, but for now there's only two arguments to the function that I'll mention:

- **file**. This should be a character string that specifies a path to the file that needs to be loaded. You can use an absolute path or a relative path to do so.
- **header**. This is a logical value indicating whether or not the first row of the file contains variable names. The default value is **TRUE**.

Therefore, to import the CSV file, the command I need is:

```
books <- read.csv( file = "booksales.csv" )
```

There are two very important points to notice here. Firstly, notice that I *didn't* try to use the **load()** function, because that function is only meant to be used for .Rdata files. If you try to use **load()** on other types of data, you get an error. Secondly, notice that when I imported the CSV file I assigned the result to a variable, which I imaginatively called **books**.⁴³ file. There's a reason for this. The idea behind an .Rdata file is that it stores a whole workspace. So, if you had the ability to look inside the file yourself you'd see that the data file keeps track of all the variables and their names. So when you **load()** the file,

⁴³Note that I didn't to this in my earlier example when loading the .Rdata

R restores all those original names. CSV files are treated differently: as far as R is concerned, the CSV only stores *one* variable, but that variable is big table. So when you import that table into the workspace, R expects *you* to give it a name.] Let's have a look at what we've got:

```
print( books )
```

```
##      Month Days Sales Stock.Levels
## 1   January    31     0      high
## 2 February    28   100      high
## 3   March     31   200       low
## 4   April     30     50      out
## 5    May      31     0      out
## 6   June      30     0      high
## 7   July      31     0      high
## 8 August     31     0      high
## 9 September   30     0      high
## 10 October    31     0      high
## 11 November   30     0      high
## 12 December   31     0      high
```

Clearly, it's worked, but the format of this output is a bit unfamiliar. We haven't seen anything like this before. What you're looking at is a *data frame*, which is a very important kind of variable in R, and one I'll discuss in Section 2.23. For now, let's just be happy that we imported the data and that it looks about right.

2.20.4 Importing data from CSV files using RStudio

Yet again, it's easier in RStudio. In the environment panel in RStudio you should see a button called "Import Dataset". Click on that, and it will give you a couple of options: select the "From Text File..." option, and it will open up a very familiar dialog box asking you to select a file: if you're on a Mac, it'll look like the usual Finder window that you use to choose a file; on Windows it looks like an Explorer window. An example of what it looks like on a Mac is shown in Figure 2.15. I'm assuming that you're familiar with your own computer, so you should have no problem finding the CSV file that you want to import! Find the one you want, then click on the "Open" button. When you do this, you'll see a window that looks like the one in Figure 2.16.

The import data set window is relatively straightforward to understand.

In the top left corner, you need to type the name of the variable you R to create. By default, that will be the same as the file name: our file is called `booksales.csv`, so RStudio suggests the name `booksales`. If you're happy

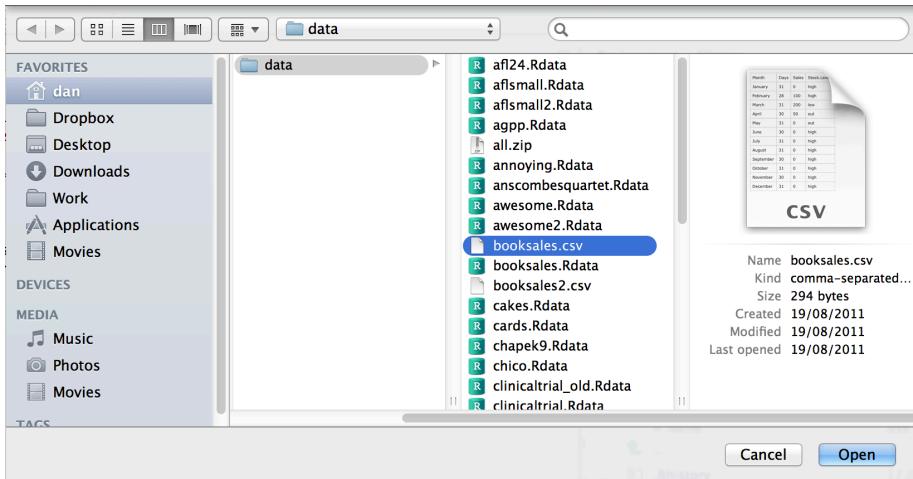


Figure 2.15: A dialog box on a Mac asking you to select the CSV file R should try to import. Mac users will recognise this immediately: it's the usual way in which a Mac asks you to find a file. Windows users won't see this: they'll see the usual explorer window that Windows always gives you when it wants you to select a file.

with that, leave it alone. If not, type something else. Immediately below this are a few things that you can tweak to make sure that the data gets imported correctly:

- Heading. Does the first row of the file contain raw data, or does it contain headings for each variable? The `booksales.csv` file has a header at the top, so I selected “yes”.
- Separator. What character is used to separate different entries? In most CSV files this will be a comma (it is “comma separated” after all). But you can change this if your file is different.
- Decimal. What character is used to specify the decimal point? In English speaking countries, this is almost always a period (i.e., `.`). That's not universally true: many European countries use a comma. So you can change that if you need to.
- Quote. What character is used to denote a block of text? That's usually going to be a double quote mark. It is for the `booksales.csv` file, so that's what I selected.

The nice thing about the RStudio window is that it shows you the raw data file at the top of the window, and it shows you a preview of the data at the bottom. If the data at the bottom doesn't look right, try changing some of the settings on the left hand side. Once you're happy, click “Import”. When you do, two commands appear in the R console:

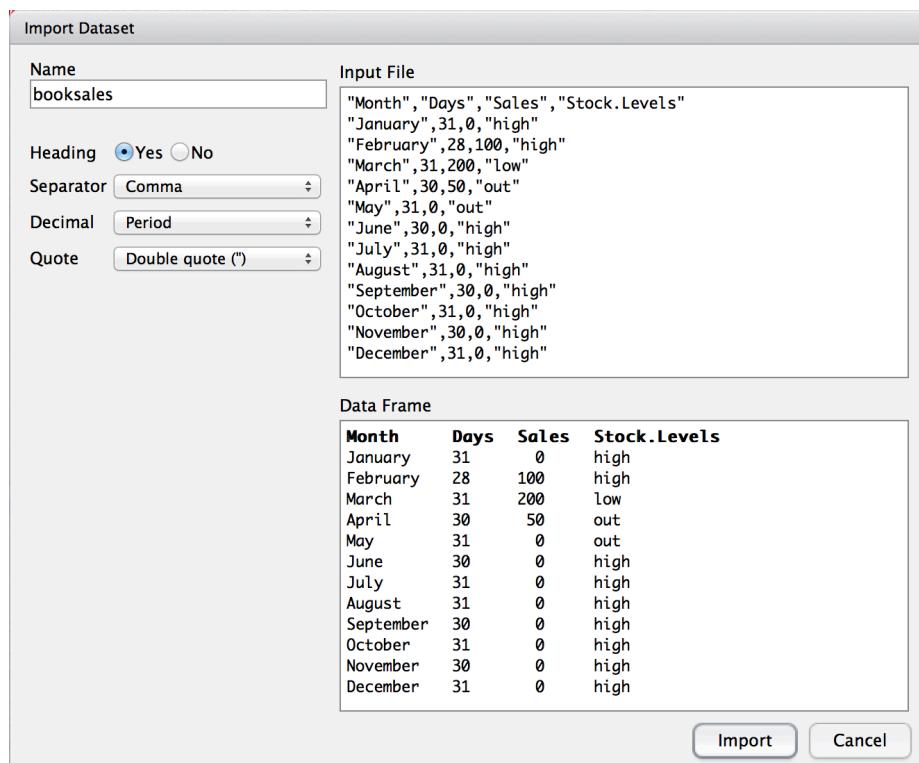


Figure 2.16: The RStudio window for importing a CSV file into R

```
booksales <- read.csv("~/Rbook/data/booksales.csv")
View(booksales)
```

The first of these commands is the one that loads the data. The second one will display a pretty table showing the data in RStudio.

2.20.5 Saving a workspace file using `save`

Not surprisingly, saving data is very similar to loading data. Although RStudio provides a simple way to save files (see below), it's worth understanding the actual commands involved. There are two commands you can use to do this, `save()` and `save.image()`. If you're happy to save *all* of the variables in your workspace into the data file, then you should use `save.image()`. And if you're happy for R to save the file into the current working directory, all you have to do is this:

```
save.image( file = "myfile.Rdata" )
```

Since `file` is the first argument, you can shorten this to `save.image("myfile.Rdata")`; and if you want to save to a different directory, then (as always) you need to be more explicit about specifying the path to the file, just as we discussed in Section 2.19. Suppose, however, I have several variables in my workspace, and I only want to save some of them. For instance, I might have this as my workspace:

```
who()
##   -- Name --   -- Class --   -- Size --
##   data        data.frame    3 x 2
##   handy       character     1
##   junk        numeric      1
```

I want to save `data` and `handy`, but not `junk`. But I don't want to delete `junk` right now, because I want to use it for something else later on. This is where the `save()` function is useful, since it lets me indicate exactly which variables I want to save. Here is one way I can use the `save` function to solve my problem:

```
save(data, handy, file = "myfile.Rdata")
```

Importantly, you *must* specify the name of the `file` argument. The reason is that if you don't do so, R will think that "`myfile.Rdata`" is actually a *variable* that you want to save, and you'll get an error message. Finally, I should mention a second way to specify which variables the `save()` function should save, which is to use the `list` argument. You do so like this:

```
save.me <- c("data", "handy")    # the variables to be saved
save( file = "booksales2.Rdata", list = save.me )    # the command to save them
```

2.20.6 Saving a workspace file using RStudio

RStudio allows you to save the workspace pretty easily. In the environment panel (Figures 2.11 and 2.12) you can see the “save” button. There’s no text, but it’s the same icon that gets used on every computer everywhere: it’s the one that looks like a floppy disk. You know, those things that haven’t been used in about 20 years. Alternatively, go to the “Session” menu and click on the “Save Workspace As...” option.⁴⁴ This will bring up the standard “save” dialog box for your operating system (e.g., on a Mac it’ll look a little bit like the loading dialog box in Figure 2.15). Type in the name of the file that you want to save it to, and all the variables in your workspace will be saved to disk. You’ll see an R command like this one

```
save.image("~/Desktop/Untitled.RData")
```

Pretty straightforward, really.

2.20.7 Other things you might want to save

Until now, we’ve talked mostly about loading and saving *data*. Other things you might want to save include:

- *The output.* Sometimes you might also want to keep a copy of all your interactions with R, including everything that you typed in and everything that R did in response. There are some functions that you can use to get R to write its output to a file rather than to print onscreen (e.g., `sink()`), but to be honest, if you do want to save the R output, the easiest thing to do is to use the mouse to select the relevant text in the R console, go to the “Edit” menu in RStudio and select “Copy”. The output has now been copied to the clipboard. Now open up your favourite text editor or word processing software, and paste it. And you’re done. However, this will only save the contents of the console, not the plots you’ve drawn (assuming you’ve drawn some). We’ll talk about saving images later on.
- *A script.* While it is possible – and sometimes handy – to save the R output as a method for keeping a copy of your statistical analyses, another option

⁴⁴A word of warning: what you *don’t* want to do is use the “File” menu. If you look in the “File” menu you will see “Save” and “Save As...” options, but they don’t save the workspace. Those options are used for dealing with *scripts*, and so they’ll produce `.R` files. We won’t get to those until Chapter ??.

that people use a lot (especially when you move beyond simple “toy” analyses) is to write *scripts*. A script is a text file in which you write out all the commands that you want R to run. You can write your script using whatever software you like. In real world data analysis writing scripts is a key skill – and as you become familiar with R you’ll probably find that most of what you do involves scripting rather than typing commands at the R prompt. However, you won’t need to do much scripting initially, so we’ll leave that until Chapter ??.

2.21 Useful things to know about variables

In Chapter 2 I talked a lot about variables, how they’re assigned and some of the things you can do with them, but there’s a lot of additional complexities. That’s not a surprise of course. However, some of those issues are worth drawing your attention to now. So that’s the goal of this section; to cover a few extra topics. As a consequence, this section is basically a bunch of things that I want to briefly mention, but don’t really fit in anywhere else. In short, I’ll talk about several different issues in this section, which are only loosely connected to one another.

2.21.1 Special values

The first thing I want to mention are some of the “special” values that you might see R produce. Most likely you’ll see them in situations where you were expecting a number, but there are quite a few other ways you can encounter them. These values are `Inf`, `NaN`, `NA` and `NULL`. These values can crop up in various different places, and so it’s important to understand what they mean.

- *Infinity (Inf)*. The easiest of the special values to explain is `Inf`, since it corresponds to a value that is infinitely large. You can also have `-Inf`. The easiest way to get `Inf` is to divide a positive number by 0:

```
1 / 0
```

```
## [1] Inf
```

In most real world data analysis situations, if you’re ending up with infinite numbers in your data, then something has gone awry. Hopefully you’ll never have to see them.

- *Not a Number (NaN)*. The special value of `NaN` is short for “not a number”, and it’s basically a reserved keyword that means “there isn’t a mathematically defined number for this”. If you can remember your high school

maths, remember that it is conventional to say that $0/0$ doesn't have a proper answer: mathematicians would say that $0/0$ is *undefined*. R says that it's not a number:

```
0 / 0
```

```
## [1] NaN
```

Nevertheless, it's still treated as a "numeric" value. To oversimplify, `NaN` corresponds to cases where you asked a proper numerical question that genuinely has *no meaningful answer*.

- *Not available* (`NA`). `NA` indicates that the value that is "supposed" to be stored here is missing. To understand what this means, it helps to recognise that the `NA` value is something that you're most likely to see when analysing data from real world experiments. Sometimes you get equipment failures, or you lose some of the data, or whatever. The point is that some of the information that you were "expecting" to get from your study is just plain missing. Note the difference between `NA` and `NaN`. For `NaN`, we really do know what's supposed to be stored; it's just that it happens to correspond to something like $0/0$ that doesn't make any sense at all. In contrast, `NA` indicates that we actually don't know what was supposed to be there. The information is *missing*.
- *No value* (`NULL`). The `NULL` value takes this "absence" concept even further. It basically asserts that the variable genuinely has no value whatsoever. This is quite different to both `NaN` and `NA`. For `NaN` we actually know what the value is, because it's something insane like $0/0$. For `NA`, we believe that there is supposed to be a value "out there", but a dog ate our homework and so we don't quite know what it is. But for `NULL` we strongly believe that there is *no value at all*.

2.21.2 Assigning names to vector elements

One thing that is sometimes a little unsatisfying about the way that R prints out a vector is that the elements come out unlabelled. Here's what I mean. Suppose I've got data reporting the quarterly profits for some company. If I just create a no-frills vector, I have to rely on memory to know which element corresponds to which event. That is:

```
profit <- c( 3.1, 0.1, -1.4, 1.1 )
profit
```

```
## [1] 3.1 0.1 -1.4 1.1
```

You can probably guess that the first element corresponds to the first quarter, the second element to the second quarter, and so on, but that's only because I've told you the back story and because this happens to be a very simple example. In general, it can be quite difficult. This is where it can be helpful to assign `names` to each of the elements. Here's how you do it:

```
names(profit) <- c("Q1", "Q2", "Q3", "Q4")
profit
```

```
##   Q1   Q2   Q3   Q4
## 3.1  0.1 -1.4  1.1
```

This is a slightly odd looking command, admittedly, but it's not too difficult to follow. All we're doing is assigning a vector of labels (character strings) to `names(profit)`. You can always delete the names again by using the command `names(profit) <- NULL`. It's also worth noting that you don't have to do this as a two stage process. You can get the same result with this command:

```
profit <- c( "Q1" = 3.1, "Q2" = 0.1, "Q3" = -1.4, "Q4" = 1.1 )
profit
```

```
##   Q1   Q2   Q3   Q4
## 3.1  0.1 -1.4  1.1
```

The important things to notice are that (a) this does make things much easier to read, but (b) the names at the top aren't the "real" data. The *value* of `profit[1]` is still 3.1; all I've done is added a *name* to `profit[1]` as well. Nevertheless, names aren't purely cosmetic, since R allows you to pull out particular elements of the vector by referring to their names:

```
profit["Q1"]
```

```
## Q1
## 3.1
```

And if I ever need to pull out the names themselves, then I just type `names(profit)`.

2.21.3 Variable classes

As we've seen, R allows you to store different kinds of data. In particular, the variables we've defined so far have either been character data (text), numeric

data, or logical data.⁴⁵ It's important that we remember what kind of information each variable stores (and even more important that R remembers) since different kinds of variables allow you to do different things to them. For instance, if your variables have numerical information in them, then it's okay to multiply them together:

```
x <- 5    # x is numeric
y <- 4    # y is numeric
x * y
```

```
## [1] 20
```

But if they contain character data, multiplication makes no sense whatsoever, and R will complain if you try to do it:

```
x <- "apples"    # x is character
y <- "oranges"   # y is character
x * y
```

```
## Error in x * y: non-numeric argument to binary operator
```

Even R is smart enough to know you can't multiply "apples" by "oranges". It knows this because the quote marks are indicators that the variable is supposed to be treated as text, not as a number.

This is quite useful, but notice that it means that R makes a big distinction between 5 and "5". Without quote marks, R treats 5 as the number five, and will allow you to do calculations with it. With the quote marks, R treats "5" as the textual character five, and doesn't recognise it as a number any more than it recognises "p" or "five" as numbers. As a consequence, there's a big difference between typing `x <- 5` and typing `x <- "5"`. In the former, we're storing the number 5; in the latter, we're storing the character "5". Thus, if we try to do multiplication with the character versions, R gets stroppy:

```
x <- "5"    # x is character
y <- "4"    # y is character
x * y
```

```
## Error in x * y: non-numeric argument to binary operator
```

Okay, let's suppose that I've forgotten what kind of data I stored in the variable `x` (which happens depressingly often). R provides a function that will let us

⁴⁵Or functions. But let's ignore functions for the moment.

find out. Or, more precisely, it provides *three* functions: `class()`, `mode()` and `typeof()`. Why the heck does it provide three functions, you might be wondering? Basically, because R actually keeps track of three different kinds of information about a variable:

1. The **`class`** of a variable is a “high level” classification, and it captures psychologically (or statistically) meaningful distinctions. For instance “2011-09-12” and “my birthday” are both text strings, but there’s an important difference between the two: one of them is a date. So it would be nice if we could get R to recognise that “2011-09-12” is a date, and allow us to do things like add or subtract from it. The class of a variable is what R uses to keep track of things like that. Because the class of a variable is critical for determining what R can or can’t do with it, the `class()` function is very handy.
2. The **`mode`** of a variable refers to the format of the information that the variable stores. It tells you whether R has stored text data or numeric data, for instance, which is kind of useful, but it only makes these “simple” distinctions. It can be useful to know about, but it’s not the main thing we care about. So I’m not going to use the `mode()` function very much.⁴⁶
3. The **`type`** of a variable is a very low level classification. We won’t use it in this book, but (for those of you that care about these details) this is where you can see the distinction between integer data, double precision numeric, etc. Almost none of you actually will care about this, so I’m not even going to bother demonstrating the `typeof()` function.

For purposes, it’s the `class()` of the variable that we care most about. Later on, I’ll talk a bit about how you can convince R to “coerce” a variable to change from one class to another (Section ??). That’s a useful skill for real world data analysis, but it’s not something that we need right now. In the meantime, the following examples illustrate the use of the `class()` function:

```
x <- "hello world"      # x is text
class(x)
```

```
## [1] "character"

x <- TRUE      # x is logical
class(x)
```

```
## [1] "logical"
```

⁴⁶Actually, I don’t think I *ever* use this in practice. I don’t know why I bother to talk about it in the book anymore.

```
x <- 100      # x is a number
class(x)
```

```
## [1] "numeric"
```

Exciting, no?

2.22 Factors

Okay, it's time to start introducing some of the data types that are somewhat more specific to statistics. If you remember back to Chapter 1.6, when we assign numbers to possible outcomes, these numbers can mean quite different things depending on what kind of variable we are attempting to measure. In particular, we commonly make the distinction between *nominal*, *ordinal*, *interval* and *ratio* scale data. How do we capture this distinction in R? Currently, we only seem to have a single numeric data type. That's probably not going to be enough, is it?

A little thought suggests that the numeric variable class in R is perfectly suited for capturing ratio scale data. For instance, if I were to measure response time (RT) for five different events, I could store the data in R like this:

```
RT <- c(342, 401, 590, 391, 554)
```

where the data here are measured in milliseconds, as is conventional in the psychological literature. It's perfectly sensible to talk about "twice the response time", $2 \times RT$, or the "response time plus 1 second", $RT + 1000$, and so both of the following are perfectly reasonable things for R to do:

```
2 * RT
```

```
## [1] 684 802 1180 782 1108
```

```
RT + 1000
```

```
## [1] 1342 1401 1590 1391 1554
```

And to a lesser extent, the "numeric" class is okay for interval scale data, as long as we remember that multiplication and division aren't terribly interesting for these sorts of variables. That is, if my IQ score is 110 and yours is 120, it's perfectly okay to say that you're 10 IQ points smarter than me⁴⁷, but it's not

⁴⁷Taking all the usual caveats that attach to IQ measurement as a given, of course.

okay to say that I'm only 92% as smart as you are, because intelligence doesn't have a natural zero.⁴⁸ We might even be willing to tolerate the use of numeric variables to represent ordinal scale variables, such as those that you typically get when you ask people to rank order items (e.g., like we do in Australian elections), though as we will see R actually has a built in tool for representing ordinal data (see Section ??) However, when it comes to nominal scale data, it becomes completely unacceptable, because almost all of the "usual" rules for what you're allowed to do with numbers don't apply to nominal scale data. It is for this reason that R has *factors*.

2.22.1 Introducing factors

Suppose, I was doing a study in which people could belong to one of three different treatment conditions. Each group of people were asked to complete the same task, but each group received different instructions. Not surprisingly, I might want to have a variable that keeps track of what group people were in. So I could type in something like this

```
group <- c(1,1,1,2,2,2,3,3,3)
```

so that `group[i]` contains the group membership of the `i`-th person in my study. Clearly, this is numeric data, but equally obviously this is a nominal scale variable. There's no sense in which "group 1" plus "group 2" equals "group 3", but nevertheless if I try to do that, R won't stop me because it doesn't know any better:

```
group + 2
```

```
## [1] 3 3 3 4 4 4 5 5 5
```

Apparently R seems to think that it's allowed to invent "group 4" and "group 5", even though they didn't actually exist. Unfortunately, R is too stupid to know any better: it thinks that 3 is an ordinary number in this context, so it sees no problem in calculating `3 + 2`. But since we're not that stupid, we'd like to stop R from doing this. We can do so by instructing R to treat `group` as a factor. This is easy to do using the `as.factor()` function.⁴⁹

⁴⁸Or, more precisely, we don't know how to measure it. Arguably, a rock has zero intelligence. But it doesn't make sense to say that the IQ of a rock is 0 in the same way that we can say that the average human has an IQ of 100. And without knowing what the IQ value is that corresponds to a literal absence of any capacity to think, reason or learn, then we really can't multiply or divide IQ scores and expect a meaningful answer.

⁴⁹Once again, this is an example of *coercing* a variable from one class to another. I'll talk about coercion in more detail in Section ??.

```
group <- as.factor(group)
group
```

```
## [1] 1 1 1 2 2 2 3 3 3
## Levels: 1 2 3
```

It looks more or less the same as before (though it's not immediately obvious what all that `Levels` rubbish is about), but if we ask R to tell us what the class of the `group` variable is now, it's clear that it has done what we asked:

```
class(group)
```

```
## [1] "factor"
```

Neat. Better yet, now that I've converted `group` to a factor, look what happens when I try to add 2 to it:

```
group + 2
```

```
## Warning in Ops.factor(group, 2): '+' not meaningful for factors
## [1] NA NA NA NA NA NA NA NA NA
```

This time even R is smart enough to know that I'm being an idiot, so it tells me off and then produces a vector of missing values. (i.e., `NA`: see Section 7.8.1).

2.22.2 Labelling the factor levels

I have a confession to make. My memory is not infinite in capacity; and it seems to be getting worse as I get older. So it kind of annoys me when I get data sets where there's a nominal scale variable called `gender`, with two levels corresponding to males and females. But when I go to print out the variable I get something like this:

```
gender
```

```
## [1] 1 1 1 1 1 2 2 2 2
## Levels: 1 2
```

Okaaaay. That's not helpful at all, and it makes me very sad. Which number corresponds to the males and which one corresponds to the females? Wouldn't it be nice if R could actually keep track of this? It's way too hard to remember which number corresponds to which gender. To fix this problem what we need to do is assign meaningful labels to the different *levels* of each factor. We can do that like this:

```
levels(group) <- c("group 1", "group 2", "group 3")
print(group)

## [1] group 1 group 1 group 1 group 2 group 2 group 2 group 3 group 3 group 3
## Levels: group 1 group 2 group 3

levels(gender) <- c("male", "female")
print(gender)

## [1] male   male   male   male   male   female female female female
## Levels: male female
```

That's much easier on the eye.

2.22.3 Moving on...

Factors are very useful things, and we'll use them a lot in this book: they're *the* main way to represent a nominal scale variable. And there are lots of nominal scale variables out there. I'll talk more about factors in Section ??, but for now you know enough to be able to get started.

2.23 Data frames

It's now time to go back and deal with the somewhat confusing thing that happened in Section 2.20.3 when we tried to open up a CSV file. Apparently we succeeded in loading the data, but it came to us in a very odd looking format. At the time, I told you that this was a *data frame*. Now I'd better explain what that means.

2.23.1 Introducing data frames

In order to understand why R has created this funny thing called a data frame, it helps to try to see what problem it solves. So let's go back to the little scenario that I used when introducing factors in Section 2.22. In that section I recorded

the `group` and `gender` for all 9 participants in my study. Let's also suppose I recorded their ages and their `score` on "Dan's Terribly Exciting Psychological Test":

```
age <- c(17, 19, 21, 37, 18, 19, 47, 18, 19)
score <- c(12, 10, 11, 15, 16, 14, 25, 21, 29)
```

Assuming no other variables are in the workspace, if I type `who()` I get this:

```
who()
```

##	-- Name --	-- Class --	-- Size --
##	age	numeric	9
##	any.sales.this.month	logical	12
##	berkeley	data.frame	39 x 3
##	berkeley.small	data.frame	46 x 2
##	coef	numeric	2
##	days.per.month	numeric	12
##	february.sales	numeric	1
##	gender	factor	9
##	greeting	character	1
##	group	factor	9
##	is.the.Party.correct	logical	1
##	months	character	12
##	projecthome	character	1
##	revenue	numeric	1
##	royalty	numeric	1
##	sales	numeric	1
##	sales.by.month	numeric	12
##	score	numeric	9
##	simpson	matrix	6 x 5
##	stock.levels	character	12
##	xlu	numeric	1

So there are four variables in the workspace, `age`, `gender`, `group` and `score`. And it just so happens that all four of them are the same size (i.e., they're all vectors with 9 elements). Aaaand it just so happens that `age[1]` corresponds to the age of the first person, and `gender[1]` is the gender of that very same person, etc. In other words, you and I both know that all four of these variables correspond to the *same* data set, and all four of them are organised in exactly the same way.

However, R *doesn't* know this! As far as it's concerned, there's no reason why the `age` variable has to be the same length as the `gender` variable; and there's no particular reason to think that `age[1]` has any special relationship to `gender[1]`

any more than it has a special relationship to `gender`[4]. In other words, when we store everything in separate variables like this, R doesn't know anything about the relationships between things. It doesn't even really know that these variables actually refer to a proper data set. The data frame fixes this: if we store our variables inside a data frame, we're telling R to treat these variables as a single, fairly coherent data set.

To see how they do this, let's create one. So how do we create a data frame? One way we've already seen: if we import our data from a CSV file, R will store it as a data frame. A second way is to create it directly from some existing variables using the `data.frame()` function. All you have to do is type a list of variables that you want to include in the data frame. The output of a `data.frame()` command is, well, a data frame. So, if I want to store all four variables from my experiment in a data frame called `expt` I can do so like this:

```
expt <- data.frame ( age, gender, group, score )
expt
```

```
##   age gender  group score
## 1  17    male group 1    12
## 2  19    male group 1    10
## 3  21    male group 1    11
## 4  37    male group 2    15
## 5  18    male group 2    16
## 6  19  female group 2    14
## 7  47  female group 3    25
## 8  18  female group 3    21
## 9  19  female group 3    29
```

Note that `expt` is a completely self-contained variable. Once you've created it, it no longer depends on the original variables from which it was constructed. That is, if we make changes to the original `age` variable, it will *not* lead to any changes to the `age` data stored in `expt`.

2.23.2 Pulling out the contents of the data frame using \$

At this point, our workspace contains only the one variable, a data frame called `expt`. But as we can see when we told R to print the variable out, this data frame contains 4 variables, each of which has 9 observations. So how do we get this information out again? After all, there's no point in storing information if you don't use it, and there's no way to use information if you can't access it. So let's talk a bit about how to pull information out of a data frame.

The first thing we might want to do is pull out one of our stored variables, let's say `score`. One thing you might try to do is ignore the fact that `score` is locked

up inside the `expt` data frame. For instance, you might try to print it out like this:

```
score
```

```
## Error in eval(expr, envir, enclos): object 'score' not found
```

This doesn't work, because R doesn't go "peeking" inside the data frame unless you explicitly tell it to do so. There's actually a very good reason for this, which I'll explain in a moment, but for now let's just assume R knows what it's doing. How do we tell R to look inside the data frame? As is always the case with R there are several ways. The simplest way is to use the `$` operator to extract the variable you're interested in, like this:

```
expt$score
```

```
## [1] 12 10 11 15 16 14 25 21 29
```

2.23.3 Getting information about a data frame

One problem that sometimes comes up in practice is that you forget what you called all your variables. Normally you might try to type `objects()` or `who()`, but neither of those commands will tell you what the names are for those variables inside a data frame! One way is to ask R to tell you what the *names* of all the variables stored in the data frame are, which you can do using the `names()` function:

```
names(expt)
```

```
## [1] "age"     "gender"   "group"    "score"
```

An alternative method is to use the `who()` function, as long as you tell it to look at the variables inside data frames. If you set `expand = TRUE` then it will not only list the variables in the workspace, but it will "expand" any data frames that you've got in the workspace, so that you can see what they look like. That is:

```
who(expand = TRUE)
```

```
##  -- Name --          -- Class --  -- Size --
##  any.sales.this.month logical      12
##  berkeley            data.frame  39 x 3
```

```

##   $women.apply      numeric    39
##   $total.admit     numeric    39
##   $number.apply     numeric    39
##   berkeley.small   data.frame 46 x 2
##   $women.apply      numeric    46
##   $total.admit     numeric    46
##   coef              numeric    2
##   days.per.month   numeric    12
##   expt              data.frame 9 x 4
##   $age               numeric    9
##   $gender             factor    9
##   $group              factor    9
##   $score              numeric    9
##   february.sales    numeric    1
##   greeting            character 1
##   is.the.Party.correct logical  1
##   months              character 12
##   projecthome         character 1
##   revenue              numeric    1
##   royalty              numeric    1
##   sales               numeric    1
##   sales.by.month    numeric    12
##   simpson             matrix    6 x 5
##   stock.levels        character 12
##   xlu                 numeric    1

```

or, since `expand` is the first argument in the `who()` function you can just type `who(TRUE)`. I'll do that a lot in this book.

2.23.4 Looking for more on data frames?

There's a lot more that can be said about data frames: they're fairly complicated beasts, and the longer you use R the more important it is to make sure you really understand them. We'll talk a lot more about them in Chapter ??.

2.24 Lists

The next kind of data I want to mention are *lists*. Lists are an extremely fundamental data structure in R, and as you start making the transition from a novice to a savvy R user you will use lists all the time. I don't use lists very often in this book – not directly – but most of the advanced data structures in R are built from lists (e.g., data frames are actually a specific type of list). Because lists are so important to how R stores things, it's useful to have a basic

understanding of them. Okay, so what is a list, exactly? Like data frames, lists are just “collections of variables.” However, unlike data frames – which are basically supposed to look like a nice “rectangular” table of data – there are no constraints on what kinds of variables we include, and no requirement that the variables have any particular relationship to one another. In order to understand what this actually *means*, the best thing to do is create a list, which we can do using the `list()` function. If I type this as my command:

```
Dan <- list( age = 34,
             nerd = TRUE,
             parents = c("Joe", "Liz")
           )
```

R creates a new list variable called `Dan`, which is a bundle of three different variables: `age`, `nerd` and `parents`. Notice, that the `parents` variable is longer than the others. This is perfectly acceptable for a list, but it wouldn’t be for a data frame. If we now print out the variable, you can see the way that R stores the list:

```
print( Dan )

## $age
## [1] 34
##
## $nerd
## [1] TRUE
##
## $parents
## [1] "Joe" "Liz"
```

As you might have guessed from those `$` symbols everywhere, the variables are stored in exactly the same way that they are for a data frame (again, this is not surprising: data frames *are* a type of list). So you will (I hope) be entirely unsurprised and probably quite bored when I tell you that you can extract the variables from the list using the `$` operator, like so:

```
Dan$nerd
```

```
## [1] TRUE
```

If you need to add new entries to the list, the easiest way to do so is to again use `$`, as the following example illustrates. If I type a command like this

```
Dan$children <- "Alex"
```

then R creates a new entry to the end of the list called `children`, and assigns it a value of "Alex". If I were now to `print()` this list out, you'd see a new entry at the bottom of the printout. Finally, it's actually possible for lists to contain other lists, so it's quite possible that I would end up using a command like `Dan$children$age` to find out how old my son is. Or I could try to remember it myself I suppose.

2.25 Formulas

The last kind of variable that I want to introduce before finally being able to start talking about statistics is the *formula*. Formulas were originally introduced into R as a convenient way to specify a particular type of statistical model (see Chapter 8) but they're such handy things that they've spread. Formulas are now used in a lot of different contexts, so it makes sense to introduce them early.

Stated simply, a formula object is a variable, but it's a special type of variable that specifies a relationship between other variables. A formula is specified using the "tilde operator" `~`. A very simple example of a formula is shown below.⁵⁰

```
formula1 <- out ~ pred
formula1

## out ~ pred
```

The *precise* meaning of this formula depends on exactly what you want to do with it, but in broad terms it means "the `out` (outcome) variable, analysed in terms of the `pred` (predictor) variable". That said, although the simplest and most common form of a formula uses the "one variable on the left, one variable on the right" format, there are others. For instance, the following examples are all reasonably common

```
formula2 <- out ~ pred1 + pred2      # more than one variable on the right
formula3 <- out ~ pred1 * pred2      # different relationship between predictors
formula4 <- ~ var1 + var2            # a 'one-sided' formula
```

and there are many more variants besides. Formulas are pretty flexible things, and so different functions will make use of different formats, depending on what the function is intended to do.

⁵⁰Note that, when I write out the formula, R doesn't check to see if the `out` and `pred` variables actually exist: it's only later on when you try to use the formula for something that this happens.

2.26 Generic functions

There's one really important thing that I omitted when I discussed functions earlier on in Section 2.7, and that's the concept of a *generic function*. The two most notable examples that you'll see in the next few chapters are `summary()` and `plot()`, although you've already seen an example of one working behind the scenes, and that's the `print()` function. The thing that makes generics different from the other functions is that their behaviour changes, often quite dramatically, depending on the `class()` of the input you give it. The easiest way to explain the concept is with an example. With that in mind, let's take a closer look at what the `print()` function actually does. I'll do this by creating a formula, and printing it out in a few different ways. First, let's stick with what we know:

```
my.formula <- blah ~ blah.blah      # create a variable of class "formula"
print( my.formula )                  # print it out using the generic print() function

## blah ~ blah.blah
```

So far, there's nothing very surprising here. But there's actually a lot going on behind the scenes here. When I type `print(my.formula)`, what actually happens is the `print()` function checks the class of the `my.formula` variable. When the function discovers that the variable it's been given is a formula, it goes looking for a function called `print.formula()`, and then delegates the whole business of printing out the variable to the `print.formula()` function.⁵¹ For what it's worth, the name for a “dedicated” function like `print.formula()` that exists only to be a special case of a generic function like `print()` is a *method*, and the name for the process in which the generic function passes off all the hard work onto a method is called *method dispatch*. You won't need to understand the details at all for this book, but you do need to know the gist of it; if only because a lot of the functions we'll use are actually generics. Anyway, to help expose a little more of the workings to you, let's bypass the `print()` function entirely and call the formula method directly:

```
print.formula( my.formula )          # print it out using the print.formula() method

## Appears to be deprecated
```

There's no difference in the output at all. But this shouldn't surprise you because it was actually the `print.formula()` method that was doing all the

⁵¹For readers with a programming background: what I'm describing is the very basics of how S3 methods work. However, you should be aware that R has two entirely distinct systems for doing object oriented programming, known as S3 and S4. Of the two, S3 is simpler and more informal, whereas S4 supports all the stuff that you might expect of a fully object oriented language. Most of the generics we'll run into in this book use the S3 system, which is convenient for me because I'm still trying to figure out S4.

hard work in the first place. The `print()` function itself is a lazy bastard that doesn't do anything other than select which of the methods is going to do the actual printing.

Okay, fair enough, but you might be wondering what would have happened if `print.formula()` didn't exist? That is, what happens if there isn't a specific method defined for the class of variable that you're using? In that case, the generic function passes off the hard work to a "default" method, whose name in this case would be `print.default()`. Let's see what happens if we bypass the `print()` formula, and try to print out `my.formula` using the `print.default()` function:

```
print.default( my.formula )      # print it out using the print.default() method

## blah ~ blah.blah
## attr(,"class")
## [1] "formula"
## attr(,".Environment")
## <environment: R_GlobalEnv>
```

Hm. You can kind of see that it is trying to print out the same formula, but there's a bunch of ugly low-level details that have also turned up on screen. This is because the `print.default()` method doesn't know anything about formulas, and doesn't know that it's supposed to be hiding the obnoxious internal gibberish that R produces sometimes.

At this stage, this is about as much as we need to know about generic functions and their methods. In fact, you can get through the entire book without learning any more about them than this, so it's probably a good idea to end this discussion here.

2.27 Getting help

The very last topic I want to mention in this chapter is where to go to find help. Obviously, I've tried to make this book as helpful as possible, but it's not even close to being a comprehensive guide, and there's thousands of things it doesn't cover. So where should you go for help?

2.27.1 How to read the help documentation

I have somewhat mixed feelings about the help documentation in R. On the plus side, there's a lot of it, and it's very thorough. On the minus side, there's a lot of it, and it's very thorough. There's so much help documentation that

it sometimes doesn't help, and most of it is written with an advanced user in mind. Often it feels like most of the help files work on the assumption that the reader already understands everything about R except for the specific topic that it's providing help for. What that means is that, once you've been using R for a long time and are beginning to get a feel for how to use it, the help documentation is awesome. These days, I find myself really liking the help files (most of them anyway). But when I first started using R I found it very dense.

To some extent, there's not much I can do to help you with this. You just have to work at it yourself; once you're moving away from being a pure beginner and are becoming a skilled user, you'll start finding the help documentation more and more helpful. In the meantime, I'll help as much as I can by trying to explain to you what you're looking at when you open a help file. To that end, let's look at the help documentation for the `load()` function. To do so, I type either of the following:

```
?load
help("load")
```

When I do that, R goes looking for the help file for the "load" topic. If it finds one, Rstudio takes it and displays it in the help panel. Alternatively, you can try a fuzzy search for a help topic

```
??load
help.search("load")
```

This will bring up a list of possible topics that you might want to follow up in. Regardless, at some point you'll find yourself looking at an actual help file. And when you do, you'll see there's a quite a lot of stuff written down there, and it comes in a pretty standardised format. So let's go through it slowly, using the "load" topic as our example. Firstly, at the very top we see this:

```
oad {base}
R Documentation
Reload Saved Datasets
Description
Reload datasets written with the function save.
```

Fairly straightforward. The next section describes how the function is used:

```
sage
```

In this instance, the usage section is actually pretty readable. It's telling you that there are two arguments to the `load()` function: the first one is called

`file`, and the second one is called `envir`. It's also telling you that there is a default value for the `envir` argument; so if the user doesn't specify what the value of `envir` should be, then R will assume that `envir = parent.frame()`. In contrast, the `file` argument has no default value at all, so the user must specify a value for it. So in one sense, this section is very straightforward.

The problem, of course, is that you don't know what the `parent.frame()` function actually does, so it's hard for you to know what the `envir = parent.frame()` bit is all about. What you could do is then go look up the help documents for the `parent.frame()` function (and sometimes that's actually a good idea), but often you'll find that the help documents for those functions are just as dense (if not more dense) than the help file that you're currently reading. As an alternative, my general approach when faced with something like this is to skim over it, see if I can make any sense of it. If so, great. If not, I find that the best thing to do is ignore it. In fact, the first time I read the help file for the `load()` function, I had no idea what any of the `envir` related stuff was about. But fortunately I didn't have to: the default setting here (i.e., `envir = parent.frame()`) is actually the thing you want in about 99% of cases, so it's safe to ignore it.

Basically, what I'm trying to say is: don't let the scary, incomprehensible parts of the help file intimidate you. Especially because there's often some parts of the help file that will make sense. Of course, I guarantee you that sometimes this strategy will lead you to make mistakes... often embarrassing mistakes. But it's still better than getting paralysed with fear.

So, let's continue on. The next part of the help documentation discusses each of the arguments, and what they're supposed to do:

arguments
`file`
 a (readable binary-mode) connection or a character string giving the name of the file to load (when tilde expansion is done).
`envir`
 the environment where the data should be loaded.
`verbose`
 should item names be printed during loading?

Okay, so what this is telling us is that the `file` argument needs to be a string (i.e., text data) which tells R the name of the file to load. It also seems to be hinting that there's other possibilities too (e.g., a "binary mode connection"), and you probably aren't quite sure what "tilde expansion" means⁵². But overall, the meaning is pretty clear.

⁵²It's extremely simple, by the way. We discussed it in Section 4.4, though I didn't call it by that name. Tilde expansion is the thing where R recognises that, in the context of specifying a file location, the tilde symbol `~` corresponds to the user home directory (e.g., `/Users/dan/`).

Turning to the `envir` argument, it's now a little clearer what the Usage section was babbling about. The `envir` argument specifies the name of an environment (see Section 4.3 if you've forgotten what environments are) into which R should place the variables when it loads the file. Almost always, this is a no-brainer: you want R to load the data into the same damn environment in which you're invoking the `load()` command. That is, if you're typing `load()` at the R prompt, then you want the data to be loaded into your workspace (i.e., the global environment). But if you're writing your own function that needs to load some data, you want the data to be loaded inside that function's private workspace. And in fact, that's exactly what the `parent.frame()` thing is all about. It's telling the `load()` function to send the data to the same place that the `load()` command itself was coming from. As it turns out, if we'd just ignored the `envir` bit we would have been totally safe. Which is nice to know.

Moving on, next up we get a detailed description of what the function actually does:

Details

`load` can load R objects saved in the current or any earlier format. It can read a compressed file (see `save`) directly from a file or from a suitable connection (including a call to `url`).

A not-open connection will be opened in mode “rb” and closed after use. Any connection other than a `gzfile` or `gzcon` connection will be wrapped in `gzcon` to allow compressed saves to be handled: note that this leaves the connection in an altered state (in particular, binary-only), and that it needs to be closed explicitly (it will not be garbage-collected).

Only R objects saved in the current format (used since R 1.4.0) can be read from a connection. If no input is available on a connection a warning will be given, but any input not in the current format will result in a error.

Loading from an earlier version will give a warning about the ‘magic number’: magic numbers 1971:1977 are from R < 0.99.0, and RD[ABX]1 from R 0.99.0 to R 1.3.1. These are all obsolete, and you are strongly recommended to re-save such files in a current format.

The `verbose` argument is mainly intended for debugging. If it is `TRUE`, then as objects from the file are loaded, their names will be printed to the console. If `verbose` is set to an integer value greater than one, additional names corresponding to attributes and other parts of individual objects will also be printed. Larger values will print names to a greater depth.

Objects can be saved with references to namespaces, usually as part of the environment of a function or formula. Such objects can be loaded even if the namespace is not available: it is replaced by a reference to the global environment with a warning. The warning identifies the first object with such a reference (but there may be more than one).

Then it tells you what the output value of the function is:

alue

A character vector of the names of objects created, invisibly.

This is usually a bit more interesting, but since the `load()` function is mainly used to load variables into the workspace rather than to return a value, it's no surprise that this doesn't do much or say much. Moving on, we sometimes see a few additional sections in the help file, which can be different depending on what the function is:

arning

Saved R objects are binary files, even those saved with `ascii = TRUE`, so ensure that they are transferred without conversion of end of line markers. `load` tries to detect such a conversion and gives an informative error message.

`load(<file>)` replaces all existing objects with the same names in the current environment (typically your workspace, `.GlobalEnv`) and hence potentially overwrites important data. It is considerably safer to use `envir =` to load into a different environment, or to `attach(file)` which `load()`s into a new entry in the search path.

Note

`file` can be a UTF-8-encoded filepath that cannot be translated to the current locale.

Yeah, yeah. Warning, warning, blah blah blah. Towards the bottom of the help file, we see something like this, which suggests a bunch of related topics that you might want to look at. These can be quite helpful:

ee Also

`save`, `download.file`; further `attach` as wrapper for `load()`.

For other interfaces to the underlying serialization format, see `unserialize` and `readRDS`.

Finally, it gives you some examples of how to use the function(s) that the help file describes. These are supposed to be proper R commands, meaning that you should be able to type them into the console yourself and they'll actually work. Sometimes it can be quite helpful to try the examples yourself. Anyway, here they are for the “`load`” help file:

xamples

As you can see, they're pretty dense, and not at all obvious to the novice user. However, they do provide good examples of the various different things that you can do with the `load()` function, so it's not a bad idea to have a look at them, and to try not to find them too intimidating.

2.27.2 Other resources

- The Rseek website (www.rseek.org). One thing that I really find annoying about the R help documentation is that it's hard to search properly. When coupled with the fact that the documentation is dense and highly technical, it's often a better idea to search or ask online for answers to your questions. With that in mind, the Rseek website is great: it's an R specific search engine. I find it really useful, and it's almost always my first port of call when I'm looking around.
- The R-help mailing list (see <http://www.r-project.org/mail.html> for details). This is the official R help mailing list. It can be very helpful, but it's *very* important that you do your homework before posting a question. The list gets a lot of traffic. While the people on the list try as hard as they can to answer questions, they do so for free, and you *really* don't want to know how much money they could charge on an hourly rate if they wanted to apply market rates. In short, they are doing you a favour, so be polite. Don't waste their time asking questions that can be easily answered by a quick search on Rseek (it's rude), make sure your question is clear, and all of the relevant information is included. In short, read the posting guidelines carefully (<http://www.r-project.org/posting-guide.html>), and make use of the `help.request()` function that R provides to check that you're actually doing what you're expected.

2.28 Summary

This chapter continued where Chapter 2 left off. The focus was still primarily on introducing basic R concepts, but this time at least you can see how those concepts are related to data analysis:

- Installing, loading and updating packages. Knowing how to extend the functionality of R by installing and using packages is critical to becoming an effective R user
- Getting around. Section 2.18 talked about how to manage your workspace and how to keep it tidy. Similarly, Section 2.19 talked about how to get R to interact with the rest of the file system.
- Loading and saving data. Finally, we encountered actual data files. Loading and saving data is obviously a crucial skill, one we discussed in Section 2.20.
- Useful things to know about variables. In particular, we talked about special values, element names and classes.
- More complex types of variables. R has a number of important variable types that will be useful when analysing real data. I talked about factors in Section 2.22, data frames in Section 2.23, lists in Section 2.24 and formulas in Section 2.25.

- Generic functions. How is it that some function seem to be able to do lots of different things? Section 2.26 tells you how.
- Getting help. Assuming that you're not looking for counselling, Section 2.27 covers several possibilities. If you are looking for counselling, well, this book really can't help you there. Sorry.

Taken together, Chapters 2 and 2.15 provide enough of a background that you can finally get started doing some statistics! Yes, there's a lot more R concepts that you ought to know (and we'll talk about some of them in Chapters?? and??), but I think that we've talked quite enough about programming for the moment. It's time to see how your experience with programming can be used to do some data analysis...

Chapter 3

Descriptive statistics

Text by Navarro (2018)

3.1 Videos

Video: Descriptive Stats: Central Tendency

Video: Descriptive Stats: Variability

Video: Descriptive Stats: Skewness and Kurtosis

3.2 Introduction

Any time that you get a new data set to look at, one of the first tasks that you have to do is find ways of summarising the data in a compact, easily understood fashion. This is what *descriptive statistics* (as opposed to inferential statistics) is all about. In fact, to many people the term “statistics” is synonymous with descriptive statistics. It is this topic that we’ll consider in this chapter, but before going into any details, let’s take a moment to get a sense of why we need descriptive statistics. To do this, let’s load the `aflsmall.Rdata` file, and use the `who()` function in the `lsr` package to see what variables are stored in the file:

```
load( "./data/aflsmall.Rdata" )
library(lsr)
who()

##      -- Name --          -- Class --    -- Size --

```

```

##   afl.finalists      factor    400
##   afl.margins        numeric   176
##   any.sales.this.month logical   12
##   berkeley           data.frame 39 x 3
##   berkeley.small     data.frame 46 x 2
##   coef                numeric    2
##   Dan                 list      4
##   days.per.month     numeric   12
##   expt                data.frame 9 x 4
##   february.sales     numeric    1
##   formula1            formula
##   formula2            formula
##   formula3            formula
##   formula4            formula
##   greeting            character  1
##   is.the.Party.correct logical  1
##   months              character 12
##   my.formula          formula
##   projecthome         character  1
##   revenue              numeric    1
##   royalty              numeric    1
##   sales                numeric    1
##   sales.by.month      numeric   12
##   simpson             matrix    6 x 5
##   stock.levels         character 12
##   xlu                  numeric    1

```

There are two variables here, `afl.finalists` and `afl.margins`. We'll focus a bit on these two variables in this chapter, so I'd better tell you what they are. Unlike most of data sets in this book, these are actually real data, relating to the Australian Football League (AFL)¹. The `afl.margins` variable contains the winning margin (number of points) for all 176 home and away games played during the 2010 season. The `afl.finalists` variable contains the names of all 400 teams that played in all 200 finals matches played during the period 1987 to 2010. Let's have a look at the `afl.margins` variable:

```
print(afl.margins)
```

```

## [1] 56 31 56 8 32 14 36 56 19 1 3 104 43 44 72 9 28
## [18] 25 27 55 20 16 16 7 23 40 48 64 22 55 95 15 49 52
## [35] 50 10 65 12 39 36 3 26 23 20 43 108 53 38 4 8 3
## [52] 13 66 67 50 61 36 38 29 9 81 3 26 12 36 37 70 1
## [69] 35 12 50 35 9 54 47 8 47 2 29 61 38 41 23 24 1

```

¹Note for non-Australians: the AFL is an Australian rules football competition. You don't need to know anything about Australian rules in order to follow this section.

```
## [86]  9 11 10 29 47 71 38 49 65 18 0 16 9 19 36 60 24
## [103] 25 44 55 3 57 83 84 35 4 35 26 22 2 14 19 30 19
## [120] 68 11 75 48 32 36 39 50 11 0 63 82 26 3 82 73 19
## [137] 33 48 8 10 53 20 71 75 76 54 44 5 22 94 29 8 98
## [154] 9 89 1 101 7 21 52 42 21 116 3 44 29 27 16 6 44
## [171] 3 28 38 29 10 10
```

This output doesn't make it easy to get a sense of what the data are actually saying. Just "looking at the data" isn't a terribly effective way of understanding data. In order to get some idea about what's going on, we need to calculate some descriptive statistics (this chapter) and draw some nice pictures (Chapter 3.9). Since the descriptive statistics are the easier of the two topics, I'll start with those, but nevertheless I'll show you a histogram of the `afl.margins` data, since it should help you get a sense of what the data we're trying to describe actually look like. But for what it's worth, this histogram – which is shown in Figure 3.1 – was generated using the `hist()` function. We'll talk a lot more about how to draw histograms in Section 3.9.2. For now, it's enough to look at the histogram and note that it provides a fairly interpretable representation of the `afl.margins` data.

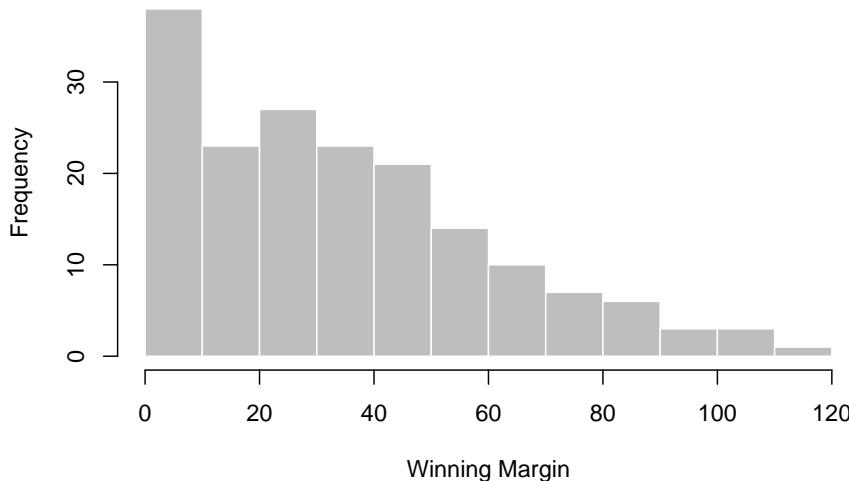


Figure 3.1: A histogram of the AFL 2010 winning margin data (the `afl.margins` variable). As you might expect, the larger the margin the less frequently you tend to see it.

3.3 Measures of central tendency

Drawing pictures of the data, as I did in Figure 3.1 is an excellent way to convey the “gist” of what the data is trying to tell you, it’s often extremely useful to try to condense the data into a few simple “summary” statistics. In most situations, the first thing that you’ll want to calculate is a measure of *central tendency*. That is, you’d like to know something about the “average” or “middle” of your data lies. The two most commonly used measures are the mean, median and mode; occasionally people will also report a trimmed mean. I’ll explain each of these in turn, and then discuss when each of them is useful.

3.3.1 The mean

The *mean* of a set of observations is just a normal, old-fashioned average: add all of the values up, and then divide by the total number of values. The first five AFL margins were 56, 31, 56, 8 and 32, so the mean of these observations is just:

$$\frac{56 + 31 + 56 + 8 + 32}{5} = \frac{183}{5} = 36.60$$

Of course, this definition of the mean isn’t news to anyone: averages (i.e., means) are used so often in everyday life that this is pretty familiar stuff. However, since the concept of a mean is something that everyone already understands, I’ll use this as an excuse to start introducing some of the mathematical notation that statisticians use to describe this calculation, and talk about how the calculations would be done in R.

The first piece of notation to introduce is N , which we’ll use to refer to the number of observations that we’re averaging (in this case $N = 5$). Next, we need to attach a label to the observations themselves. It’s traditional to use X for this, and to use subscripts to indicate which observation we’re actually talking about. That is, we’ll use X_1 to refer to the first observation, X_2 to refer to the second observation, and so on, all the way up to X_N for the last one. Or, to say the same thing in a slightly more abstract way, we use X_i to refer to the i -th observation. Just to make sure we’re clear on the notation, the following table lists the 5 observations in the `afl.margins` variable, along with the mathematical symbol used to refer to it, and the actual value that the observation corresponds to:

the observation	its symbol	the observed value
winning margin, game 1	$\$X_1\$$	56 points
winning margin, game 2	$\$X_2\$$	31 points
winning margin, game 3	$\$X_3\$$	56 points
winning margin, game 4	$\$X_4\$$	8 points
winning margin, game 5	$\$X_5\$$	32 points

Okay, now let's try to write a formula for the mean. By tradition, we use \bar{X} as the notation for the mean. So the calculation for the mean could be expressed using the following formula:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_{N-1} + X_N}{N}$$

This formula is entirely correct, but it's terribly long, so we make use of the **summation symbol** Σ to shorten it.² If I want to add up the first five observations, I could write out the sum the long way, $X_1 + X_2 + X_3 + X_4 + X_5$ or I could use the summation symbol to shorten it to this:

$$\sum_{i=1}^5 X_i$$

Taken literally, this could be read as “the sum, taken over all i values from 1 to 5, of the value X_i ”. But basically, what it means is “add up the first five observations”. In any case, we can use this notation to write out the formula for the mean, which looks like this:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

In all honesty, I can't imagine that all this mathematical notation helps clarify the concept of the mean at all. In fact, it's really just a fancy way of writing out the same thing I said in words: add all the values up, and then divide by the total number of items. However, that's not really the reason I went into all that detail. My goal was to try to make sure that everyone reading this book is clear on the notation that we'll be using throughout the book: \bar{X} for the mean, Σ for the idea of summation, X_i for the i th observation, and N for the total number of observations. We're going to be re-using these symbols a fair bit, so it's important that you understand them well enough to be able to “read” the equations, and to be able to see that it's just saying “add up lots of things and then divide by another thing”.

3.3.2 Calculating the mean in R

Okay that's the maths, how do we get the magic computing box to do the work for us? If you really wanted to, you could do this calculation directly in R. For the first 5 AFL scores, do this just by typing it in as if R were a calculator...

²The choice to use Σ to denote summation isn't arbitrary: it's the Greek upper case letter sigma, which is the analogue of the letter S in that alphabet. Similarly, there's an equivalent symbol used to denote the multiplication of lots of numbers: because multiplications are also called “products”, we use the Π symbol for this; the Greek upper case pi, which is the analogue of the letter P.

```
(56 + 31 + 56 + 8 + 32) / 5
```

```
## [1] 36.6
```

... in which case R outputs the answer 36.6, just as if it were a calculator. However, that's not the only way to do the calculations, and when the number of observations starts to become large, it's easily the most tedious. Besides, in almost every real world scenario, you've already got the actual numbers stored in a variable of some kind, just like we have with the `afl.margins` variable. Under those circumstances, what you want is a function that will just add up all the values stored in a numeric vector. That's what the `sum()` function does. If we want to add up all 176 winning margins in the data set, we can do so using the following command:³

```
sum( afl.margins )
```

```
## [1] 6213
```

If we only want the sum of the first five observations, then we can use square brackets to pull out only the first five elements of the vector. So the command would now be:

```
sum( afl.margins[1:5] )
```

```
## [1] 183
```

To calculate the mean, we now tell R to divide the output of this summation by five, so the command that we need to type now becomes the following:

```
sum( afl.margins[1:5] ) / 5
```

```
## [1] 36.6
```

Although it's pretty easy to calculate the mean using the `sum()` function, we can do it in an even easier way, since R also provides us with the `mean()` function. To calculate the mean for all 176 games, we would use the following command:

³Note that, just as we saw with the combine function `c()` and the remove function `rm()`, the `sum()` function has unnamed arguments. I'll talk about unnamed arguments later in Section ??, but for now let's just ignore this detail.

```
mean( x = afl.margins )
```

```
## [1] 35.30114
```

However, since `x` is the first argument to the function, I could have omitted the argument name. In any case, just to show you that there's nothing funny going on, here's what we would do to calculate the mean for the first five observations:

```
mean( afl.margins[1:5] )
```

```
## [1] 36.6
```

As you can see, this gives exactly the same answers as the previous calculations.

3.3.3 The median

The second measure of central tendency that people use a lot is the *median*, and it's even easier to describe than the mean. The median of a set of observations is just the middle value. As before let's imagine we were interested only in the first 5 AFL winning margins: 56, 31, 56, 8 and 32. To figure out the median, we sort these numbers into ascending order:

8, 31, **32**, 56, 56

From inspection, it's obvious that the median value of these 5 observations is 32, since that's the middle one in the sorted list (I've put it in bold to make it even more obvious). Easy stuff. But what should we do if we were interested in the first 6 games rather than the first 5? Since the sixth game in the season had a winning margin of 14 points, our sorted list is now

8, 14, **31**, **32**, 56, 56

and there are *two* middle numbers, 31 and 32. The median is defined as the average of those two numbers, which is of course 31.5. As before, it's very tedious to do this by hand when you've got lots of numbers. To illustrate this, here's what happens when you use R to sort all 176 winning margins. First, I'll use the `sort()` function (discussed in Chapter ??) to display the winning margins in increasing numerical order:

```
sort( x = afl.margins )
```

```
##   [1]  0  0  1  1  1  1  2  2  3  3  3  3  3  3  3  4
##  [18]  4  5  6  7  7  8  8  8  8  9  9  9  9  9  9  10
```

```
## [35] 10 10 10 10 11 11 11 12 12 12 13 14 14 15 16 16 16
## [52] 16 18 19 19 19 19 19 20 20 20 21 21 22 22 22 23 23
## [69] 23 24 24 25 25 26 26 26 26 27 27 28 28 29 29 29 29
## [86] 29 29 30 31 32 32 33 35 35 35 35 36 36 36 36 36 36
## [103] 37 38 38 38 38 39 39 40 41 42 43 43 44 44 44 44
## [120] 44 47 47 47 48 48 48 49 49 50 50 50 50 52 52 53 53
## [137] 54 54 55 55 55 56 56 56 57 60 61 61 63 64 65 65 66
## [154] 67 68 70 71 71 72 73 75 75 76 81 82 82 83 84 89 94
## [171] 95 98 101 104 108 116
```

The middle values are 30 and 31, so the median winning margin for 2010 was 30.5 points. In real life, of course, no-one actually calculates the median by sorting the data and then looking for the middle value. In real life, we use the `median` command:

```
median( x = afl.margins )
```

```
## [1] 30.5
```

which outputs the median value of 30.5.

3.3.4 Mean or median? What's the difference?

Knowing how to calculate means and medians is only a part of the story. You also need to understand what each one is saying about the data, and what that implies for when you should use each one. This is illustrated in Figure 3.2 the mean is kind of like the “centre of gravity” of the data set, whereas the median is the “middle value” in the data. What this implies, as far as which one you should use, depends a little on what type of data you’ve got and what you’re trying to achieve. As a rough guide:

- If your data are nominal scale, you probably shouldn’t be using either the mean or the median. Both the mean and the median rely on the idea that the numbers assigned to values are meaningful. If the numbering scheme is arbitrary, then it’s probably best to use the mode (Section 3.3.7) instead.
- If your data are ordinal scale, you’re more likely to want to use the median than the mean. The median only makes use of the order information in your data (i.e., which numbers are bigger), but doesn’t depend on the precise numbers involved. That’s exactly the situation that applies when your data are ordinal scale. The mean, on the other hand, makes use of the precise numeric values assigned to the observations, so it’s not really appropriate for ordinal data.

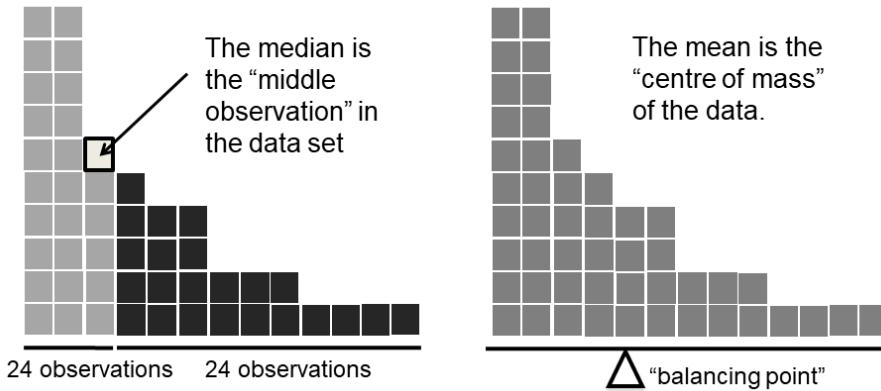


Figure 3.2: An illustration of the difference between how the mean and the median should be interpreted. The mean is basically the “centre of gravity” of the data set: if you imagine that the histogram of the data is a solid object, then the point on which you could balance it (as if on a see-saw) is the mean. In contrast, the median is the middle observation. Half of the observations are smaller, and half of the observations are larger.

- For interval and ratio scale data, either one is generally acceptable. Which one you pick depends a bit on what you’re trying to achieve. The mean has the advantage that it uses all the information in the data (which is useful when you don’t have a lot of data), but it’s very sensitive to extreme values, as we’ll see in Section 3.3.6.

Let’s expand on that last part a little. One consequence is that there’s systematic differences between the mean and the median when the histogram is asymmetric (skewed; see Section 3.5). This is illustrated in Figure 3.2 notice that the median (right hand side) is located closer to the “body” of the histogram, whereas the mean (left hand side) gets dragged towards the “tail” (where the extreme values are). To give a concrete example, suppose Bob (income \$50,000), Kate (income \$60,000) and Jane (income \$65,000) are sitting at a table: the average income at the table is \$58,333 and the median income is \$60,000. Then Bill sits down with them (income \$100,000,000). The average income has now jumped to \$25,043,750 but the median rises only to \$62,500. If you’re interested in looking at the overall income at the table, the mean might be the right answer; but if you’re interested in what counts as a typical income at the table, the median would be a better choice here.

3.3.5 A real life example

To try to get a sense of why you need to pay attention to the differences between the mean and the median, let's consider a real life example. Since I tend to mock journalists for their poor scientific and statistical knowledge, I should give credit where credit is due. This is from an excellent article on the ABC news website⁴ 24 September, 2010:

Senior Commonwealth Bank executives have travelled the world in the past couple of weeks with a presentation showing how Australian house prices, and the key price to income ratios, compare favourably with similar countries. “Housing affordability has actually been going sideways for the last five to six years,” said Craig James, the chief economist of the bank’s trading arm, CommSec.

This probably comes as a huge surprise to anyone with a mortgage, or who wants a mortgage, or pays rent, or isn’t completely oblivious to what’s been going on in the Australian housing market over the last several years. Back to the article:

CBA has waged its war against what it believes are housing doomsayers with graphs, numbers and international comparisons. In its presentation, the bank rejects arguments that Australia’s housing is relatively expensive compared to incomes. It says Australia’s house price to household income ratio of 5.6 in the major cities, and 4.3 nationwide, is comparable to many other developed nations. It says San Francisco and New York have ratios of 7, Auckland’s is 6.7, and Vancouver comes in at 9.3.

More excellent news! Except, the article goes on to make the observation that...

Many analysts say that has led the bank to use misleading figures and comparisons. If you go to page four of CBA’s presentation and read the source information at the bottom of the graph and table, you would notice there is an additional source on the international comparison – Demographia. However, if the Commonwealth Bank had also used Demographia’s analysis of Australia’s house price to income ratio, it would have come up with a figure closer to 9 rather than 5.6 or 4.3

That’s, um, a rather serious discrepancy. One group of people say 9, another says 4-5. Should we just split the difference, and say the truth lies somewhere in between? Absolutely not: this is a situation where there is a right answer and a wrong answer. Demographia are correct, and the Commonwealth Bank is incorrect. As the article points out

⁴www.abc.net.au/news/stories/2010/09/24/3021480.htm

[An] obvious problem with the Commonwealth Bank's domestic price to income figures is they compare average incomes with median house prices (unlike the Demographia figures that compare median incomes to median prices). The median is the mid-point, effectively cutting out the highs and lows, and that means the average is generally higher when it comes to incomes and asset prices, because it includes the earnings of Australia's wealthiest people. To put it another way: the Commonwealth Bank's figures count Ralph Norris' multi-million dollar pay packet on the income side, but not his (no doubt) very expensive house in the property price figures, thus understating the house price to income ratio for middle-income Australians.

Couldn't have put it better myself. The way that Demographia calculated the ratio is the right thing to do. The way that the Bank did it is incorrect. As for why an extremely quantitatively sophisticated organisation such as a major bank made such an elementary mistake, well... I can't say for sure, since I have no special insight into their thinking, but the article itself does happen to mention the following facts, which may or may not be relevant:

[As] Australia's largest home lender, the Commonwealth Bank has one of the biggest vested interests in house prices rising. It effectively owns a massive swathe of Australian housing as security for its home loans as well as many small business loans.

My, my.

3.3.6 Trimmed mean

One of the fundamental rules of applied statistics is that the data are messy. Real life is never simple, and so the data sets that you obtain are never as straightforward as the statistical theory says.⁵ This can have awkward consequences. To illustrate, consider this rather strange looking data set:

$$-100, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

If you were to observe this in a real life data set, you'd probably suspect that something funny was going on with the -100 value. It's probably an *outlier*, a value that doesn't really belong with the others. You might consider removing it from the data set entirely, and in this particular case I'd probably agree with

⁵Or at least, the basic statistical theory – these days there is a whole subfield of statistics called *robust statistics* that tries to grapple with the messiness of real data and develop theory that can cope with it.

that course of action. In real life, however, you don't always get such cut-and-dried examples. For instance, you might get this instead:

-15, 2, 3, 4, 5, 6, 7, 8, 9, 12

The -15 looks a bit suspicious, but not anywhere near as much as that -100 did. In this case, it's a little trickier. It *might* be a legitimate observation, it might not.

When faced with a situation where some of the most extreme-valued observations might not be quite trustworthy, the mean is not necessarily a good measure of central tendency. It is highly sensitive to one or two extreme values, and is thus not considered to be a *robust* measure. One remedy that we've seen is to use the median. A more general solution is to use a "trimmed mean". To calculate a trimmed mean, what you do is "discard" the most extreme examples on both ends (i.e., the largest and the smallest), and then take the mean of everything else. The goal is to preserve the best characteristics of the mean and the median: just like a median, you aren't highly influenced by extreme outliers, but like the mean, you "use" more than one of the observations. Generally, we describe a trimmed mean in terms of the percentage of observation on either side that are discarded. So, for instance, a 10% trimmed mean discards the largest 10% of the observations *and* the smallest 10% of the observations, and then takes the mean of the remaining 80% of the observations. Not surprisingly, the 0% trimmed mean is just the regular mean, and the 50% trimmed mean is the median. In that sense, trimmed means provide a whole family of central tendency measures that span the range from the mean to the median.

For our toy example above, we have 10 observations, and so a 10% trimmed mean is calculated by ignoring the largest value (i.e., 12) and the smallest value (i.e., -15) and taking the mean of the remaining values. First, let's enter the data

```
dataset <- c( -15, 2, 3, 4, 5, 6, 7, 8, 9, 12 )
```

Next, let's calculate means and medians:

```
mean( x = dataset )
## [1] 4.1

median( x = dataset )
## [1] 5.5
```

That's a fairly substantial difference, but I'm tempted to think that the mean is being influenced a bit too much by the extreme values at either end of the

data set, especially the -15 one. So let's just try trimming the mean a bit. If I take a 10% trimmed mean, we'll drop the extreme values on either side, and take the mean of the rest:

```
mean( x = dataset, trim = .1)
```

```
## [1] 5.5
```

which in this case gives exactly the same answer as the median. Note that, to get a 10% trimmed mean you write `trim = .1`, not `trim = 10`. In any case, let's finish up by calculating the 5% trimmed mean for the `afl.margins` data,

```
mean( x = afl.margins, trim = .05)
```

```
## [1] 33.75
```

3.3.7 Mode

The mode of a sample is very simple: it is the value that occurs most frequently. To illustrate the mode using the AFL data, let's examine a different aspect to the data set. Who has played in the most finals? The `afl.finalists` variable is a factor that contains the name of every team that played in any AFL final from 1987-2010, so let's have a look at it. To do this we will use the `head()` command. `head()` is useful when you're working with a `data.frame` with a lot of rows since you can use it to tell you how many rows to return. There have been a lot of finals in this period so printing `afl.finalists` using `print(afl.finalists)` will just fill us the screen. The command below tells R we just want the first 25 rows of the `data.frame`.

```
head(afl.finalists, 25)
```

```
## [1] Hawthorn    Melbourne   Carlton     Melbourne   Hawthorn
## [6] Carlton     Melbourne   Carlton     Hawthorn   Melbourne
## [11] Melbourne   Hawthorn   Melbourne   Essendon   Hawthorn
## [16] Geelong     Geelong    Hawthorn   Collingwood Melbourne
## [21] Collingwood West Coast Collingwood Essendon   Collingwood
## 17 Levels: Adelaide Brisbane Carlton Collingwood Essendon ... Western Bulldogs
```

There are actually 400 entries (aren't you glad we didn't print them all?). We could read through all 400, and count the number of occasions on which each team name appears in our list of finalists, thereby producing a *frequency table*. However, that would be mindless and boring: exactly the sort of task that computers are great at. So let's use the `table()` function (discussed in more detail in Section ??) to do this task for us:

```
table( afl.finalists )

## afl.finalists
##      Adelaide      Brisbane      Carlton      Collingwood
##          26           25           26           28
##      Essendon      Fitzroy      Fremantle     Geelong
##          32            0           6           39
##      Hawthorn    Melbourne  North Melbourne  Port Adelaide
##          27           28           28           17
##      Richmond     St Kilda      Sydney      West Coast
##          6            24           26           38
## Western Bulldogs
##          24
```

Now that we have our frequency table, we can just look at it and see that, over the 24 years for which we have data, Geelong has played in more finals than any other team. Thus, the mode of the `finalists` data is "Geelong". The core packages in R don't have a function for calculating the mode⁶. However, I've included a function in the `lsr` package that does this. The function is called `modeOf()`, and here's how you use it:

```
modeOf( x = afl.finalists )

## [1] "Geelong"
```

There's also a function called `maxFreq()` that tells you what the modal frequency is. If we apply this function to our `finalists` data, we obtain the following:

```
maxFreq( x = afl.finalists )

## [1] 39
```

Taken together, we observe that Geelong (39 finals) played in more finals than any other team during the 1987-2010 period.

One last point to make with respect to the mode. While it's generally true that the mode is most often calculated when you have nominal scale data (because means and medians are useless for those sorts of variables), there are some situations in which you really do want to know the mode of an ordinal, interval or ratio scale variable. For instance, let's go back to thinking about our `afl.margins` variable. This variable is clearly ratio scale (if it's not clear

⁶As we saw earlier, it *does* have a function called `mode()`, but it does something completely different.

to you, it may help to re-read Section ??), and so in most situations the mean or the median is the measure of central tendency that you want. But consider this scenario... a friend of yours is offering a bet. They pick a football game at random, and (without knowing who is playing) you have to guess the *exact* margin. If you guess correctly, you win \$50. If you don't, you lose \$1. There are no consolation prizes for "almost" getting the right answer. You have to guess exactly the right margin⁷ For this bet, the mean and the median are completely useless to you. It is the mode that you should bet on. So, we calculate this modal value

```
mode0f( x = afl.margins )
## [1] 3

maxFreq( x = afl.margins )
## [1] 8
```

So the 2010 data suggest you should bet on a 3 point margin, and since this was observed in 8 of the 176 game (4.5% of games) the odds are firmly in your favour.

3.4 Measures of variability

The statistics that we've discussed so far all relate to *central tendency*. That is, they all talk about which values are "in the middle" or "popular" in the data. However, central tendency is not the only type of summary statistic that we want to calculate. The second thing that we really want is a measure of the ***variability*** of the data. That is, how "spread out" are the data? How "far" away from the mean or median do the observed values tend to be? For now, let's assume that the data are interval or ratio scale, so we'll continue to use the `afl.margins` data. We'll use this data to discuss several different measures of spread, each with different strengths and weaknesses.

3.4.1 Range

The ***range*** of a variable is very simple: it's the biggest value minus the smallest value. For the AFL winning margins data, the maximum value is 116, and the minimum value is 0. We can calculate these values in R using the `max()` and `min()` functions:

⁷This is called a "0-1 loss function", meaning that you either win (1) or you lose (0), with no middle ground.

```
max( afl.margins )
## [1] 116
min( afl.margins )
## [1] 0
```

where I've omitted the output because it's not interesting. The other possibility is to use the `range()` function; which outputs both the minimum value and the maximum value in a vector, like this:

```
range( afl.margins )
## [1] 0 116
```

Although the range is the simplest way to quantify the notion of “variability”, it’s one of the worst. Recall from our discussion of the mean that we want our summary measure to be robust. If the data set has one or two extremely bad values in it, we’d like our statistics not to be unduly influenced by these cases. If we look once again at our toy example of a data set containing very extreme outliers...

–100, 2, 3, 4, 5, 6, 7, 8, 9, 10

... it is clear that the range is not robust, since this has a range of 110, but if the outlier were removed we would have a range of only 8.

3.4.2 Interquartile range

The *interquartile range* (IQR) is like the range, but instead of calculating the difference between the biggest and smallest value, it calculates the difference between the 25th quantile and the 75th quantile. Probably you already know what a *quantile* is (they’re more commonly called percentiles), but if not: the 10th percentile of a data set is the smallest number x such that 10% of the data is less than x . In fact, we’ve already come across the idea: the median of a data set is its 50th quantile / percentile! R actually provides you with a way of calculating quantiles, using the (surprise, surprise) `quantile()` function. Let’s use it to calculate the median AFL winning margin:

```
quantile( x = afl.margins, probs = .5)
## 50%
## 30.5
```

And not surprisingly, this agrees with the answer that we saw earlier with the `median()` function. Now, we can actually input lots of quantiles at once, by specifying a vector for the `probs` argument. So lets do that, and get the 25th and 75th percentile:

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

##    25%    75%
## 12.75 50.50
```

And, by noting that $50.5 - 12.75 = 37.75$, we can see that the interquartile range for the 2010 AFL winning margins data is 37.75. Of course, that seems like too much work to do all that typing, so R has a built in function called `IQR()` that we can use:

```
IQR( x = afl.margins )

## [1] 37.75
```

While it's obvious how to interpret the range, it's a little less obvious how to interpret the IQR. The simplest way to think about it is like this: the interquartile range is the range spanned by the “middle half” of the data. That is, one quarter of the data falls below the 25th percentile, one quarter of the data is above the 75th percentile, leaving the “middle half” of the data lying in between the two. And the IQR is the range covered by that middle half.

3.4.3 Mean absolute deviation

The two measures we've looked at so far, the range and the interquartile range, both rely on the idea that we can measure the spread of the data by looking at the quantiles of the data. However, this isn't the only way to think about the problem. A different approach is to select a meaningful reference point (usually the mean or the median) and then report the “typical” deviations from that reference point. What do we mean by “typical” deviation? Usually, the mean or median value of these deviations! In practice, this leads to two different measures, the “mean absolute deviation (from the mean)” and the “median absolute deviation (from the median)”. From what I've read, the measure based on the median seems to be used in statistics, and does seem to be the better of the two, but to be honest I don't think I've seen it used much in psychology. The measure based on the mean does occasionally show up in psychology though. In this section I'll talk about the first one, and I'll come back to talk about the second one later.

Since the previous paragraph might sound a little abstract, let's go through the ***mean absolute deviation*** from the mean a little more slowly. One useful thing about this measure is that the name actually tells you exactly how to calculate it. Let's think about our AFL winning margins data, and once again we'll start by pretending that there's only 5 games in total, with winning margins of 56, 31, 56, 8 and 32. Since our calculations rely on an examination of the deviation from some reference point (in this case the mean), the first thing we need to calculate is the mean, \bar{X} . For these five observations, our mean is $\bar{X} = 36.6$. The next step is to convert each of our observations X_i into a deviation score. We do this by calculating the difference between the observation X_i and the mean \bar{X} . That is, the deviation score is defined to be $X_i - \bar{X}$. For the first observation in our sample, this is equal to $56 - 36.6 = 19.4$. Okay, that's simple enough. The next step in the process is to convert these deviations to absolute deviations. As we discussed earlier when talking about the `abs()` function in R (Section 2.7), we do this by converting any negative values to positive ones. Mathematically, we would denote the absolute value of -3 as $|-3|$, and so we say that $|-3| = 3$. We use the absolute value function here because we don't really care whether the value is higher than the mean or lower than the mean, we're just interested in how *close* it is to the mean. To help make this process as obvious as possible, the table below shows these calculations for all five observations:

the observation	its symbol	the observed value
winning margin, game 1	X_1	56 points
winning margin, game 2	X_2	31 points
winning margin, game 3	X_3	56 points
winning margin, game 4	X_4	8 points
winning margin, game 5	X_5	32 points

Now that we have calculated the absolute deviation score for every observation in the data set, all that we have to do to calculate the mean of these scores. Let's do that:

$$\frac{19.4 + 5.6 + 19.4 + 28.6 + 4.6}{5} = 15.52$$

And we're done. The mean absolute deviation for these five scores is 15.52.

However, while our calculations for this little example are at an end, we do have a couple of things left to talk about. Firstly, we should really try to write down a proper mathematical formula. But in order to do this I need some mathematical notation to refer to the mean absolute deviation. Irritatingly, “mean absolute deviation” and “median absolute deviation” have the same acronym (MAD), which leads to a certain amount of ambiguity, and since R tends to use MAD to refer to the median absolute deviation, I'd better come up with something different for the mean absolute deviation. Sigh. What I'll do is use AAD instead, short for *average* absolute deviation. Now that we have some unambiguous

notation, here's the formula that describes what we just calculated:

$$(X) = \frac{1}{N} \sum_{i=1}^N |X_i - \bar{X}|$$

The last thing we need to talk about is how to calculate AAD in R. One possibility would be to do everything using low level commands, laboriously following the same steps that I used when describing the calculations above. However, that's pretty tedious. You'd end up with a series of commands that might look like this:

```
X <- c(56, 31, 56, 8, 32)      # enter the data
X.bar <- mean(X)               # step 1. the mean of the data
AD <- abs(X - X.bar)          # step 2. the absolute deviations from the mean
AAD <- mean(AD)                # step 3. the mean absolute deviations
print(AAD)                     # print the results

## [1] 15.52
```

Each of those commands is pretty simple, but there's just too many of them. And because I find that to be too much typing, the `lsr` package has a very simple function called `aad()` that does the calculations for you. If we apply the `aad()` function to our data, we get this:

```
library(ls)
aad(X)
```

```
## [1] 15.52
```

No surprises there.

3.4.4 Variance

Although the mean absolute deviation measure has its uses, it's not the best measure of variability to use. From a purely mathematical perspective, there are some solid reasons to prefer squared deviations rather than absolute deviations. If we do that, we obtain a measure is called the *variance*, which has a lot of really nice statistical properties that I'm going to ignore,⁸ and $\text{Var}(Y)$ respectively. Now imagine I want to define a new variable Z that is the sum of the two, $Z = X + Y$. As it turns out, the variance of Z is equal to $\text{Var}(X) +$

⁸Well, I will very briefly mention the one that I think is coolest, for a very particular definition of "cool", that is. Variances are *additive*. Here's what that means: suppose I have two variables X and Y , whose variances are Var

Table 3.1: Basic arithmetic operations in R. These five operators are used very frequently throughout the text, so it's important to be familiar with them at the outset.

Notation [English]	<code>\$i\$</code> [which game]	<code>\$X_i\$</code> [value]	<code>\$X_i - \bar{X}\$</code> [deviation from mean]	<code>\$(X)\$</code>
1		56	19.4	376.36
2		31	-5.6	31.36
3		56	19.4	376.36
4		8	-28.6	817.96
5		32	-4.6	21.16

$\text{Var}(Y)$. This is a *very* useful property, but it's not true of the other measures that I talk about in this section.] and one massive psychological flaw that I'm going to make a big deal out of in a moment. The variance of a data set X is sometimes written as $\text{Var}(X)$, but it's more commonly denoted s^2 (the reason for this will become clearer shortly). The formula that we use to calculate the variance of a set of observations is as follows:

$$\text{Var}(X) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$$

$$\text{Var}(X) = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}$$

As you can see, it's basically the same formula that we used to calculate the mean absolute deviation, except that instead of using "absolute deviations" we use "squared deviations". It is for this reason that the variance is sometimes referred to as the "mean square deviation".

Now that we've got the basic idea, let's have a look at a concrete example. Once again, let's use the first five AFL games as our data. If we follow the same approach that we took last time, we end up with the following table:

That last column contains all of our squared deviations, so all we have to do is average them. If we do that by typing all the numbers into R by hand...

```
( 376.36 + 31.36 + 376.36 + 817.96 + 21.16 ) / 5
```

```
## [1] 324.64
```

... we end up with a variance of 324.64. Exciting, isn't it? For the moment, let's ignore the burning question that you're all probably thinking (i.e., what the heck does a variance of 324.64 actually mean?) and instead talk a bit more about how to do the calculations in R, because this will reveal something very weird.

As always, we want to avoid having to type in a whole lot of numbers ourselves. And as it happens, we have the vector `X` lying around, which we created in the previous section. With this in mind, we can calculate the variance of `X` by using the following command,

```
mean( (X - mean(X))^2)
## [1] 324.64
```

and as usual we get the same answer as the one that we got when we did everything by hand. However, I *still* think that this is too much typing. Fortunately, R has a built in function called `var()` which does calculate variances. So we could also do this...

```
var(X)
## [1] 405.8
```

and you get the same... no, wait... you get a completely *different* answer. That's just weird. Is R broken? Is this a typo? Is Dan an idiot?

As it happens, the answer is no.⁹ It's not a typo, and R is not making a mistake. To get a feel for what's happening, let's stop using the tiny data set containing only 5 data points, and switch to the full set of 176 games that we've got stored in our `afl.margins` vector. First, let's calculate the variance by using the formula that I described above:

```
mean( (afl.margins - mean(afl.margins))^2)
## [1] 675.9718
```

Now let's use the `var()` function:

```
var( afl.margins )
## [1] 679.8345
```

Hm. These two numbers are very similar this time. That seems like too much of a coincidence to be a mistake. And of course it isn't a mistake. In fact, it's very simple to explain what R is doing here, but slightly trickier to explain *why* R is doing it. So let's start with the "what". What R is doing is evaluating a

⁹With the possible exception of the third question.

slightly different formula to the one I showed you above. Instead of averaging the squared deviations, which requires you to divide by the number of data points N , R has chosen to divide by $N - 1$. In other words, the formula that R is using is this one

$$\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$$

It's easy enough to verify that this is what's happening, as the following command illustrates:

```
sum( (X-mean(X))^2 ) / 4
```

```
## [1] 405.8
```

This is the same answer that R gave us originally when we calculated `var(x)` originally. So that's the *what*. The real question is *why* R is dividing by $N-1$ and not by N . After all, the variance is supposed to be the *mean* squared deviation, right? So shouldn't we be dividing by N , the actual number of observations in the sample? Well, yes, we should. However, as we'll discuss in Chapter 4.2, there's a subtle distinction between "describing a sample" and "making guesses about the population from which the sample came". Up to this point, it's been a distinction without a difference. Regardless of whether you're describing a sample or drawing inferences about the population, the mean is calculated exactly the same way. Not so for the variance, or the standard deviation, or for many other measures besides. What I outlined to you initially (i.e., take the actual average, and thus divide by N) assumes that you literally intend to calculate the variance of the sample. Most of the time, however, you're not terribly interested in the sample *in and of itself*. Rather, the sample exists to tell you something about the world. If so, you're actually starting to move away from calculating a "sample statistic", and towards the idea of estimating a "population parameter". However, I'm getting ahead of myself. For now, let's just take it on faith that R knows what it's doing, and we'll revisit the question later on when we talk about estimation in Chapter 4.2.

Okay, one last thing. This section so far has read a bit like a mystery novel. I've shown you how to calculate the variance, described the weird " $N-1$ " thing that R does and hinted at the reason why it's there, but I haven't mentioned the single most important thing... how do you *interpret* the variance? Descriptive statistics are supposed to describe things, after all, and right now the variance is really just a gibberish number. Unfortunately, the reason why I haven't given you the human-friendly interpretation of the variance is that there really isn't one. This is the most serious problem with the variance. Although it has some elegant mathematical properties that suggest that it really is a fundamental quantity for expressing variation, it's completely useless if you want to communicate with an actual human... variances are completely uninterpretable in terms of

the original variable! All the numbers have been squared, and they don't mean anything anymore. This is a huge issue. For instance, according to the table I presented earlier, the margin in game 1 was "376.36 points-squared higher than the average margin". This is *exactly* as stupid as it sounds; and so when we calculate a variance of 324.64, we're in the same situation. I've watched a lot of footy games, and never has anyone referred to "points squared". It's *not* a real unit of measurement, and since the variance is expressed in terms of this gibberish unit, it is totally meaningless to a human.

3.4.5 Standard deviation

Okay, suppose that you like the idea of using the variance because of those nice mathematical properties that I haven't talked about, but – since you're a human and not a robot – you'd like to have a measure that is expressed in the same units as the data itself (i.e., points, not points-squared). What should you do? The solution to the problem is obvious: take the square root of the variance, known as the **standard deviation**, also called the "root mean squared deviation", or RMSD. This solves out problem fairly neatly: while nobody has a clue what "a variance of 324.68 points-squared" really means, it's much easier to understand "a standard deviation of 18.01 points", since it's expressed in the original units. It is traditional to refer to the standard deviation of a sample of data as s , though "sd" and "std dev." are also used at times. Because the standard deviation is equal to the square root of the variance, you probably won't be surprised to see that the formula is:

$$s = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2}$$

and the R function that we use to calculate it is `sd()`. However, as you might have guessed from our discussion of the variance, what R actually calculates is slightly different to the formula given above. Just like the we saw with the variance, what R calculates is a version that divides by $N - 1$ rather than N . For reasons that will make sense when we return to this topic in Chapter@refch:estimation I'll refer to this new quantity as $\hat{\sigma}$ (read as: "sigma hat"), and the formula for this is

$$\hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2}$$

With that in mind, calculating standard deviations in R is simple:

```
sd( afl.margins )
```

```
## [1] 26.07364
```

Interpreting standard deviations is slightly more complex. Because the standard deviation is derived from the variance, and the variance is a quantity that has little to no meaning that makes sense to us humans, the standard deviation doesn't have a simple interpretation. As a consequence, most of us just rely on a simple rule of thumb: in general, you should expect 68% of the data to fall within 1 standard deviation of the mean, 95% of the data to fall within 2 standard deviation of the mean, and 99.7% of the data to fall within 3 standard deviations of the mean. This rule tends to work pretty well most of the time, but it's not exact: it's actually calculated based on an *assumption* that the histogram is symmetric and "bell shaped".¹⁰ As you can tell from looking at the AFL winning margins histogram in Figure 3.1, this isn't exactly true of our data! Even so, the rule is approximately correct. As it turns out, 65.3% of the AFL margins data fall within one standard deviation of the mean. This is shown visually in Figure 3.3.

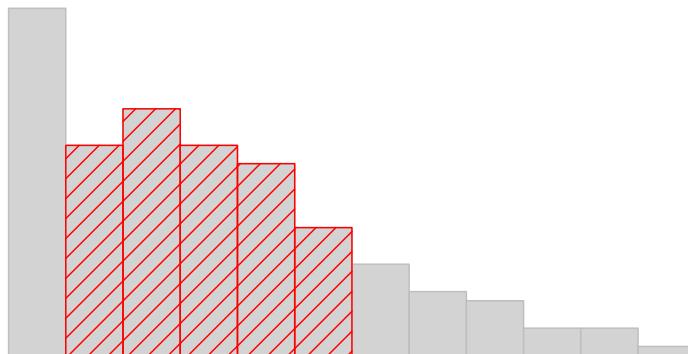


Figure 3.3: An illustration of the standard deviation, applied to the AFL winning margins data. The shaded bars in the histogram show how much of the data fall within one standard deviation of the mean. In this case, 65.3% of the data set lies within this range, which is pretty consistent with the "approximately 68% rule" discussed in the main text.

¹⁰Strictly, the assumption is that the data are *normally* distributed, which is an important concept that we'll discuss more in Chapter ??, and will turn up over and over again later in the book.

3.4.6 Median absolute deviation

The last measure of variability that I want to talk about is the *median absolute deviation* (MAD). The basic idea behind MAD is very simple, and is pretty much identical to the idea behind the mean absolute deviation (Section 3.4.3). The difference is that you use the median everywhere. If we were to frame this idea as a pair of R commands, they would look like this:

```
# mean absolute deviation from the mean:  
mean( abs(afl.margins - mean(afl.margins)) )  
  
## [1] 21.10124  
  
# *median* absolute deviation from the *median*:  
median( abs(afl.margins - median(afl.margins)) )  
  
## [1] 19.5
```

This has a straightforward interpretation: every observation in the data set lies some distance away from the typical value (the median). So the MAD is an attempt to describe a *typical deviation from a typical value* in the data set. It wouldn't be unreasonable to interpret the MAD value of 19.5 for our AFL data by saying something like this:

The median winning margin in 2010 was 30.5, indicating that a typical game involved a winning margin of about 30 points. However, there was a fair amount of variation from game to game: the MAD value was 19.5, indicating that a typical winning margin would differ from this median value by about 19-20 points.

As you'd expect, R has a built in function for calculating MAD, and you will be shocked no doubt to hear that it's called `mad()`. However, it's a little bit more complicated than the functions that we've been using previously. If you want to use it to calculate MAD in the exact same way that I have described it above, the command that you need to use specifies two arguments: the data set itself `x`, and a `constant` that I'll explain in a moment. For our purposes, the constant is 1, so our command becomes

```
mad( x = afl.margins, constant = 1 )  
  
## [1] 19.5
```

Apart from the weirdness of having to type that `constant = 1` part, this is pretty straightforward.

Okay, so what exactly is this `constant = 1` argument? I won't go into all the details here, but here's the gist. Although the "raw" MAD value that I've described above is completely interpretable on its own terms, that's not actually how it's used in a lot of real world contexts. Instead, what happens a lot is that the researcher *actually* wants to calculate the standard deviation. However, in the same way that the mean is very sensitive to extreme values, the standard deviation is vulnerable to the exact same issue. So, in much the same way that people sometimes use the median as a "robust" way of calculating "something that is like the mean", it's not uncommon to use MAD as a method for calculating "something that is like the standard deviation". Unfortunately, the *raw* MAD value doesn't do this. Our raw MAD value is 19.5, and our standard deviation was 26.07. However, what some clever person has shown is that, under certain assumptions¹¹, you can multiply the raw MAD value by 1.4826 and obtain a number that is directly comparable to the standard deviation. As a consequence, the default value of `constant` is 1.4826, and so when you use the `mad()` command without manually setting a value, here's what you get:

```
mad( afl.margins )
```

```
## [1] 28.9107
```

I should point out, though, that if you want to use this "corrected" MAD value as a robust version of the standard deviation, you really are relying on the assumption that the data are (or at least, are "supposed to be" in some sense) symmetric and basically shaped like a bell curve. That's really *not* true for our `afl.margins` data, so in this case I wouldn't try to use the MAD value this way.

3.4.7 Which measure to use?

We've discussed quite a few measures of spread (range, IQR, MAD, variance and standard deviation), and hinted at their strengths and weaknesses. Here's a quick summary:

- *Range*. Gives you the full spread of the data. It's very vulnerable to outliers, and as a consequence it isn't often used unless you have good reasons to care about the extremes in the data.
- *Interquartile range*. Tells you where the "middle half" of the data sits. It's pretty robust, and complements the median nicely. This is used a lot.

¹¹The assumption again being that the data are normally-distributed!

- *Mean absolute deviation.* Tells you how far “on average” the observations are from the mean. It’s very interpretable, but has a few minor issues (not discussed here) that make it less attractive to statisticians than the standard deviation. Used sometimes, but not often.
- *Variance.* Tells you the average squared deviation from the mean. It’s mathematically elegant, and is probably the “right” way to describe variation around the mean, but it’s completely uninterpretable because it doesn’t use the same units as the data. Almost never used except as a mathematical tool; but it’s buried “under the hood” of a very large number of statistical tools.
- *Standard deviation.* This is the square root of the variance. It’s fairly elegant mathematically, and it’s expressed in the same units as the data so it can be interpreted pretty well. In situations where the mean is the measure of central tendency, this is the default. This is by far the most popular measure of variation.
- *Median absolute deviation.* The typical (i.e., median) deviation from the median value. In the raw form it’s simple and interpretable; in the corrected form it’s a robust way to estimate the standard deviation, for some kinds of data sets. Not used very often, but it does get reported sometimes.

In short, the IQR and the standard deviation are easily the two most common measures used to report the variability of the data; but there are situations in which the others are used. I’ve described all of them in this book because there’s a fair chance you’ll run into most of these somewhere.

3.5 Skew and kurtosis

Text by David Schuster

There are two more descriptive statistics that you will sometimes see reported in the psychological literature, known as skew and kurtosis. These two measures are useful to diagnose normality (although kurtosis is less useful, in practice). If we find evidence of skewness or kurtosis, we may suspect that our distribution is not normal. Later in this course, we will see that it is generally not a problem for us if one of our sample distributions are non-normal. Regardless, skewness and kurtosis can tell us about the shape of our data.

Example by Navarro (2018)

```
## [1] -0.9139277
## [1] -0.008227129
## [1] 0.9096447
```

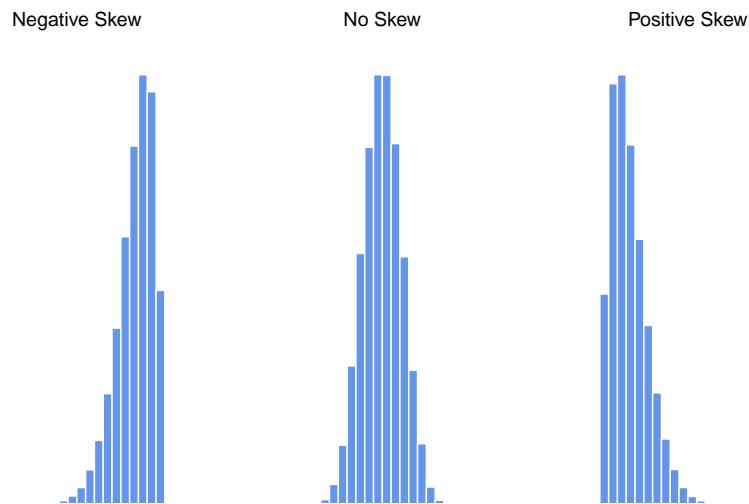


Figure 3.4: An illustration of skewness. On the left we have a negatively skewed data set (skewness = $-.93$), in the middle we have a data set with no skew (technically, skewness = $-.006$), and on the right we have a positively skewed data set (skewness = $.93$).

Since it's the more interesting of the two, let's start by talking about the *skewness*. Skewness is basically a measure of asymmetry, and the easiest way to explain it is by drawing some pictures. As Figure 3.4 illustrates, if the data tend to have a lot of extreme small values (i.e., the lower tail is “longer” than the upper tail) and not so many extremely large values (left panel), then we say that the data are *negatively skewed*. On the other hand, if there are more extremely large values than extremely small ones (right panel) we say that the data are *positively skewed*. That's the qualitative idea behind skewness. The actual formula for the skewness of a data set is as follows

$$\text{skewness}(X) = \frac{1}{N\hat{\sigma}^3} \sum_{i=1}^N (X_i - \bar{X})^3$$

where N is the number of observations, \bar{X} is the sample mean, and $\hat{\sigma}$ is the standard deviation (the “divide by $N - 1$ ” version, that is). Perhaps more helpfully, it might be useful to point out that the `psych` package contains a `skew()` function that you can use to calculate skewness. So if we wanted to use this function to calculate the skewness of the `afl.margins` data, we'd first need to load the package

```
library( psych )
```

which now makes it possible to use the following command:

```
skew( x = afl.margins )
```

```
## [1] 0.7671555
```

Not surprisingly, it turns out that the AFL winning margins data is fairly skewed.

3.5.1 More detail on skewness measures

Text by David Schuster

Unfortunately, this characterization of skewness is a bit of a simplification. R's `psych` package does have a skewness statistic, but it does not match the formula Navarro provided in this section. I spent an afternoon disentangling all the available skewness statistics. Rather than get too detailed with this (if you want that detail, look for Doane & Seward's 2011 article on measuring skewness), allow me to summarize:

- First, my overall recommendation would be to visually examine a Q-Q plot and histogram, then run a Shapiro-Wilks test, understanding that none of these methods are perfect.

- Skewness is defined mathematically as the third **moment**. Moments are statistical parameters. The first moment is mean, the second moment is variance, the third moment is skewness, and the fourth moment is kurtosis
- The mathematical definition is of limited use to the researcher, because it does not account for sample size or variation in samples (sampling error)
- Statisticians have come up with several more useful measures of skewness. Their performance differences are not very important in practice, and they tend to give the same result when you have large samples.
- It matters less which skewness value you report. It is more important that you are clear about which statistic you used. Frustratingly, stats packages and Excel implement a single method but do not make it clear which method they use. This leads to people saying, “I measured skewness using SPSS,” which the SPSS people probably really like.
- For future reference more than use right now, I researched which measures are used in which packages.

3.5.1.1 Use statistics to describe skew

```

x <- afl.margins # use the same example
n <- length(x) # sample size
sum_cubed <- sum((x - mean(x))^3) # sum of cubed deviations from the mean
pop_sd <- sqrt(var(x) * (n-1)/n) # population sd
sample_sd <- sd(x)
pop_sd_cubed <- pop_sd^3 # population sd cubed

gamma_1 <- sum_cubed / pop_sd_cubed # Pearson's moment coefficient of skewness
g_1 <- sum_cubed / (pop_sd_cubed * n) # Fisher-Pearson coefficient of skewness (Type = 1)
G_1 <- (n / ((n-1)*(n-2))) * sum(((x - mean(x)) / sample_sd)^3) # adjusted Fisher-Pearson coefficient of skewness (Type = 3)
b_1 <- g_1 * ((n-1)/n)^(3/2) # (Type = 3, MINITAB, stat.desc, skew(), psych package)
SE_skew_SPSS = sqrt(6*n*(n-1) / ((n-2)*(n+1)*(n+3))) # SPSS skewness SE per https://www.rdocumentation.org/packages/e1071/functions/se.skew

gamma_1

## [1] 136.1783

g_1

## [1] 0.7737405

```

G_1

```
## [1] 0.7804075
```

b_1

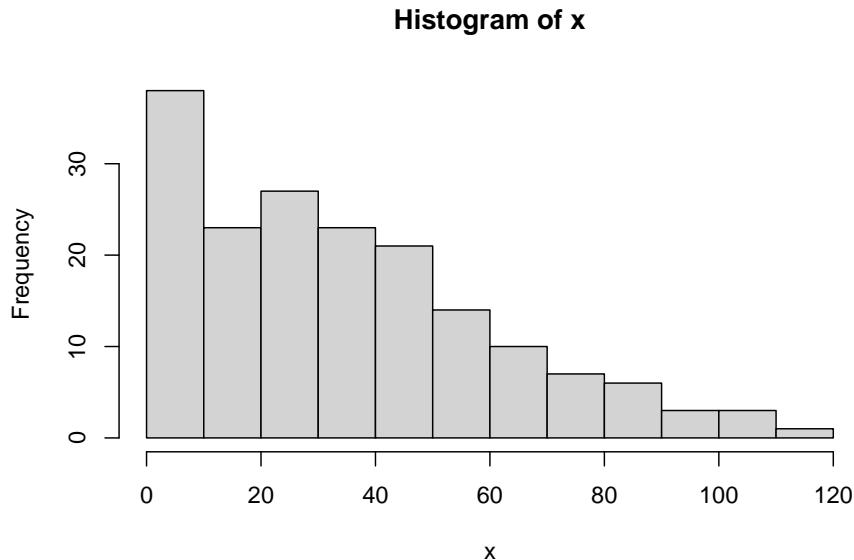
```
## [1] 0.7671555
```

The point of all of this is to be aware that when you use the psych package and `skew()` or `stat.desc` commands, you are reporting β_1 . SPSS reports G_1 as the skewness statistic as well as a standard error statistic. You can replicate this, if you want:

1. Find G_1 using `skew(x, type = 2)`
2. Find SE_{skew} using the R formula: `sqrt(6*n*(n-1) / ((n-2)*(n+1)*(n+3)))`
3. Divide: G_1/SE_{skew} .
4. If the result is between -2 and 2, you can infer a lack of skewness. Values -2 indicate a concerning amount of negative skew; values above 2 indicate a concerning amount of positive skew.

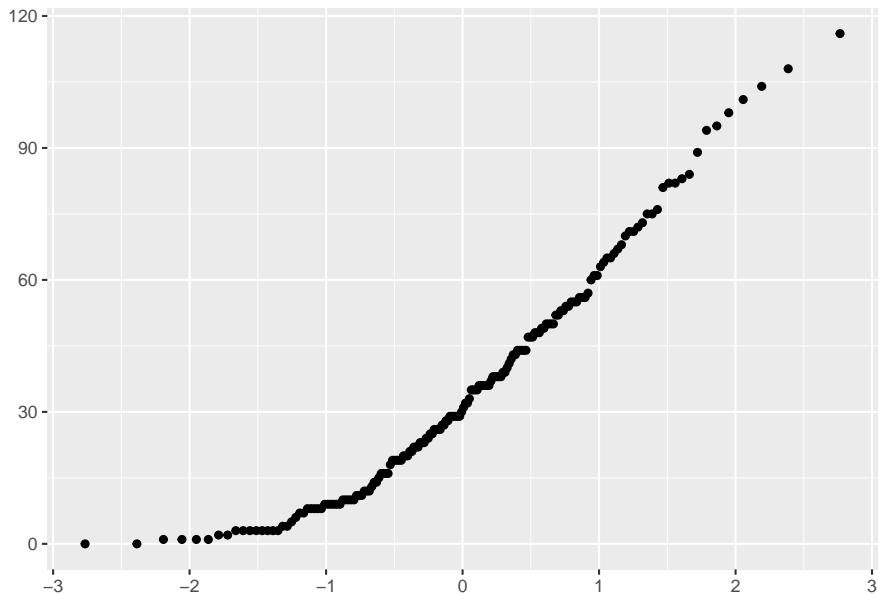
3.5.1.2 Examine the histogram

```
hist(x)
```



3.5.1.3 Look for normality visually in a Q-Q plot (quantile-quantile plot)

```
library("ggplot2")  
  
##  
## Attaching package: 'ggplot2'  
  
## The following objects are masked from 'package:psych':  
##  
##     %+%, alpha  
  
qplot(sample = x) # qplot Q-Q plots
```



- In a normal distribution, the dots of the plot will form a near-straight line.
- Remember that quantiles divide a continuous probability distribution into intervals with equal probabilities.
- A Q-Q plot is based on quantiles of data versus theoretical normal distribution.
- In contrast a P-P plot (not shown here) is based on the cumulative density function (probability) of data versus a theoretical normal distribution.
- Q-Q is better at detecting differences near the middle of the distribution than a P-P plot.

3.5.1.4 The Shapiro-Wilk test tests the null hypothesis that data are from a normal distribution

If $p < .05$, reject the null and conclude the data are not normal. If $p > .05$, retain the null and officially make no inference about the distribution. Also, beware that not all deviations from normality may be detected, and, in large samples, the test may be too sensitive. Report as $W(df) = \#, p = \#$.

```
shapiro.test(x)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: x  
## W = 0.9398, p-value = 9.481e-07
```

The final measure that is sometimes referred to, though very rarely in practice, is the **kurtosis** of a data set. Put simply, kurtosis is a measure of the “pointiness” of a data set, as illustrated in Figure 3.5.

```
## [1] -0.9586669
```

```
## [1] 0.01672576
```

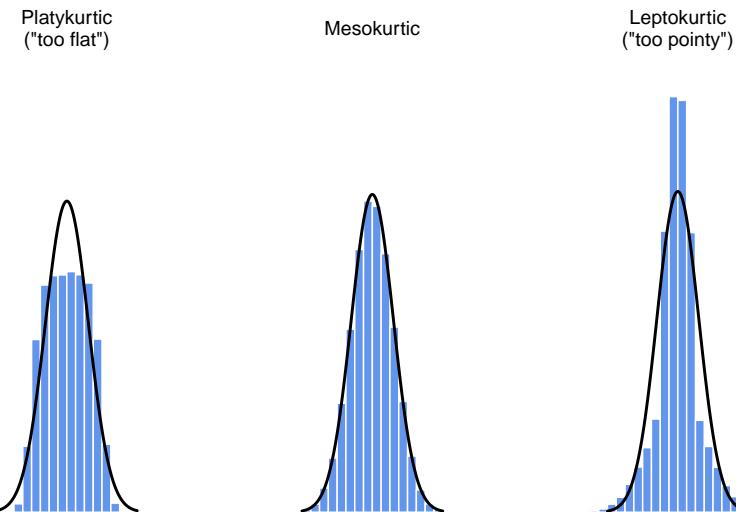


Figure 3.5: An illustration of kurtosis. On the left, we have a “platykurtic” data set (kurtosis = -0.95), meaning that the data set is “too flat”. In the middle we have a “mesokurtic” data set (kurtosis is almost exactly 0), which means that the pointiness of the data is just about right. Finally, on the right, we have a “leptokurtic” data set (kurtosis = 2.12) indicating that the data set is “too pointy”. Note that kurtosis is measured with respect to a normal curve (black line)

```
## [1] 2.011041
```

By convention, we say that the “normal curve” (black lines) has zero kurtosis, so the pointiness of a data set is assessed relative to this curve. In this Figure, the data on the left are not pointy enough, so the kurtosis is negative and we call the data *platykurtic*. The data on the right are too pointy, so the kurtosis is positive and we say that the data is *leptokurtic*. But the data in the middle are just pointy enough, so we say that it is *mesokurtic* and has kurtosis zero. This is summarised in the table below:

informal term	technical name	kurtosis value
too flat	platykurtic	negative
just pointy enough	mesokurtic	zero
too pointy	leptokurtic	positive

The equation for kurtosis is pretty similar in spirit to the formulas we’ve seen already for the variance and the skewness; except that where the variance involved squared deviations and the skewness involved cubed deviations, the kurtosis involves raising the deviations to the fourth power:¹²

$$\text{kurtosis}(X) = \frac{1}{N\hat{\sigma}^4} \sum_{i=1}^N (X_i - \bar{X})^4 - 3$$

I know, it’s not terribly interesting to me either. More to the point, the **psych** package has a function called **kurtosi()** that you can use to calculate the kurtosis of your data. For instance, if we were to do this for the AFL margins,

```
kurtosi( x = afl.margins )
```

```
## [1] 0.02962633
```

we discover that the AFL winning margins data are just pointy enough.

3.6 Getting an overall summary of a variable

Up to this point in the chapter I’ve explained several different summary statistics that are commonly used when analysing data, along with specific functions that you can use in R to calculate each one. However, it’s kind of annoying to have to separately calculate means, medians, standard deviations, skews etc. Wouldn’t it be nice if R had some helpful functions that would do all these tedious calculations at once? Something like **summary()** or **describe()**, perhaps? Why yes, yes it would. So much so that both of these functions exist. The **summary()** function is in the **base** package, so it comes with every installation of R. The **describe()** function is part of the **psych** package, which we loaded earlier in the chapter.

¹²The “–3” part is something that statisticians tack on to ensure that the normal curve has kurtosis zero. It looks a bit stupid, just sticking a “–3” at the end of the formula, but there are good mathematical reasons for doing this.

3.6.1 “Summarising” a variable

The `summary()` function is an easy thing to use, but a tricky thing to understand in full, since it’s a generic function (see Section 2.26). The basic idea behind the `summary()` function is that it prints out some useful information about whatever object (i.e., variable, as far as we’re concerned) you specify as the `object` argument. As a consequence, the behaviour of the `summary()` function differs quite dramatically depending on the class of the object that you give it. Let’s start by giving it a *numeric* object:

```
summary( object = afl.margins )

##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##   0.00  12.75  30.50  35.30  50.50 116.00
```

For numeric variables, we get a whole bunch of useful descriptive statistics. It gives us the minimum and maximum values (i.e., the range), the first and third quartiles (25th and 75th percentiles; i.e., the IQR), the mean and the median. In other words, it gives us a pretty good collection of descriptive statistics related to the central tendency and the spread of the data.

Okay, what about if we feed it a logical vector instead? Let’s say I want to know something about how many “blowouts” there were in the 2010 AFL season. I operationalise the concept of a blowout (see Chapter 1.6) as a game in which the winning margin exceeds 50 points. Let’s create a logical variable `blowouts` in which the *i*-th element is TRUE if that game was a blowout according to my definition,

```
blowouts <- afl.margins > 50
blowouts

## [1] TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE
## [12] TRUE FALSE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE
## [23] FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE TRUE FALSE FALSE
## [34] TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [45] FALSE TRUE TRUE FALSE FALSE FALSE FALSE TRUE TRUE FALSE
## [56] TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE
## [67] TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE
## [78] FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [89] FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE
## [100] FALSE TRUE FALSE FALSE FALSE TRUE FALSE TRUE TRUE TRUE FALSE
## [111] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE
## [122] TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE
## [133] FALSE TRUE TRUE FALSE FALSE FALSE FALSE TRUE FALSE TRUE
## [144] TRUE TRUE TRUE FALSE FALSE TRUE FALSE FALSE TRUE FALSE
```

```
## [155] TRUE FALSE TRUE FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE
## [166] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

So that's what the `blowouts` variable looks like. Now let's ask R for a `summary()`

```
summary( object = blowouts )
```

```
##      Mode   FALSE    TRUE
## logical     132      44
```

In this context, the `summary()` function gives us a count of the number of `TRUE` values, the number of `FALSE` values, and the number of missing values (i.e., the `NAs`). Pretty reasonable behaviour.

Next, let's try to give it a factor. If you recall, I've defined the `afl.finalists` vector as a factor, so let's use that:

```
summary( object = afl.finalists )
```

```
##          Adelaide      Brisbane      Carlton      Collingwood
##                  26              25              26                  28
##          Essendon      Fitzroy      Fremantle      Geelong
##                  32                  0                  6                  39
##          Hawthorn      Melbourne  North Melbourne      Port Adelaide
##                  27                  28                  28                  17
##          Richmond      St Kilda      Sydney      West Coast
##                  6                  24                  26                  38
##  Western Bulldogs
##                  24
```

For factors, we get a frequency table, just like we got when we used the `table()` function. Interestingly, however, if we convert this to a character vector using the `as.character()` function (see Section ??, we don't get the same results:

```
f2 <- as.character( afl.finalists )
summary( object = f2 )
```

```
##      Length     Class      Mode
##          400 character character
```

This is one of those situations I was referring to in Section 2.22, in which it is helpful to declare your nominal scale variable as a factor rather than a character vector. Because I've defined `afl.finalists` as a factor, R *knows* that it should treat it as a nominal scale variable, and so it gives you a much more detailed (and helpful) summary than it would have if I'd left it as a character vector.

3.6.2 “Summarising” a data frame

Okay what about data frames? When you pass a data frame to the `summary()` function, it produces a slightly condensed summary of each variable inside the data frame. To give you a sense of how this can be useful, let’s try this for a new data set, one that you’ve never seen before. The data is stored in the `clinicaltrial.Rdata` file, and we’ll use it a lot in Chapter ?? (you can find a complete description of the data at the start of that chapter). Let’s load it, and see what we’ve got:

```
load( "./data/clinicaltrial.Rdata" )
who(TRUE)

##      -- Name --      -- Class --      -- Size --
##   clin.trial     data.frame    18 x 3
##   $drug          factor        18
##   $therapy       factor        18
##   $mood.gain    numeric       18
```

There’s a single data frame called `clin.trial` which contains three variables, `drug`, `therapy` and `mood.gain`. Presumably then, this data is from a clinical trial of some kind, in which people were administered different drugs; and the researchers looked to see what the drugs did to their mood. Let’s see if the `summary()` function sheds a little more light on this situation:

```
summary( clin.trial )

##           drug         therapy      mood.gain
## placebo :6  no.therapy:9  Min.   :0.1000
## anxifree:6   CBT       :9  1st Qu.:0.4250
## joyzepam:6
##                               Median :0.8500
##                               Mean   :0.8833
##                               3rd Qu.:1.3000
##                               Max.   :1.8000
```

Evidently there were three drugs: a placebo, something called “anxitfree” and something called “joyzepam”; and there were 6 people administered each drug. There were 9 people treated using cognitive behavioural therapy (CBT) and 9 people who received no psychological treatment. And we can see from looking at the summary of the `mood.gain` variable that most people did show a mood gain (mean = .88), though without knowing what the scale is here it’s hard to say much more than that. Still, that’s not too bad. Overall, I feel that I learned something from that.

3.6.3 “Describing” a data frame

The `describe()` function (in the `psych` package) is a little different, and it’s really only intended to be useful when your data are interval or ratio scale. Unlike the `summary()` function, it calculates the same descriptive statistics for any type of variable you give it. By default, these are:

- `var`. This is just an index: 1 for the first variable, 2 for the second variable, and so on.
- `n`. This is the sample size: more precisely, it’s the number of non-missing values.
- `mean`. This is the sample mean (Section 3.3.1).
- `sd`. This is the (bias corrected) standard deviation (Section 3.4.5).
- `median`. The median (Section 3.3.3).
- `trimmed`. This is trimmed mean. By default it’s the 10% trimmed mean (Section 3.3.6).
- `mad`. The median absolute deviation (Section 3.4.6).
- `min`. The minimum value.
- `max`. The maximum value.
- `range`. The range spanned by the data (Section 3.4.1).
- `skew`. The skewness (Section 3.5).
- `kurtosis`. The kurtosis (Section 3.5).
- `se`. The standard error of the mean (Chapter 4.2).

Notice that these descriptive statistics generally only make sense for data that are interval or ratio scale (usually encoded as numeric vectors). For nominal or ordinal variables (usually encoded as factors), most of these descriptive statistics are not all that useful. What the `describe()` function does is convert factors and logical variables to numeric vectors in order to do the calculations. These variables are marked with `*` and most of the time, the descriptive statistics for those variables won’t make much sense. If you try to feed it a data frame that includes a character vector as a variable, it produces an error.

With those caveats in mind, let’s use the `describe()` function to have a look at the `clin.trial` data frame. Here’s what we get:

```
describe( x = clin.trial )

##           vars   n  mean    sd median trimmed  mad min max range skew
## drug*        1 18  2.00  0.84    2.00    2.00 1.48 1.0 3.0    2.0 0.00
## therapy*     2 18  1.50  0.51    1.50    1.50 0.74 1.0 2.0    1.0 0.00
## mood.gain    3 18  0.88  0.53    0.85    0.88 0.67 0.1 1.8    1.7 0.13
##                  kurtosis   se
## drug*        -1.66  0.20
## therapy*     -2.11  0.12
## mood.gain    -1.44  0.13
```

As you can see, the output for the asterisked variables is pretty meaningless, and should be ignored. However, for the `mood.gain` variable, there's a lot of useful information.

3.7 Descriptive statistics separately for each group

It is very commonly the case that you find yourself needing to look at descriptive statistics, broken down by some grouping variable. This is pretty easy to do in R, and there are three functions in particular that are worth knowing about: `by()`, `describeBy()` and `aggregate()`. Let's start with the `describeBy()` function, which is part of the `psych` package. The `describeBy()` function is very similar to the `describe()` function, except that it has an additional argument called `group` which specifies a grouping variable. For instance, let's say, I want to look at the descriptive statistics for the `clin.trial` data, broken down separately by `therapy` type. The command I would use here is:

```
describeBy( x=clin.trial, group=clin.trial$therapy )

##
##  Descriptive statistics by group
##  group: no.therapy
##          vars n mean   sd median trimmed  mad min max range skew kurtosis
## drug*      1 9 2.00 0.87     2.0    2.00 1.48 1.0 3.0    2.0 0.00   -1.81
## therapy*   2 9 1.00 0.00     1.0    1.00 0.00 1.0 1.0    0.0  NaN     NaN
## mood.gain  3 9 0.72 0.59     0.5    0.72 0.44 0.1 1.7    1.6 0.51   -1.59
##          se
## drug*      0.29
## therapy*   0.00
## mood.gain  0.20
## -----
##  group: CBT
##          vars n mean   sd median trimmed  mad min max range skew
## drug*      1 9 2.00 0.87     2.0    2.00 1.48 1.0 3.0    2.0 0.00
## therapy*   2 9 2.00 0.00     2.0    2.00 0.00 2.0 2.0    0.0  NaN
## mood.gain  3 9 1.04 0.45     1.1    1.04 0.44 0.3 1.8    1.5 -0.03
##          kurtosis se
## drug*      -1.81 0.29
## therapy*    NaN 0.00
## mood.gain  -1.12 0.15
```

As you can see, the output is essentially identical to the output that the `describe()` function produce, except that the output now gives you means,

standard deviations etc separately for the CBT group and the no.therapy group. Notice that, as before, the output displays asterisks for factor variables, in order to draw your attention to the fact that the descriptive statistics that it has calculated won't be very meaningful for those variables. Nevertheless, this command has given us some really useful descriptive statistics mood.gain variable, broken down as a function of therapy.

A somewhat more general solution is offered by the `by()` function. There are three arguments that you need to specify when using this function: the `data` argument specifies the data set, the `INDICES` argument specifies the grouping variable, and the `FUN` argument specifies the name of a function that you want to apply separately to each group. To give a sense of how powerful this is, you can reproduce the `describeBy()` function by using a command like this:

```
by( data=clin.trial, INDICES=clin.trial$therapy, FUN=describe )

## clin.trial$therapy: no.therapy
##          vars n mean   sd median trimmed  mad min max range skew kurtosis
## drug*      1 9 2.00 0.87    2.0    2.00 1.48 1.0 3.0    2.0 0.00 -1.81
## therapy*   2 9 1.00 0.00    1.0    1.00 0.00 1.0 1.0    0.0  NaN   NaN
## mood.gain  3 9 0.72 0.59    0.5    0.72 0.44 0.1 1.7    1.6 0.51 -1.59
##          se
## drug*     0.29
## therapy*  0.00
## mood.gain 0.20
## -----
## clin.trial$therapy: CBT
##          vars n mean   sd median trimmed  mad min max range skew
## drug*      1 9 2.00 0.87    2.0    2.00 1.48 1.0 3.0    2.0 0.00
## therapy*   2 9 2.00 0.00    2.0    2.00 0.00 2.0 2.0    0.0  NaN
## mood.gain  3 9 1.04 0.45    1.1    1.04 0.44 0.3 1.8    1.5 -0.03
##          kurtosis se
## drug*     -1.81 0.29
## therapy*   NaN  0.00
## mood.gain -1.12 0.15
```

This will produce the exact same output as the command shown earlier. However, there's nothing special about the `describe()` function. You could just as easily use the `by()` function in conjunction with the `summary()` function. For example:

```
by( data=clin.trial, INDICES=clin.trial$therapy, FUN=summary )

## clin.trial$therapy: no.therapy
##      drug      therapy      mood.gain
```

```

## placebo :3  no.therapy:9   Min.   :0.1000
## anxifree:3   CBT       :0   1st Qu.:0.3000
## joyzepam:3
##                               Median :0.5000
##                               Mean   :0.7222
##                               3rd Qu.:1.3000
##                               Max.   :1.7000
## -----
## clin.trial$therapy: CBT
##           drug      therapy   mood.gain
## placebo :3  no.therapy:0   Min.   :0.300
## anxifree:3   CBT       :9   1st Qu.:0.800
## joyzepam:3
##                               Median :1.100
##                               Mean   :1.044
##                               3rd Qu.:1.300
##                               Max.   :1.800

```

Again, this output is pretty easy to interpret. It's the output of the `summary()` function, applied separately to CBT group and the `no.therapy` group. For the two factors (`drug` and `therapy`) it prints out a frequency table, whereas for the numeric variable (`mood.gain`) it prints out the range, interquartile range, mean and median.

What if you have multiple grouping variables? Suppose, for example, you would like to look at the average mood gain separately for all possible combinations of drug and therapy. It is actually possible to do this using the `by()` and `describeBy()` functions, but I usually find it more convenient to use the `aggregate()` function in this situation. There are again three arguments that you need to specify. The `formula` argument is used to indicate which variable you want to analyse, and which variables are used to specify the groups. For instance, if you want to look at `mood.gain` separately for each possible combination of `drug` and `therapy`, the formula you want is `mood.gain ~ drug + therapy`. The `data` argument is used to specify the data frame containing all the data, and the `FUN` argument is used to indicate what function you want to calculate for each group (e.g., the `mean`). So, to obtain group means, use this command:

```

aggregate( formula = mood.gain ~ drug + therapy, # mood.gain by drug/therapy combina
            data = clin.trial,                      # data is in the clin.trial data fr
            FUN = mean                                # print out group means
)

##          drug      therapy mood.gain
## 1 placebo no.therapy  0.300000
## 2 anxifree no.therapy  0.400000
## 3 joyzepam no.therapy 1.466667
## 4 placebo      CBT  0.600000

```

```
## 5 anxifree      CBT  1.033333
## 6 joyzepam     CBT  1.500000
```

or, alternatively, if you want to calculate the standard deviations for each group, you would use the following command (argument names omitted this time):

```
aggregate( mood.gain ~ drug + therapy, clin.trial, sd )
```

```
##          drug    therapy mood.gain
## 1 placebo no.therapy 0.2000000
## 2 anxifree no.therapy 0.2000000
## 3 joyzepam no.therapy 0.2081666
## 4 placebo      CBT 0.3000000
## 5 anxifree      CBT 0.2081666
## 6 joyzepam      CBT 0.2645751
```

3.8 Good descriptive statistics are descriptive!

The death of one man is a tragedy. The death of millions is a statistic.

– Josef Stalin, Potsdam 1945

950,000 – 1,200,000

– Estimate of Soviet repression deaths, 1937-1938 (Ellman, 2002)

Stalin’s infamous quote about the statistical character death of millions is worth giving some thought. The clear intent of his statement is that the death of an individual touches us personally and its force cannot be denied, but that the deaths of a multitude are incomprehensible, and as a consequence mere statistics, more easily ignored. I’d argue that Stalin was half right. A statistic is an abstraction, a description of events beyond our personal experience, and so hard to visualise. Few if any of us can imagine what the deaths of millions is “really” like, but we can imagine one death, and this gives the lone death its feeling of immediate tragedy, a feeling that is missing from Ellman’s cold statistical description.

Yet it is not so simple: without numbers, without counts, without a description of what happened, we have *no chance* of understanding what really happened, no opportunity event to try to summon the missing feeling. And in truth, as I write this, sitting in comfort on a Saturday morning, half a world and a whole lifetime away from the Gulags, when I put the Ellman estimate next to the Stalin quote a dull dread settles in my stomach and a chill settles over me.

The Stalinist repression is something truly beyond my experience, but with a combination of statistical data and those recorded personal histories that have come down to us, it is not entirely beyond my comprehension. Because what Ellman's numbers tell us is this: over a two year period, Stalinist repression wiped out the equivalent of every man, woman and child currently alive in the city where I live. Each one of those deaths had its own story, was its own tragedy, and only some of those are known to us now. Even so, with a few carefully chosen statistics, the scale of the atrocity starts to come into focus.

Thus it is no small thing to say that the first task of the statistician and the scientist is to summarise the data, to find some collection of numbers that can convey to an audience a sense of what has happened. This is the job of descriptive statistics, but it's not a job that can be told solely using the numbers. You are a data analyst, not a statistical software package. Part of your job is to take these *statistics* and turn them into a *description*. When you analyse data, it is not sufficient to list off a collection of numbers. Always remember that what you're really trying to do is communicate with a human audience. The numbers are important, but they need to be put together into a meaningful story that your audience can interpret. That means you need to think about framing. You need to think about context. And you need to think about the individual events that your statistics are summarising.

3.9 Drawing graphs

Above all else show the data.

—Edward Tufte¹³

Visualising data is one of the most important tasks facing the data analyst. It's important for two distinct but closely related reasons. Firstly, there's the matter of drawing “presentation graphics”: displaying your data in a clean, visually appealing fashion makes it easier for your reader to understand what you're trying to tell them. Equally important, perhaps even more important, is the fact that drawing graphs helps *you* to understand the data. To that end, it's important to draw “exploratory graphics” that help you learn about the data as you go about analysing it. These points might seem pretty obvious, but I cannot count the number of times I've seen people forget them.

To give a sense of the importance of this chapter, I want to start with a classic illustration of just how powerful a good graph can be. To that end, Figure 3.6 shows a redrawing of one of the most famous data visualisations of all time: John Snow's 1854 map of cholera deaths. The map is elegant in its simplicity. In the background we have a street map, which helps orient the viewer. Over

¹³The origin of this quote is Tufte's lovely book *The Visual Display of Quantitative Information*.

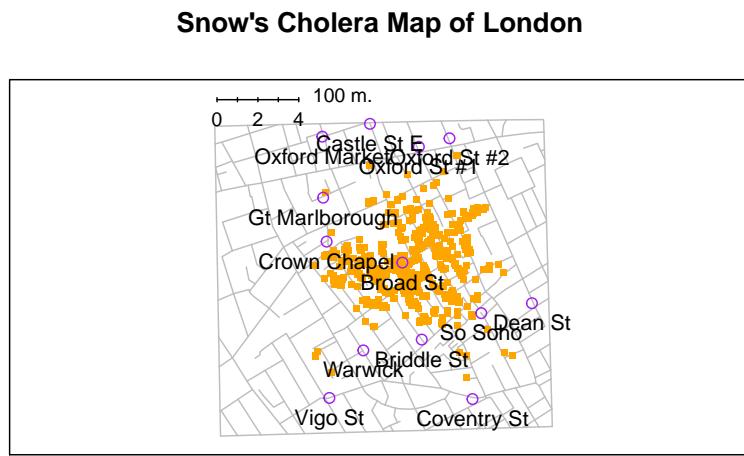


Figure 3.6: A stylised redrawing of John Snow’s original cholera map. Each small dot represents the location of a cholera case, and each large circle shows the location of a well. As the plot makes clear, the cholera outbreak is centred very closely on the Broad St pump. This image uses the data from the `HistData` package, and was drawn using minor alterations to the commands provided in the help files. Note that Snow’s original hand drawn map used different symbols and labels, but you get the idea.

the top, we see a large number of small dots, each one representing the location of a cholera case. The larger symbols show the location of water pumps, labelled by name. Even the most casual inspection of the graph makes it very clear that the source of the outbreak is almost certainly the Broad Street pump. Upon viewing this graph, Dr Snow arranged to have the handle removed from the pump, ending the outbreak that had killed over 500 people. Such is the power of a good data visualisation.

The goals in this chapter are twofold: firstly, to discuss several fairly standard graphs that we use a lot when analysing and presenting data, and secondly, to show you how to create these graphs in R. The graphs themselves tend to be pretty straightforward, so in that respect this chapter is pretty simple. Where people usually struggle is learning how to produce graphs, and especially, learning how to produce good graphs.¹⁴ Fortunately, learning how to draw graphs in R is reasonably simple, as long as you're not too picky about what your graph looks like. What I mean when I say this is that R has a lot of *very* good graphing functions, and most of the time you can produce a clean, high-quality graphic without having to learn very much about the low-level details of how R handles graphics. Unfortunately, on those occasions when you do want to do something non-standard, or if you need to make highly specific changes to the figure, you actually do need to learn a fair bit about the these details; and those details are both complicated and boring. With that in mind, the structure of this chapter is as follows: I'll start out by giving you a very quick overview of how graphics work in R. I'll then discuss several different kinds of graph and how to draw them, as well as showing the basics of how to customise these plots.

Dave note: Navarro (2018) included a fair amount of under-the-hood detail about how graphs are drawn in R. We do not need that level of theory for now. Instead, I'm including just the least you need to quickly visualize your data.

3.9.1 An introduction to plotting

Before I discuss any specialised graphics, let's start by drawing a few very simple graphs just to get a feel for what it's like to draw pictures using R. To that end, let's create a small vector `Fibonacci` that contains a few numbers we'd like R to draw for us. Then, we'll ask R to `plot()` those numbers. The result is Figure 3.7.

¹⁴I should add that this isn't unique to R. Like everything in R there's a pretty steep learning curve to learning how to draw graphs, and like always there's a massive payoff at the end in terms of the quality of what you can produce. But to be honest, I've seen the same problems show up regardless of what system people use. I suspect that the hardest thing to do is to force yourself to take the time to think deeply about what your graphs are doing. I say that in full knowledge that only about half of my graphs turn out as well as they ought to. Understanding what makes a good graph is easy: actually designing a good graph is *hard*.

```
Fibonacci <- c( 1,1,2,3,5,8,13 )
plot( Fibonacci )
```

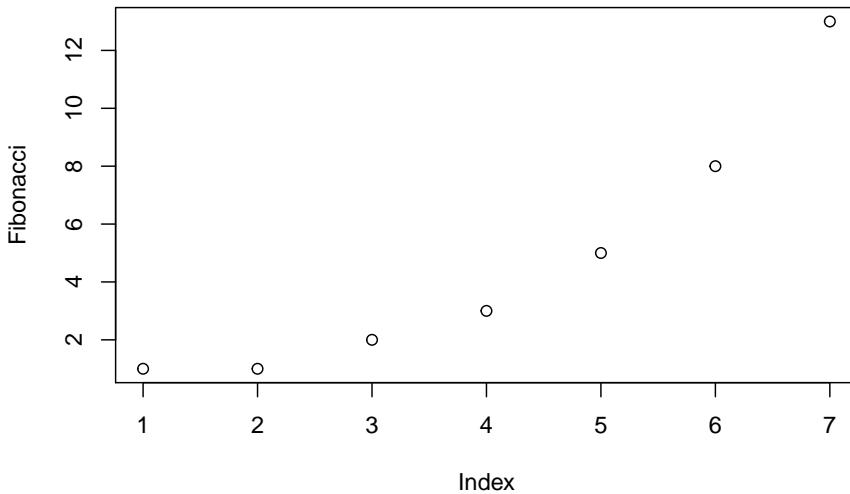


Figure 3.7: Our first plot

As you can see, what R has done is plot the *values* stored in the `Fibonacci` variable on the vertical axis (y-axis) and the corresponding *index* on the horizontal axis (x-axis). In other words, since the 4th element of the vector has a value of 3, we get a dot plotted at the location (4,3). That's pretty straightforward, and the image in Figure 3.7 is probably pretty close to what you would have had in mind when I suggested that we plot the `Fibonacci` data.

3.9.1.1 Customising the title and the axis labels

One of the first things that you'll find yourself wanting to do when customising your plot is to label it better. You might want to specify more appropriate axis labels, add a title or add a subtitle. The arguments that you need to specify to make this happen are:

- `main`. A character string containing the title.
- `sub`. A character string containing the subtitle.
- `xlab`. A character string containing the x-axis label.
- `ylab`. A character string containing the y-axis label.

These aren't graphical parameters, they're arguments to the high-level function. However, because the high-level functions all rely on the same low-level function to do the drawing¹⁵ the names of these arguments are identical for pretty much every high-level function I've come across. Let's have a look at what happens when we make use of all these arguments. Here's the command. The picture that this draws is shown in Figure 3.8.

```
plot( x = Fibonacci,
      main = "You specify title using the 'main' argument",
      sub = "The subtitle appears here! (Use the 'sub' argument for this)",
      xlab = "The x-axis label is 'xlab'",
      ylab = "The y-axis label is 'ylab'"
    )
```

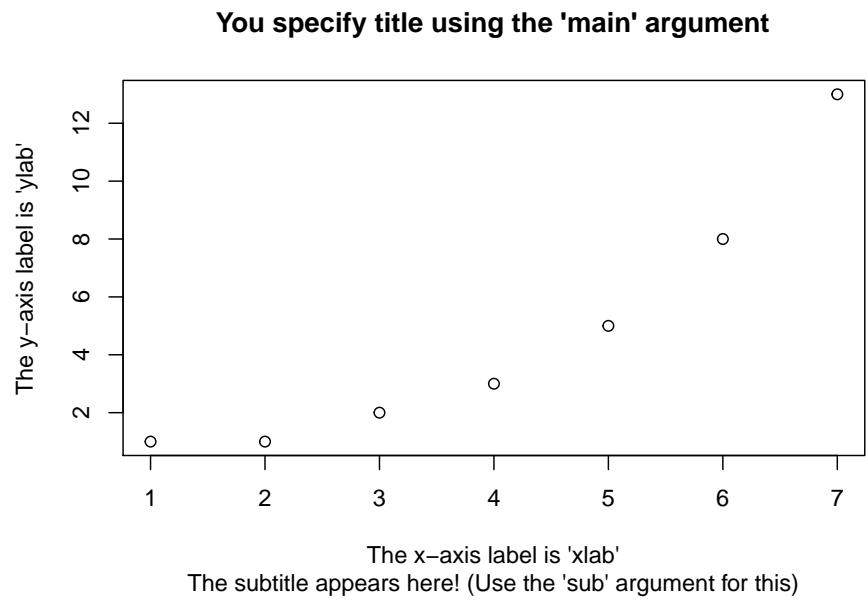


Figure 3.8: How to add your own title, subtitle, x-axis label and y-axis label to the plot.

It's more or less as you'd expect. The plot itself is identical to the one we drew in Figure 3.7, except for the fact that we've changed the axis labels, and added a title and a subtitle. Even so, there's a couple of interesting features worth calling your attention to. Firstly, notice that the subtitle is drawn below the plot, which I personally find annoying; as a consequence I almost never use

¹⁵The low-level function that does this is called `title()` in case you ever need to know, and you can type `?title` to find out a bit more detail about what these arguments do.

subtitles. You may have a different opinion, of course, but the important thing is that you remember where the subtitle actually goes. Secondly, notice that R has decided to use boldface text and a larger font size for the title. This is one of my most hated default settings in R graphics, since I feel that it draws too much attention to the title. Generally, while I do want my reader to look at the title, I find that the R defaults are a bit overpowering, so I often like to change the settings. To that end, there are a bunch of graphical parameters that you can use to customise the font style:

- *Font styles:* `font.main`, `font.sub`, `font.lab`, `font.axis`. These four parameters control the font style used for the plot title (`font.main`), the subtitle (`font.sub`), the axis labels (`font.lab`: note that you can't specify separate styles for the x-axis and y-axis without using low level commands), and the numbers next to the tick marks on the axis (`font.axis`). Somewhat irritatingly, these arguments are numbers instead of meaningful names: a value of 1 corresponds to plain text, 2 means boldface, 3 means italic and 4 means bold italic.
- *Font colours:* `col.main`, `col.sub`, `col.lab`, `col.axis`. These parameters do pretty much what the name says: each one specifies a **colour** in which to type each of the different bits of text. Conveniently, R has a very large number of named colours (type `colours()` to see a list of over 650 colour names that R knows), so you can use the English language name of the colour to select it.¹⁶ Thus, the parameter value here string like "`red`", "`gray25`" or "`springgreen4`" (yes, R really does recognise four different shades of "spring green").
- *Font size:* `cex.main`, `cex.sub`, `cex.lab`, `cex.axis`. Font size is handled in a slightly curious way in R. The "cex" part here is short for "character expansion", and it's essentially a magnification value. By default, all of these are set to a value of 1, except for the font title: `cex.main` has a default magnification of 1.2, which is why the title font is 20% bigger than the others.
- *Font family:* `family`. This argument specifies a font family to use: the simplest way to use it is to set it to "`sans`", "`serif`", or "`mono`", corresponding to a san serif font, a serif font, or a monospaced font. If you want to, you can give the name of a specific font, but keep in mind that different operating systems use different fonts, so it's probably safest to keep it simple. Better yet, unless you have some deep objections to the R defaults, just ignore this parameter entirely. That's what I usually do.

To give you a sense of how you can use these parameters to customise your titles, the following command can be used to draw Figure 3.9:

¹⁶On the off chance that this isn't enough freedom for you, you can select a colour directly as a "red, green, blue" specification using the `rgb()` function, or as a "hue, saturation, value" specification using the `hsv()` function.

```
plot( x = Fibonacci,
      main = "The first 7 Fibonacci numbers", # the title
      xlab = "Position in the sequence",       # x-axis label
      ylab = "The Fibonacci number",          # y-axis
      font.main = 1,
      cex.main = 1,
      font.axis = 2,
      col.lab = "gray50" )
```

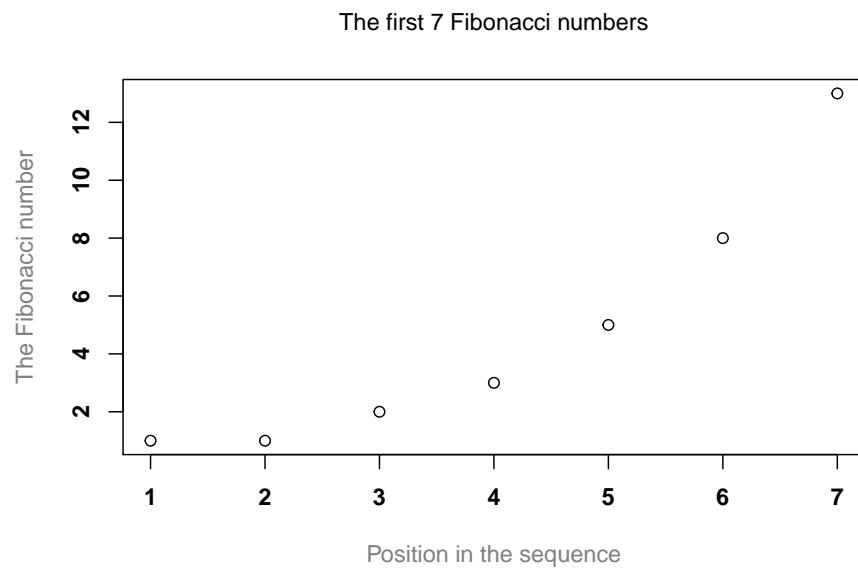


Figure 3.9: How to customise the appearance of the titles and labels.

Although this command is quite long, it's not complicated: all it does is override a bunch of the default parameter values. The only difficult aspect to this is that you have to remember what each of these parameters is called, and what all the different values are. And in practice I never remember: I have to look up the help documentation every time, or else look it up in this book.

3.9.2 Histograms

Now that we've tamed (or possibly fled from) the beast that is R graphical parameters, let's talk more seriously about some real life graphics that you'll want to draw. We begin with the humble *histogram*. Histograms are one of the simplest and most useful ways of visualising data. They make most sense when

you have an interval or ratio scale (e.g., the `afl.margins` data from Chapter 3 and what you want to do is get an overall impression of the data. Most of you probably know how histograms work, since they're so widely used, but for the sake of completeness I'll describe them. All you do is divide up the possible values into **bins**, and then count the number of observations that fall within each bin. This count is referred to as the frequency of the bin, and is displayed as a bar: in the AFL winning margins data, there are 33 games in which the winning margin was less than 10 points, and it is this fact that is represented by the height of the leftmost bar in Figure 3.10. Drawing this histogram in R is pretty straightforward. The function you need to use is called `hist()`, and it has pretty reasonable default settings. In fact, Figure 3.10 is exactly what you get if you just type this:

```
hist( afl.margins )
```

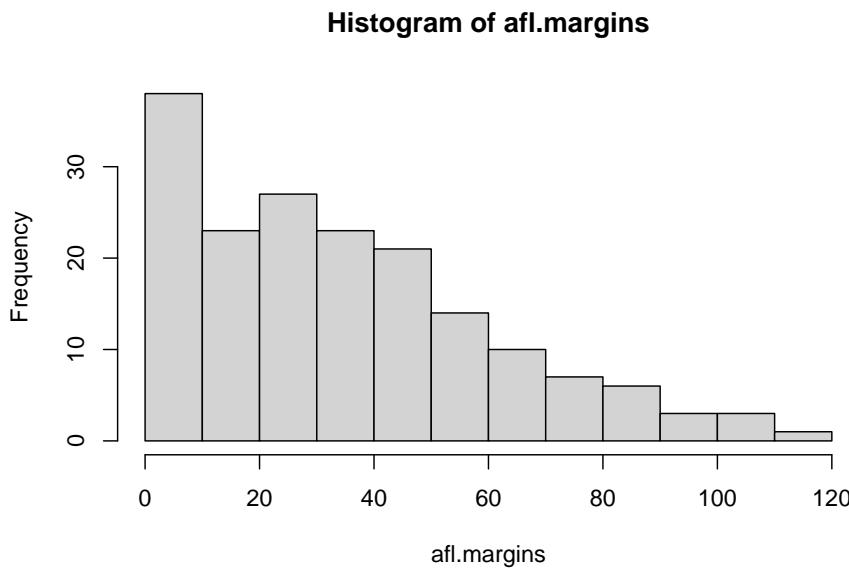


Figure 3.10: The default histogram that R produces

Although this image would need a lot of cleaning up in order to make a good presentation graphic (i.e., one you'd include in a report), it nevertheless does a pretty good job of describing the data. In fact, the big strength of a histogram is that (properly used) it does show the entire spread of the data, so you can get a pretty good sense about what it looks like. The downside to histograms is that they aren't very compact: unlike some of the other plots I'll talk about it's hard to cram 20-30 histograms into a single image without overwhelming the

viewer. And of course, if your data are nominal scale (e.g., the `afl.finalists` data) then histograms are useless.

The main subtlety that you need to be aware of when drawing histograms is determining where the `breaks` that separate bins should be located, and (relatedly) how many breaks there should be. In Figure 3.10, you can see that R has made pretty sensible choices all by itself: the breaks are located at 0, 10, 20, ... 120, which is exactly what I would have done had I been forced to make a choice myself. On the other hand, consider the two histograms in Figure 3.11 and 3.12, which I produced using the following two commands:

```
hist( x = afl.margins, breaks = 3 )
```

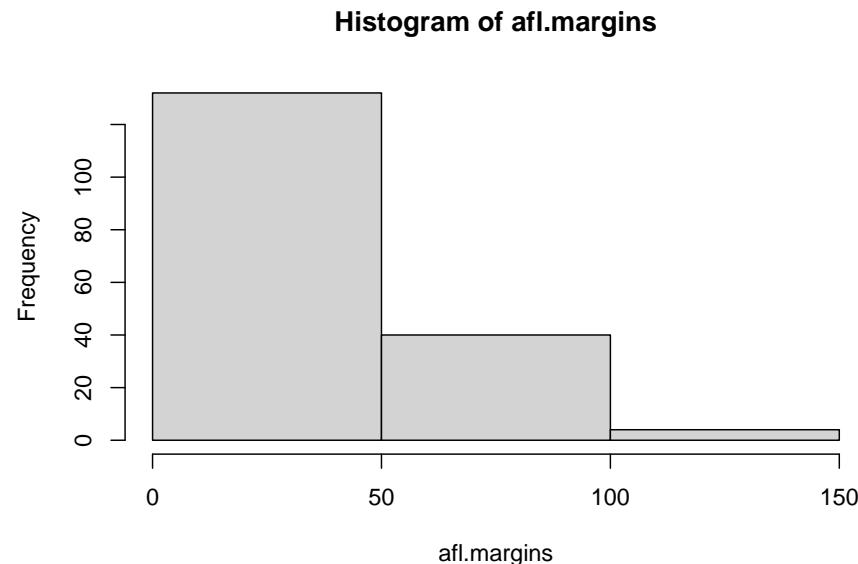


Figure 3.11: A histogram with too few bins

```
hist( x = afl.margins, breaks = 0:116 )
```

In Figure 3.12, the bins are only 1 point wide. As a result, although the plot is very informative (it displays the entire data set with no loss of information at all!) the plot is very hard to interpret, and feels quite cluttered. On the other hand, the plot in Figure 3.11 has a bin width of 50 points, and has the opposite problem: it's very easy to “read” this plot, but it doesn't convey a lot of information. One gets the sense that this histogram is hiding too much. In short, the way in which you specify the `breaks` has a big effect on what

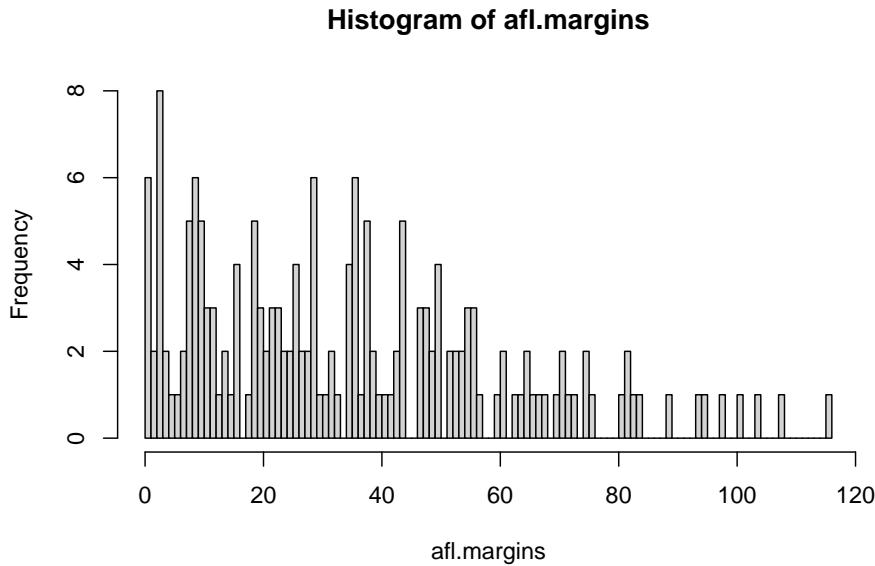


Figure 3.12: A histogram with too many bins

the histogram looks like, so it's important to make sure you choose the breaks sensibly. In general R does a pretty good job of selecting the breaks on its own, since it makes use of some quite clever tricks that statisticians have devised for automatically selecting the right bins for a histogram, but nevertheless it's usually a good idea to play around with the breaks a bit to see what happens.

There is one fairly important thing to add regarding how the `breaks` argument works. There are two different ways you can specify the breaks. You can either specify *how many* breaks you want (which is what I did for panel b when I typed `breaks = 3`) and let R figure out where they should go, or you can provide a vector that tells R exactly where the breaks should be placed (which is what I did for panel c when I typed `breaks = 0:116`). The behaviour of the `hist()` function is slightly different depending on which version you use. If all you do is tell it *how many* breaks you want, R treats it as a “suggestion” not as a demand. It assumes you want “approximately 3” breaks, but if it doesn't think that this would look very pretty on screen, it picks a different (but similar) number. It does this for a sensible reason – it tries to make sure that the breaks are located at sensible values (like 10) rather than stupid ones (like 7.224414). And most of the time R is right: usually, when a human researcher says “give me 3 breaks”, he or she really does mean “give me approximately 3 breaks, and don't put them in stupid places”. However, sometimes R is dead wrong. Sometimes you really do mean “exactly 3 breaks”, and you know precisely where you want them to go. So you need to invoke “real person privilege”, and order R to do what it's

bloody well told. In order to do that, you *have* to input the full vector that tells R exactly where you want the breaks. If you do that, R will go back to behaving like the nice little obedient calculator that it's supposed to be.

3.9.2.1 Visual style of your histogram

Okay, so at this point we can draw a basic histogram, and we can alter the number and even the location of the `breaks`. However, the visual style of the histograms shown in Figures 3.10, 3.11, and 3.12 could stand to be improved. We can fix this by making use of some of the other arguments to the `hist()` function. Most of the things you might want to try doing have already been covered in Section 3.9.1, but there's a few new things:

- *Shading lines:* `density`, `angle`. You can add diagonal lines to shade the bars: the `density` value is a number indicating how many lines per inch R should draw (the default value of `NULL` means no lines), and the `angle` is a number indicating how many degrees from horizontal the lines should be drawn at (default is `angle = 45` degrees).
- *Specifics regarding colours:* `col`, `border`. You can also change the colours: in this instance the `col` parameter sets the colour of the shading (either the shading lines if there are any, or else the colour of the interior of the bars if there are not), and the `border` argument sets the colour of the edges of the bars.
- *Labelling the bars:* `labels`. You can also attach labels to each of the bars using the `labels` argument. The simplest way to do this is to set `labels = TRUE`, in which case R will add a number just above each bar, that number being the exact number of observations in the bin. Alternatively, you can choose the labels yourself, by inputting a vector of strings, e.g., `labels = c("label 1","label 2","etc")`

Not surprisingly, this doesn't exhaust the possibilities. If you type `help("hist")` or `?hist` and have a look at the help documentation for histograms, you'll see a few more options. A histogram that makes use of the histogram-specific customisations as well as several of the options we discussed in Section 3.9.1 is shown in Figure 3.13. The R command that I used to draw it is this:

```
hist( x = afl.margins,
      main = "2010 AFL margins", # title of the plot
      xlab = "Margin",           # set the x-axis label
      density = 10,              # draw shading lines: 10 per inch
      angle = 40,                # set the angle of the shading lines is 40 degrees
      border = "gray20",          # set the colour of the borders of the bars
      col = "gray80",             # set the colour of the shading lines
```

```

    labels = TRUE,
    ylim = c(0,40)           # add frequency labels to each bar
)                           # change the scale of the y-axis

```

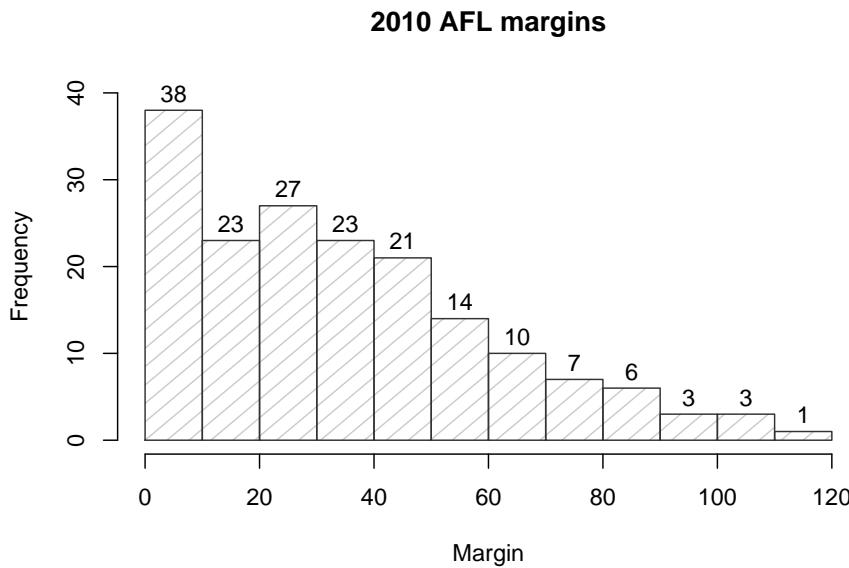


Figure 3.13: A histogram with histogram specific customisations

Overall, this is a much nicer histogram than the default ones.

3.9.3 Boxplots

Another alternative to histograms is a **boxplot**, sometimes called a “box and whiskers” plot. Like histograms, they’re most suited to interval or ratio scale data. The idea behind a boxplot is to provide a simple visual depiction of the median, the interquartile range, and the range of the data. And because they do so in a fairly compact way, boxplots have become a very popular statistical graphic, especially during the exploratory stage of data analysis when you’re trying to understand the data yourself. Let’s have a look at how they work, again using the `afl.margins` data as our example. Firstly, let’s actually calculate these numbers ourselves using the `summary()` function:¹⁷

¹⁷R being what it is, it’s no great surprise that there’s also a `fivenum()` function that does much the same thing.

```
summary( afl.margins )
```

```
##   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##   0.00 12.75 30.50 35.30 50.50 116.00
```

So how does a boxplot capture these numbers? The easiest way to describe what a boxplot looks like is just to draw one. The function for doing this in R is (surprise, surprise) `boxplot()`. As always there's a lot of optional arguments that you can specify if you want, but for the most part you can just let R choose the defaults for you. That said, I'm going to override one of the defaults to start with by specifying the `range` option, but for the most part you won't want to do this (I'll explain why in a minute). With that as preamble, let's try the following command:

```
boxplot( x = afl.margins, range = 100 )
```

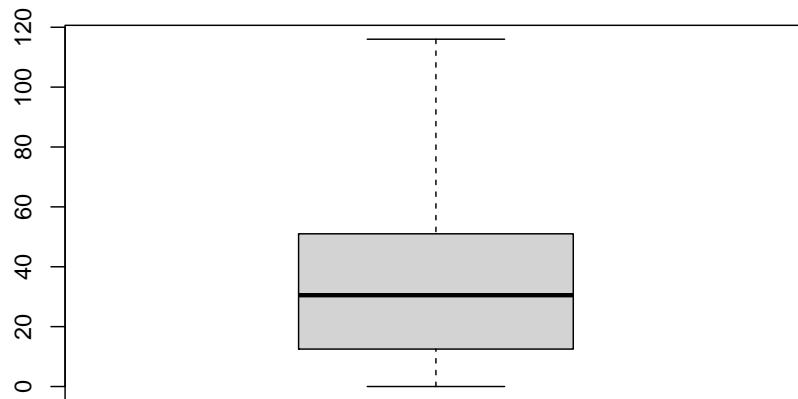


Figure 3.14: A basic boxplot

What R draws is shown in Figure 3.14, the most basic boxplot possible. When you look at this plot, this is how you should interpret it: the thick line in the middle of the box is the median; the box itself spans the range from the 25th percentile to the 75th percentile; and the “whiskers” cover the full range from

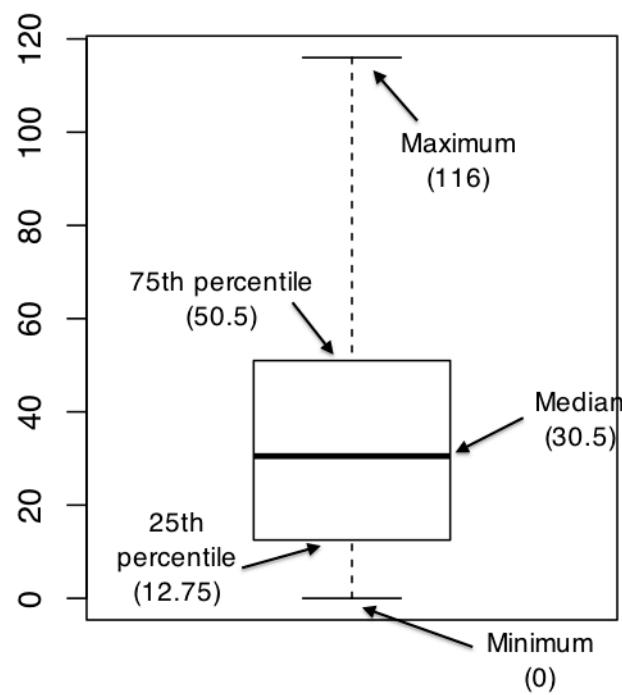


Figure 3.15: An annotated boxplot

the minimum value to the maximum value. This is summarised in the annotated plot in Figure 3.15.

In practice, this isn't quite how boxplots usually work. In most applications, the “whiskers” don't cover the full range from minimum to maximum. Instead, they actually go out to the most extreme data point that doesn't exceed a certain bound. By default, this value is 1.5 times the interquartile range, corresponding to a `range` value of 1.5. Any observation whose value falls outside this range is plotted as a circle instead of being covered by the whiskers, and is commonly referred to as an *outlier*. For our AFL margins data, there is one observation (a game with a margin of 116 points) that falls outside this range. As a consequence, the upper whisker is pulled back to the next largest observation (a value of 108), and the observation at 116 is plotted as a circle. This is illustrated in Figure 3.16. Since the default value is `range = 1.5` we can draw this plot using the simple command

```
boxplot( afl.margins )
```

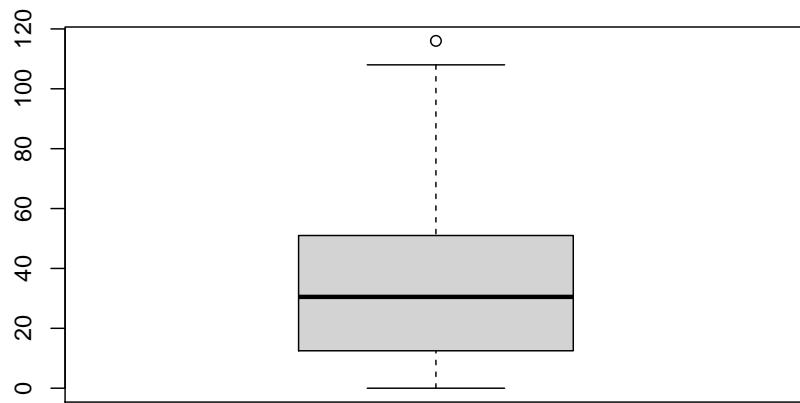


Figure 3.16: By default, R will only extent the whiskers a distance of 1.5 times the interquartile range, and will plot any points that fall outside that range separately

3.9.3.1 Visual style of your boxplot

I'll talk a little more about the relationship between boxplots and outliers in the Section 3.9.3.2, but before I do let's take the time to clean this figure up. Boxplots in R are extremely customisable. In addition to the usual range of graphical parameters that you can tweak to make the plot look nice, you can also exercise nearly complete control over every element to the plot. Consider the boxplot in Figure 3.17: in this version of the plot, not only have I added labels (`xlab`, `ylab`) and removed the stupid border (`frame.plot`), I've also dimmed all of the graphical elements of the boxplot except the central bar that plots the median (`border`) so as to draw more attention to the median rather than the rest of the boxplot.

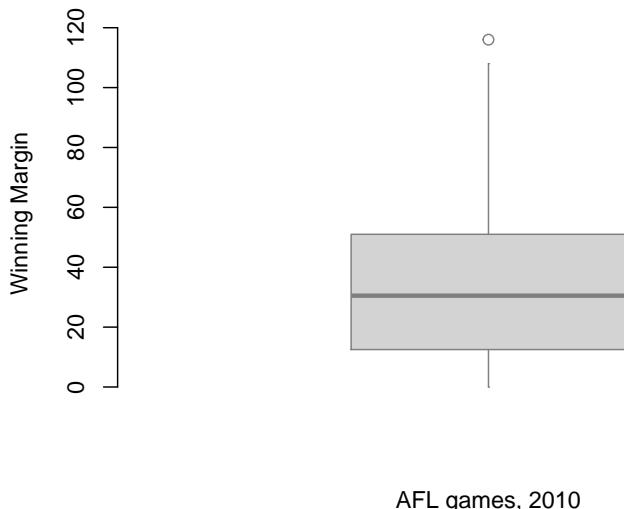


Figure 3.17: A boxplot with boxplot specific customisations

You've seen all these options in previous sections in this chapter, so hopefully those customisations won't need any further explanation. However, I've done two new things as well: I've deleted the cross-bars at the top and bottom of the whiskers (known as the "staples" of the plot), and converted the whiskers themselves to solid lines. The arguments that I used to do this are called by the ridiculous names of `staplewex` and `whisklty`,¹⁸ and I'll explain these in a

¹⁸I realise there's a kind of logic to the way R names are constructed, but they still sound dumb. When I typed this sentence, all I could think was that it sounded like the name of a kids movie if it had been written by Lewis Carroll: "The frabjous gambolles of Staplewex and Whisklty" or something along those lines.

moment.

But first, here's the actual command I used to draw this figure:

```
boxplot( x = afl.margins,           # the data
         xlab = "AFL games, 2010",   # x-axis label
         ylab = "Winning Margin",    # y-axis label
         border = "grey50",          # dim the border of the box
         frame.plot = FALSE,        # don't draw a frame
         staplewex = 0,              # don't draw staples
         whisklty = 1               # solid line for whisker
     )
```

Overall, I think the resulting boxplot is a huge improvement in visual design over the default version. In my opinion at least, there's a fairly minimalist aesthetic that governs good statistical graphics. Ideally, every visual element that you add to a plot should convey part of the message. If your plot includes things that don't actually help the reader learn anything new, you should consider removing them. Personally, I can't see the point of the cross-bars on a standard boxplot, so I've deleted them.

3.9.3.2 Using box plots to detect outliers

Because the boxplot automatically (unless you change the `range` argument) separates out those observations that lie within a certain range, people often use them as an informal method for detecting *outliers*: observations that are “suspiciously” distant from the rest of the data. Here's an example. Suppose that I'd drawn the boxplot for the AFL margins data, and it came up looking like Figure 3.18.

It's pretty clear that something funny is going on with one of the observations. Apparently, there was one game in which the margin was over 300 points! That doesn't sound right to me. Now that I've become suspicious, it's time to look a bit more closely at the data. One function that can be handy for this is the `which()` function; it takes as input a vector of logicals, and outputs the indices of the TRUE cases. This is particularly useful in the current context because it lets me do this:

```
suspicious.cases <- afl.margins > 300
which( suspicious.cases )
```

```
## [1] 137
```

although in real life I probably wouldn't bother creating the `suspicious.cases` variable: I'd just cut out the middle man and use a command like `which(`

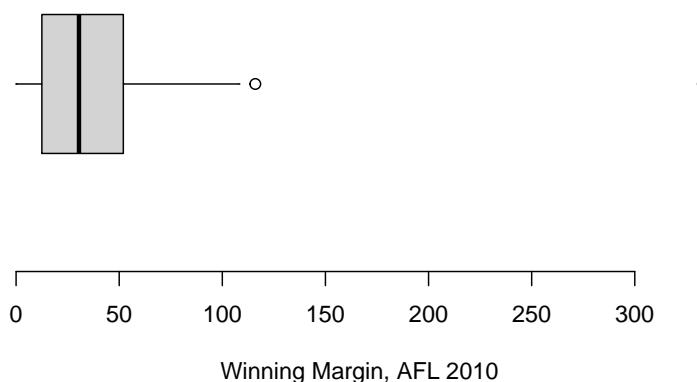


Figure 3.18: A boxplot showing one very suspicious outlier! I've drawn this plot in a similar, minimalist style to the one in Figure 3.17, but I've used the `horizontal` argument to draw it sideways in order to save space.

`afl.margins > 300`). In any case, what this has done is shown me that the outlier corresponds to game 137. Then, I find the recorded margin for that game:

```
afl.margins[137]
```

```
## [1] 333
```

Hm. That definitely doesn't sound right. So then I go back to the original data source (the internet!) and I discover that the actual margin of that game was 33 points. Now it's pretty clear what happened. Someone must have typed in the wrong number. Easily fixed, just by typing `afl.margins[137] <- 33`. While this might seem like a silly example, I should stress that this kind of thing actually happens a lot. Real world data sets are often riddled with stupid errors, especially when someone had to type something into a computer at some point. In fact, there's actually a name for this phase of data analysis, since in practice it can waste a huge chunk of our time: *data cleaning*. It involves searching for typos, missing data and all sorts of other obnoxious errors in raw data files.¹⁹

What about the real data? Does the value of 116 constitute a funny observation not? Possibly. As it turns out the game in question was Fremantle v Hawthorn, and was played in round 21 (the second last home and away round of the season). Fremantle had already qualified for the final series and for them the outcome of the game was irrelevant; and the team decided to rest several of their star players. As a consequence, Fremantle went into the game severely underpowered. In contrast, Hawthorn had started the season very poorly but had ended on a massive winning streak, and for them a win could secure a place in the finals. With the game played on Hawthorn's home turf²⁰ and with so many unusual factors at play, it is perhaps no surprise that Hawthorn annihilated Fremantle by 116 points. Two weeks later, however, the two teams met again in an elimination final on Fremantle's home ground, and Fremantle won comfortably by 30 points.²¹

¹⁹Sometimes it's convenient to have the boxplot automatically label the outliers for you. The original `boxplot()` function doesn't allow you to do this; however, the `Boxplot()` function in the `car` package does. The design of the `Boxplot()` function is very similar to `boxplot()`. It just adds a few new arguments that allow you to tweak the labelling scheme. I'll leave it to the reader to check this out.

²⁰Sort of. The game was played in Launceston, which is a de facto home away from home for Hawthorn.

²¹Contrast this situation with the next largest winning margin in the data set, which was Geelong's 108 point demolition of Richmond in round 6 at their home ground, Kardinia Park. Geelong have been one of the most dominant teams over the last several years, a period during which they strung together an incredible 29-game winning streak at Kardinia Park. Richmond have been useless for several years. This is in no meaningful sense an outlier. Geelong have been winning by these margins (and Richmond losing by them) for quite some time. Frankly I'm surprised that the result wasn't more lopsided: as happened to Melbourne in 2011 when Geelong won by a modest 186 points.

So, should we exclude the game from subsequent analyses? If this were a psychology experiment rather than an AFL season, I'd be quite tempted to exclude it because there's pretty strong evidence that Fremantle weren't really trying very hard: and to the extent that my research question is based on an assumption that participants are genuinely trying to do the task. On the other hand, in a lot of studies we're actually interested in seeing the full range of possible behaviour, and that includes situations where people decide not to try very hard: so excluding that observation would be a bad idea. In the context of the AFL data, a similar distinction applies. If I'd been trying to make tips about who would perform well in the finals, I would have (and in fact did) disregard the Round 21 massacre, because it's way too misleading. On the other hand, if my interest is solely in the home and away season itself, I think it would be a shame to throw away information pertaining to one of the most distinctive (if boring) games of the year. In other words, the decision about whether to include outliers or exclude them depends heavily on *why* you think the data look they way they do, and what you want to use the data *for*. Statistical tools can provide an automatic method for suggesting candidates for deletion, but you really need to exercise good judgment here. As I've said before, R is a mindless automaton. It doesn't watch the footy, so it lacks the broader context to make an informed decision. You are *not* a mindless automaton, so you should exercise judgment: if the outlier looks legitimate to you, then keep it. In any case, I'll return to the topic again in Section 8.13, so let's return to our discussion of how to draw boxplots.

3.9.3.3 Drawing multiple boxplots

One last thing. What if you want to draw multiple boxplots at once? Suppose, for instance, I wanted separate boxplots showing the AFL margins not just for 2010, but for every year between 1987 and 2010. To do that, the first thing we'll have to do is find the data. These are stored in the `aflsmall2.Rdata` file. So let's load it and take a quick peek at what's inside:

```
load( "aflsmall2.Rdata" )
who( TRUE )
#   -- Name --    -- Class --    -- Size --
#   afl2          data.frame    4296 x 2
#   $margin        numeric      4296
#   $year          numeric      4296
```

Notice that `afl2` data frame is pretty big. It contains 4296 games, which is far more than I want to see printed out on my computer screen. To that end, R provides you with a few useful functions to print out only a few of the row in the data frame. The first of these is `head()` which prints out the first 6 rows, of the data frame, like this:

```
head( afl2 )

##   margin year
## 1     33 1987
## 2     59 1987
## 3     45 1987
## 4     91 1987
## 5     39 1987
## 6      1 1987
```

You can also use the `tail()` function to print out the last 6 rows. The `car` package also provides a handy little function called `some()` which prints out a random subset of the rows.

In any case, the important thing is that we have the `afl2` data frame which contains the variables that we're interested in. What we want to do is have R draw boxplots for the `margin` variable, plotted separately for each separate `year`. The way to do this using the `boxplot()` function is to input a `formula` rather than a variable as the input. In this case, the formula we want is `margin ~ year`. So our boxplot command now looks like this. The result is shown in Figure 3.19.²²

```
boxplot( formula = margin ~ year,
         data = afl2
)
```

Even this, the default version of the plot, gives a sense of why it's sometimes useful to choose boxplots instead of histograms. Even before taking the time to turn this basic output into something more readable, it's possible to get a good sense of what the data look like from year to year without getting overwhelmed with too much detail. Now imagine what would have happened if I'd tried to cram 24 histograms into this space: no chance at all that the reader is going to learn anything useful.

That being said, the default boxplot leaves a great deal to be desired in terms of visual clarity. The outliers are too visually prominent, the dotted lines look messy, and the interesting content (i.e., the behaviour of the median and the interquartile range across years) gets a little obscured. Fortunately, this is easy

²²Actually, there's other ways to do this. If the input argument `x` is a list object (see Section 2.24, the `boxplot()` function will draw a separate boxplot for each variable in that list. Relatedly, since the `plot()` function – which we'll discuss shortly – is a generic (see Section 2.26, you might not be surprised to learn that one of its special cases is a boxplot: specifically, if you use `plot()` where the first argument `x` is a factor and the second argument `y` is numeric, then the result will be a boxplot, showing the values in `y`, with a separate boxplot for each level. For instance, something like `plot(x = afl2$year, y = afl2$margin)` would work.

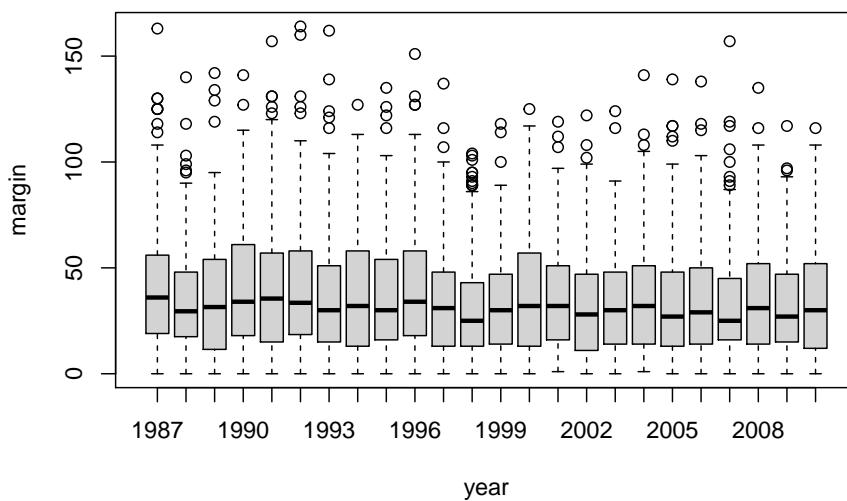


Figure 3.19: Boxplots showing the AFL winning margins for the 24 years from 1987 to 2010 inclusive. This is the default plot created by R, with no annotations added and no changes to the visual design. It's pretty readable, though at a minimum you'd want to include some basic annotations labelling the axes. Compare and contrast with Figure 3.20

to fix, since we've already covered a lot of tools you can use to customise your output. After playing around with several different versions of the plot, the one I settled on is shown in Figure 3.20. The command I used to produce it is long, but not complicated:

```
boxplot( formula = margin ~ year,    # the formula
         data = afl2,                 # the data set
         xlab = "AFL season",        # x axis label
         ylab = "Winning Margin",    # y axis label
         frame.plot = FALSE,         # don't draw a frame
         staplewex = 0,               # don't draw staples
         staplecol = "white",        # (fixes a tiny display issue)
         boxwex = .75,                # narrow the boxes slightly
         boxfill = "grey80",          # lightly shade the boxes
         whisklty = 1,                # solid line for whiskers
         whiskcol = "grey70",         # dim the whiskers
         boxcol = "grey70",           # dim the box borders
         outcol = "grey70",           # dim the outliers
         outpch = 20,                  # outliers as solid dots
         outcex = .5,                  # shrink the outliers
         medlty = "blank",            # no line for the medians
         medpch = 20,                  # instead, draw solid dots
         medlwd = 1.5                 # make them larger
     )
```

Of course, given that the command is that long, you might have guessed that I didn't spend ages typing all that rubbish in over and over again. Instead, I wrote a script, which I kept tweaking until it produced the figure that I wanted. We'll talk about scripts later in Section ??, but given the length of the command I thought I'd remind you that there's an easier way of trying out different commands than typing them all in over and over.

3.9.4 Bar graphs

Another form of graph that you often want to plot is the *bar graph*. The main function that you can use in R to draw them is the `barplot()` function.²³ And to illustrate the use of the function, I'll use the `finalists` variable that I introduced in Section 3.3.7. What I want to do is draw a bar graph that displays the number of finals that each team has played in over the time spanned by the `afl` data set. So, let's start by creating a vector that contains this information. I'll use the `tabulate()` function to do this (which will be discussed properly in Section ??, since it creates a simple numeric vector:

²³Once again, it's worth noting the link to the generic `plot()` function. If the `x` argument to `plot()` is a factor (and no `y` argument is given), the result is a bar graph. So you could use `plot(afl.finalists)` and get the same output as `barplot(afl.finalists)`.

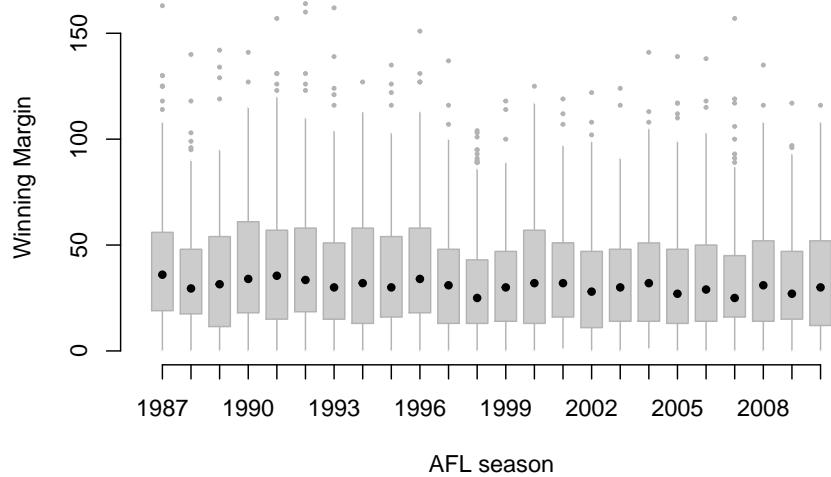


Figure 3.20: A cleaned up version of Figure 3.19. Notice that I've used a very minimalist design for the boxplots, so as to focus the eye on the medians. I've also converted the medians to solid dots, to convey a sense that year to year variation in the median should be thought of as a single coherent plot (similar to what we did when plotting the `Fibonacci` variable earlier). The size of outliers has been shrunk, because they aren't actually very interesting. In contrast, I've added a fill colour to the boxes, to make it easier to look at the trend in the interquartile range across years.

```

freq <- tabulate( afl.finalists )
print( freq )

## [1] 26 25 26 28 32  0  6 39 27 28 28 17  6 24 26 38 24

```

This isn't exactly the prettiest of frequency tables, of course. I'm only doing it this way so that you can see the `barplot()` function in its "purest" form: when the input is just an ordinary numeric vector. That being said, I'm obviously going to need the team names to create some labels, so let's create a variable with those. I'll do this using the `levels()` function, which outputs the names of all the levels of a factor (see Section 2.22):

```

teams <- levels( afl.finalists )
print( teams )

## [1] "Adelaide"          "Brisbane"           "Carlton"
## [4] "Collingwood"        "Essendon"            "Fitzroy"
## [7] "Fremantle"          "Geelong"             "Hawthorn"
## [10] "Melbourne"          "North Melbourne"    "Port Adelaide"
## [13] "Richmond"           "St Kilda"            "Sydney"
## [16] "West Coast"         "Western Bulldogs"

```

Okay, so now that we have the information we need, let's draw our bar graph. The main argument that you need to specify for a bar graph is the `height` of the bars, which in our case correspond to the values stored in the `freq` variable:

```

barplot( height = freq )  # specifying the argument name
barplot( freq )           # the lazier version

```

Either of these two commands will produce the simple bar graph shown in Figure 3.21.

As you can see, R has drawn a pretty minimal plot. It doesn't have any labels, obviously, because we didn't actually tell the `barplot()` function what the labels are! To do this, we need to specify the `names.arg` argument. The `names.arg` argument needs to be a vector of character strings containing the text that needs to be used as the label for each of the items. In this case, the `teams` vector is exactly what we need, so the command we're looking for is:

```
barplot( height = freq, names.arg = teams )
```

This is an improvement, but not much of an improvement. R has only included a few of the labels, because it can't fit them in the plot. This is the same behaviour

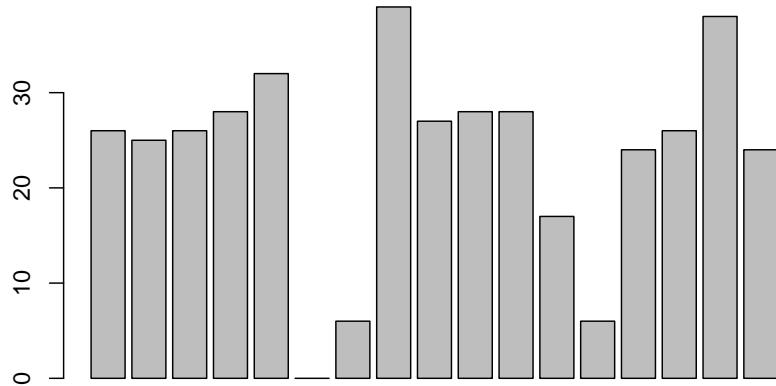


Figure 3.21: the simplest version of a bargraph, containing the data but no labels

we saw earlier with the multiple-boxplot graph in Figure 3.19. However, in Figure 3.19 it wasn't an issue: it's pretty obvious from inspection that the two unlabelled plots in between 1987 and 1990 must correspond to the data from 1988 and 1989. However, the fact that `barplot()` has omitted the names of every team in between Adelaide and Fitzroy is a lot more problematic.

The simplest way to fix this is to rotate the labels, so that the text runs vertically not horizontally. To do this, we need to alter set the `las` parameter, which I discussed briefly in Section 3.9.1. What I'll do is tell R to rotate the text so that it's always perpendicular to the axes (i.e., I'll set `las = 2`). When I do that, as per the following command...

```
barplot(height = freq, # the frequencies
        names.arg = teams, # the label
        las = 2)           # rotate the labels
```

... the result is the bar graph shown in Figure 3.23. We've fixed the problem, but we've created a new one: the axis labels don't quite fit anymore. To fix this, we have to be a bit cleverer again. A simple fix would be to use shorter names rather than the full name of all teams, and in many situations that's probably the right thing to do. However, at other times you really do need to create a bit more space to add your labels, so I'll show you how to do that.

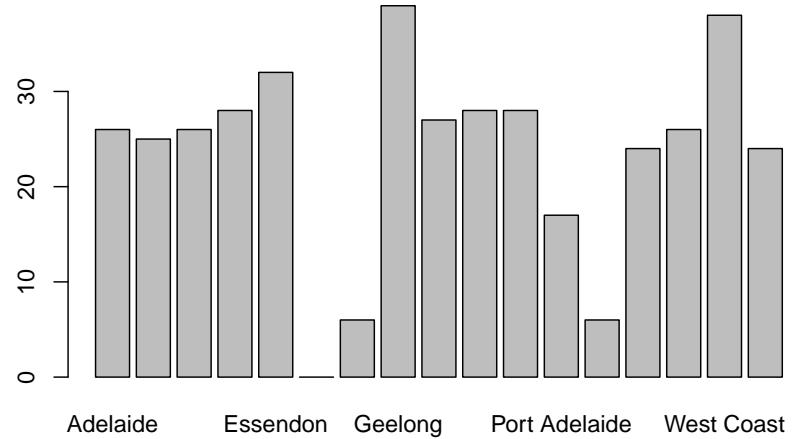


Figure 3.22: we've added the labels, but because the text runs horizontally R only includes a few of them

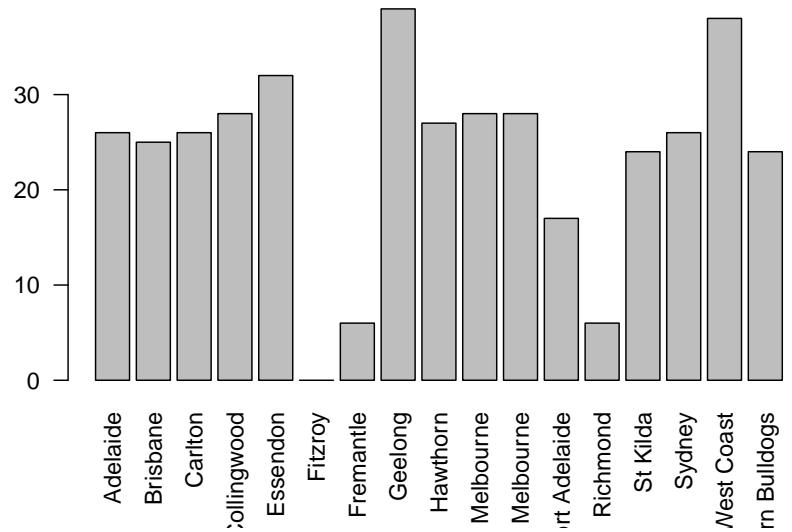


Figure 3.23: we've rotated the labels, but now the text is too long to fit

3.9.5 Saving image files using R and Rstudio

Hold on, you might be thinking. What's the good of being able to draw pretty pictures in R if I can't save them and send them to friends to brag about how awesome my data is? How do I save the picture? This is another one of those situations where the easiest thing to do is to use the RStudio tools.

If you're running R through Rstudio, then the easiest way to save your image is to click on the "Export" button in the Plot panel (i.e., the area in Rstudio where all the plots have been appearing). When you do that you'll see a menu that contains the options "Save Plot as PDF" and "Save Plot as Image". Either version works. Both will bring up dialog boxes that give you a few options that you can play with, but besides that it's pretty simple.

This works pretty nicely for most situations. So, unless you're filled with a burning desire to learn the low level details, feel free to skip the rest of this section.

3.9.6 Summary

Calculating some basic descriptive statistics is one of the very first things you do when analysing real data, and descriptive statistics are much simpler to understand than inferential statistics, so like every other statistics textbook I've started with descriptives. In this chapter, we talked about the following topics:

- *Measures of central tendency.* Broadly speaking, central tendency measures tell you where the data are. There's three measures that are typically reported in the literature: the mean, median and mode. (Section 3.3)
- *Measures of variability.* In contrast, measures of variability tell you about how "spread out" the data are. The key measures are: range, standard deviation, interquartile reange (Section 3.4)
- *Getting summaries of variables in R.* Since this book focuses on doing data analysis in R, we spent a bit of time talking about how descriptive statistics are computed in R. (Section 2.14 and 3.7)
- *Basic overview to R graphics.* In Section ?? we talked about how graphics in R are organised, and then moved on to the basics of how they're drawn in Section 3.9.1.
- *Common plots.* Much of the chapter was focused on standard graphs that statisticians like to produce: histograms (Section 3.9.2), boxplots (Section 3.9.3), and bar graphs (Section 3.9.4).

A traditional first course in statistics spends only a small proportion of the class on descriptive statistics, maybe one or two lectures at most. The vast majority of the lecturer's time is spent on inferential statistics, because that's

where all the hard stuff is. That makes sense, but it hides the practical everyday importance of choosing good descriptives.

Dave note: With this chapter, I condensed two of Navarro (2018)'s chapters: descriptive stats and visualizations. For our course, the tools we need at this moment are the ones that help us describe single variables. We will introduce additional tools for summarizing and visualizing bivariate (two variable) data a bit later in the course, when we need them. And, if you do not feel comfortable with every customization option presented here, do not be concerned. For now, if you can generate histograms, bar charts, and the like, you have what you need. We will learn more about customizing these graphs for presentations later, as well.

Chapter 4

Inferential statistics: The Central Limit Theorem

4.1 Videos

Video: Are you a Bayesian or a Frequentist?

Video: Sampling Methods

Video: Estimation and Confidence Intervals

Video: The central limit theorem

4.2 Introduction

Text by Navarro (2018)

The role of descriptive statistics is to concisely summarise what we *do* know. In contrast, the purpose of inferential statistics is to “learn what we do not know from what we do”. We are in a good position to think about the problem of statistical inference. What kinds of things would we like to learn about? And how do we learn them? These are the questions that lie at the heart of inferential statistics, and they are traditionally divided into two “big ideas”: estimation and hypothesis testing. The goal in this chapter is to introduce the first of these big ideas, estimation theory, but I’m going to witter on about probability and sampling theory first because estimation theory doesn’t make sense until you understand probability and sampling . As a consequence, this chapter divides naturally into two parts Sections 4.5 through 4.7 are focused on sampling theory, and Sections 4.8 and 4.9 make use of sampling theory to discuss how statisticians think about estimation.

4.3 How are probability and statistics different?

Text by Navarro (2018) Before we start talking about probability theory, it's helpful to spend a moment thinking about the relationship between probability and statistics. The two disciplines are closely related but they're not identical. Probability theory is "the doctrine of chances". It's a branch of mathematics that tells you how often different kinds of events will happen. For example, all of these questions are things you can answer using probability theory:

- What are the chances of a fair coin coming up heads 10 times in a row?
- If I roll two six sided dice, how likely is it that I'll roll two sixes?
- How likely is it that five cards drawn from a perfectly shuffled deck will all be hearts?
- What are the chances that I'll win the lottery?

Notice that all of these questions have something in common. In each case the "truth of the world" is known, and my question relates to the "what kind of events" will happen. In the first question I *know* that the coin is fair, so there's a 50% chance that any individual coin flip will come up heads. In the second question, I *know* that the chance of rolling a 6 on a single die is 1 in 6. In the third question I *know* that the deck is shuffled properly. And in the fourth question, I *know* that the lottery follows specific rules. You get the idea. The critical point is that probabilistic questions start with a known ***model*** of the world, and we use that model to do some calculations. The underlying model can be quite simple. For instance, in the coin flipping example, we can write down the model like this:

$$P(\text{heads}) = 0.5$$

which you can read as "the probability of heads is 0.5". As we'll see later, in the same way that percentages are numbers that range from 0% to 100%, probabilities are just numbers that range from 0 to 1. When using this probability model to answer the first question, I don't actually know exactly what's going to happen. Maybe I'll get 10 heads, like the question says. But maybe I'll get three heads. That's the key thing: in probability theory, the *model* is known, but the *data* are not.

So that's probability. What about statistics? Statistical questions work the other way around. In statistics, we *do not* know the truth about the world. All we have is the data, and it is from the data that we want to *learn* the truth about the world. Statistical questions tend to look more like these:

- If my friend flips a coin 10 times and gets 10 heads, are they playing a trick on me?
- If five cards off the top of the deck are all hearts, how likely is it that the deck was shuffled? - If the lottery commissioner's spouse wins the lottery, how likely is it that the lottery was rigged?

This time around, the only thing we have are data. What I *know* is that I saw my friend flip the coin 10 times and it came up heads every time. And what I want to *infer* is whether or not I should conclude that what I just saw was actually a fair coin being flipped 10 times in a row, or whether I should suspect that my friend is playing a trick on me. The data I have look like this:

H H H H H H H H H H

and what I'm trying to do is work out which “model of the world” I should put my trust in. If the coin is fair, then the model I should adopt is one that says that the probability of heads is 0.5; that is, $P(\text{heads}) = 0.5$. If the coin is not fair, then I should conclude that the probability of heads is *not* 0.5, which we would write as $P(\text{heads}) \neq 0.5$. In other words, the statistical inference problem is to figure out which of these probability models is right. Clearly, the statistical question isn’t the same as the probability question, but they’re deeply connected to one another. Because of this, a good introduction to statistical theory will start with a discussion of what probability is and how it works.

4.4 What does probability mean?

Text by Navarro (2018)

Let’s start with the first of these questions. What is “probability”? It might seem surprising to you, but while statisticians and mathematicians (mostly) agree on what the *rules* of probability are, there’s much less of a consensus on what the word really *means*. It seems weird because we’re all very comfortable using words like “chance”, “likely”, “possible” and “probable”, and it doesn’t seem like it should be a very difficult question to answer. If you had to explain “probability” to a five year old, you could do a pretty good job. But if you’ve ever had that experience in real life, you might walk away from the conversation feeling like you didn’t quite get it right, and that (like many everyday concepts) it turns out that you don’t *really* know what it’s all about.

So I’ll have a go at it. Let’s suppose I want to bet on a soccer game between two teams of robots, *Arduino Arsenal* and *C Milan*. After thinking about it, I decide that there is an 80% probability that *Arduino Arsenal* winning. What do I mean by that? Here are three possibilities...

- They’re robot teams, so I can make them play over and over again, and if I did that, *Arduino Arsenal* would win 8 out of every 10 games on average.
- For any given game, I would only agree that betting on this game is only “fair” if a \$1 bet on *C Milan* gives a \$5 payoff (i.e. I get my \$1 back plus a \$4 reward for being correct), as would a \$4 bet on *Arduino Arsenal* (i.e., my \$4 bet plus a \$1 reward).

- My subjective “belief” or “confidence” in an *Arduino Arsenal* victory is four times as strong as my belief in a *C Milan* victory.

Each of these seems sensible. However they’re not identical, and not every statistician would endorse all of them. The reason is that there are different statistical ideologies (yes, really!) and depending on which one you subscribe to, you might say that some of those statements are meaningless or irrelevant. In this section, I give a brief introduction the two main approaches that exist in the literature. These are by no means the only approaches, but they’re the two big ones.

4.4.1 The frequentist view

The first of the two major approaches to probability, and the more dominant one in statistics, is referred to as the *frequentist view*, and it defines probability as a *long-run frequency*. Suppose we were to try flipping a fair coin, over and over again. By definition, this is a coin that has $P(H) = 0.5$. What might we observe? One possibility is that the first 20 flips might look like this:

T, H, H, H, H, T, T, H, H, H, H, T, H, H, T, T, T, T, T, H

In this case 11 of these 20 coin flips (55%) came up heads. Now suppose that I’d been keeping a running tally of the number of heads (which I’ll call N_H) that I’ve seen, across the first N flips, and calculate the proportion of heads N_H/N every time. Here’s what I’d get (I did literally flip coins to produce this!):

number.of.flips	number.of.heads	proportion
1	0	0.00
2	1	0.50
3	2	0.67
4	3	0.75
5	4	0.80
6	4	0.67
7	4	0.57
8	5	0.63
9	6	0.67
10	7	0.70
11	8	0.73
12	8	0.67
13	9	0.69
14	10	0.71
15	10	0.67
16	10	0.63
17	10	0.59
18	10	0.56
19	10	0.53
20	11	0.55

Notice that at the start of the sequence, the *proportion* of heads fluctuates wildly, starting at .00 and rising as high as .80. Later on, one gets the impression that it dampens out a bit, with more and more of the values actually being pretty close to the “right” answer of .50. This is the frequentist definition of probability in a nutshell: flip a fair coin over and over again, and as N grows large (approaches infinity, denoted $N \rightarrow \infty$), the proportion of heads will converge to 50%. There are some subtle technicalities that the mathematicians care about, but qualitatively speaking, that’s how the frequentists define probability. Unfortunately, I don’t have an infinite number of coins, or the infinite patience required to flip a coin an infinite number of times. However, I do have a computer, and computers excel at mindless repetitive tasks. So I asked my computer to simulate flipping a coin 1000 times, and then drew a picture of what happens to the proportion N_H/N as N increases. Actually, I did it four times, just to make sure it wasn’t a fluke. The results are shown in Figure 4.1. As you can see, the *proportion of observed heads* eventually stops fluctuating, and settles down; when it does, the number at which it finally settles is the true probability of heads.

The frequentist definition of probability has some desirable characteristics. Firstly, it is objective: the probability of an event is *necessarily* grounded in the world. The only way that probability statements can make sense is if they refer to (a sequence of) events that occur in the physical universe.¹ Secondly, it

¹This doesn’t mean that frequentists can’t make hypothetical statements, of course; it’s just that if you want to make a statement about probability, then it must be possible to

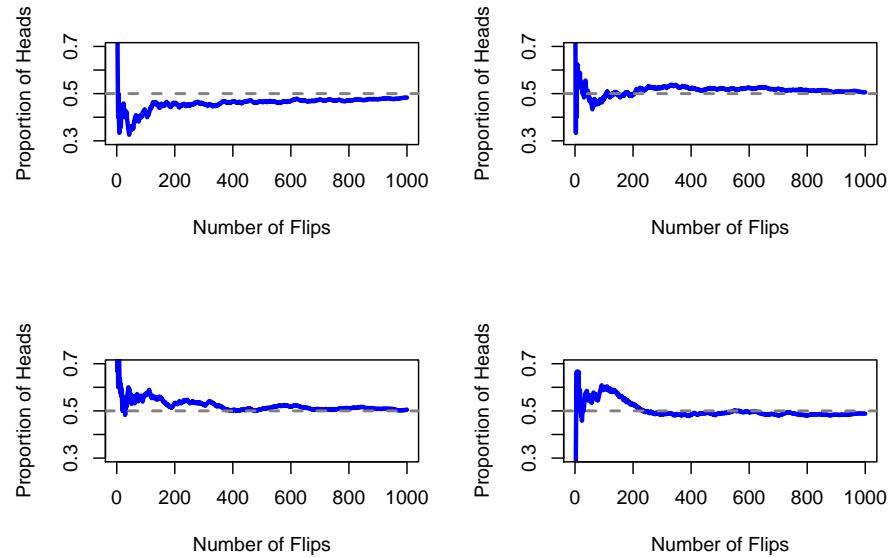


Figure 4.1: An illustration of how frequentist probability works. If you flip a fair coin over and over again, the proportion of heads that you've seen eventually settles down, and converges to the true probability of 0.5. Each panel shows four different simulated experiments: in each case, we pretend we flipped a coin 1000 times, and kept track of the proportion of flips that were heads as we went along. Although none of these sequences actually ended up with an exact value of .5, if we'd extended the experiment for an infinite number of coin flips they would have.

is unambiguous: any two people watching the same sequence of events unfold, trying to calculate the probability of an event, must inevitably come up with the same answer. However, it also has undesirable characteristics. Firstly, infinite sequences don't exist in the physical world. Suppose you picked up a coin from your pocket and started to flip it. Every time it lands, it impacts on the ground. Each impact wears the coin down a bit; eventually, the coin will be destroyed. So, one might ask whether it really makes sense to pretend that an "infinite" sequence of coin flips is even a meaningful concept, or an objective one. We can't say that an "infinite sequence" of events is a real thing in the physical universe, because the physical universe doesn't allow infinite anything. More seriously, the frequentist definition has a narrow scope. There are lots of things out there that human beings are happy to assign probability to in everyday language, but cannot (even in theory) be mapped onto a hypothetical sequence of events. For instance, if a meteorologist comes on TV and says, "the probability of rain in Adelaide on 2 November 2048 is 60%" we humans are happy to accept this. But it's not clear how to define this in frequentist terms. There's only one city of Adelaide, and only 2 November 2048. There's no infinite sequence of events here, just a once-off thing. Frequentist probability genuinely *forbids* us from making probability statements about a single event. From the frequentist perspective, it will either rain tomorrow or it will not; there is no "probability" that attaches to a single non-repeatable event. Now, it should be said that there are some very clever tricks that frequentists can use to get around this. One possibility is that what the meteorologist means is something like this: "There is a category of days for which I predict a 60% chance of rain; if we look only across those days for which I make this prediction, then on 60% of those days it will actually rain". It's very weird and counterintuitive to think of it this way, but you do see frequentists do this sometimes. And it *will* come up later in this book (see Section 4.9).

4.4.2 The Bayesian view

The *Bayesian view* of probability is often called the subjectivist view, and it is a minority view among statisticians, but one that has been steadily gaining traction for the last several decades. There are many flavours of Bayesianism, making hard to say exactly what "the" Bayesian view is. The most common way of thinking about subjective probability is to define the probability of an event as the *degree of belief* that an intelligent and rational agent assigns to that truth of that event. From that perspective, probabilities don't exist in the world, but rather in the thoughts and assumptions of people and other intelligent beings. However, in order for this approach to work, we need some way of operationalising "degree of belief". One way that you can do this is to formalise it in terms of "rational gambling", though there are many other ways. Suppose that I believe that there's a 60% probability of rain tomorrow.

redescribe that statement in terms of a sequence of potentially observable events, and the relative frequencies of different outcomes that appear within that sequence.

If someone offers me a bet: if it rains tomorrow, then I win \$5, but if it doesn't rain then I lose \$5. Clearly, from my perspective, this is a pretty good bet. On the other hand, if I think that the probability of rain is only 40%, then it's a bad bet to take. Thus, we can operationalise the notion of a "subjective probability" in terms of what bets I'm willing to accept.

What are the advantages and disadvantages to the Bayesian approach? The main advantage is that it allows you to assign probabilities to any event you want to. You don't need to be limited to those events that are repeatable. The main disadvantage (to many people) is that we can't be purely objective – specifying a probability requires us to specify an entity that has the relevant degree of belief. This entity might be a human, an alien, a robot, or even a statistician, but there has to be an intelligent agent out there that believes in things. To many people this is uncomfortable: it seems to make probability arbitrary. While the Bayesian approach does require that the agent in question be rational (i.e., obey the rules of probability), it does allow everyone to have their own beliefs; I can believe the coin is fair and you don't have to, even though we're both rational. The frequentist view doesn't allow any two observers to attribute different probabilities to the same event: when that happens, then at least one of them must be wrong. The Bayesian view does not prevent this from occurring. Two observers with different background knowledge can legitimately hold different beliefs about the same event. In short, where the frequentist view is sometimes considered to be too narrow (forbids lots of things that we want to assign probabilities to), the Bayesian view is sometimes thought to be too broad (allows too many differences between observers).

4.4.3 What's the difference? And who is right?

Now that you've seen each of these two views independently, it's useful to make sure you can compare the two. Go back to the hypothetical robot soccer game at the start of the section. What do you think a frequentist and a Bayesian would say about these three statements? Which statement would a frequentist say is the correct definition of probability? Which one would a Bayesian do? Would some of these statements be meaningless to a frequentist or a Bayesian? If you've understood the two perspectives, you should have some sense of how to answer those questions.

Okay, assuming you understand the different, you might be wondering which of them is *right*? Honestly, I don't know that there is a right answer. As far as I can tell there's nothing mathematically incorrect about the way frequentists think about sequences of events, and there's nothing mathematically incorrect about the way that Bayesians define the beliefs of a rational agent. In fact, when you dig down into the details, Bayesians and frequentists actually agree about a lot of things. Many frequentist methods lead to decisions that Bayesians agree a rational agent would make. Many Bayesian methods have very good frequentist properties.

For the most part, I'm a pragmatist so I'll use any statistical method that I trust. As it turns out, that makes me prefer Bayesian methods, for reasons I'll explain towards the end of the book, but I'm not fundamentally opposed to frequentist methods. Not everyone is quite so relaxed. For instance, consider Sir Ronald Fisher, one of the towering figures of 20th century statistics and a vehement opponent to all things Bayesian, whose paper on the mathematical foundations of statistics referred to Bayesian probability as "an impenetrable jungle [that] arrests progress towards precision of statistical concepts" Fisher (1922). Or the psychologist Paul Meehl, who suggests that relying on frequentist methods could turn you into "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring" Meehl (1967). The history of statistics, as you might gather, is not devoid of entertainment.

In any case, while I personally prefer the Bayesian view, the majority of statistical analyses are based on the frequentist approach. My reasoning is pragmatic: the goal of this book is to cover roughly the same territory as a typical undergraduate stats class in psychology, and if you want to understand the statistical tools used by most psychologists, you'll need a good grasp of frequentist methods. I promise you that this isn't wasted effort. Even if you end up wanting to switch to the Bayesian perspective, you really should read through at least one book on the "orthodox" frequentist view. And since R is the most widely used statistical language for Bayesians, you might as well read a book that uses R. Besides, I won't completely ignore the Bayesian perspective. Every now and then I'll add some commentary from a Bayesian point of view.

4.5 Samples, populations and sampling

Text by Navarro (2018)

In the prelude to Part I discussed the riddle of induction, and highlighted the fact that *all* learning requires you to make assumptions. Accepting that this is true, our first task to come up with some fairly general assumptions about data that make sense. This is where **sampling theory** comes in. If probability theory is the foundations upon which all statistical theory builds, sampling theory is the frame around which you can build the rest of the house. Sampling theory plays a huge role in specifying the assumptions upon which your statistical inferences rely. And in order to talk about "making inferences" the way statisticians think about it, we need to be a bit more explicit about what it is that we're drawing inferences *from* (the sample) and what it is that we're drawing inferences *about* (the population).

In almost every situation of interest, what we have available to us as researchers is a **sample** of data. We might have run experiment with some number of participants; a polling company might have phoned some number of people to ask questions about voting intentions; etc. Regardless: the data set available to us is finite, and incomplete. We can't possibly get every person in the world to

do our experiment; a polling company doesn't have the time or the money to ring up every voter in the country etc. In our earlier discussion of descriptive statistics (Chapter 3, this sample was the only thing we were interested in. Our only goal was to find ways of describing, summarising and graphing that sample. This is about to change.

4.5.1 Defining a population

A sample is a concrete thing. You can open up a data file, and there's the data from your sample. A *population*, on the other hand, is a more abstract idea. It refers to the set of all possible people, or all possible observations, that you want to draw conclusions about, and is generally *much* bigger than the sample. In an ideal world, the researcher would begin the study with a clear idea of what the population of interest is, since the process of designing a study and testing hypotheses about the data that it produces does depend on the population about which you want to make statements. However, that doesn't always happen in practice: usually the researcher has a fairly vague idea of what the population is and designs the study as best he/she can on that basis.

Sometimes it's easy to state the population of interest. For instance, in the "polling company" example that opened the chapter, the population consisted of all voters enrolled at the a time of the study – millions of people. The sample was a set of 1000 people who all belong to that population. In most situations the situation is much less simple. In a typical a psychological experiment, determining the population of interest is a bit more complicated. Suppose I run an experiment using 100 undergraduate students as my participants. My goal, as a cognitive scientist, is to try to learn something about how the mind works. So, which of the following would count as "the population":

- All of the undergraduate psychology students at the University of Adelaide?
- Undergraduate psychology students in general, anywhere in the world?
- Australians currently living?
- Australians of similar ages to my sample?
- Anyone currently alive?
- Any human being, past, present or future?
- Any biological organism with a sufficient degree of intelligence operating in a terrestrial environment?
- Any intelligent being?

Each of these defines a real group of mind-possessing entities, all of which might be of interest to me as a cognitive scientist, and it's not at all clear which one ought to be the true population of interest. As another example, consider the Wellesley-Croker game that we discussed in the prelude. The sample here is a specific sequence of 12 wins and 0 losses for Wellesley. What is the population?

- All outcomes until Wellesley and Croker arrived at their destination?
- All outcomes if Wellesley and Croker had played the game for the rest of their lives?
- All outcomes if Wellseley and Croker lived forever and played the game until the world ran out of hills?
- All outcomes if we created an infinite set of parallel universes and the Wellesely/Croker pair made guesses about the same 12 hills in each universe?

Again, it's not obvious what the population is.

4.5.2 Simple random samples

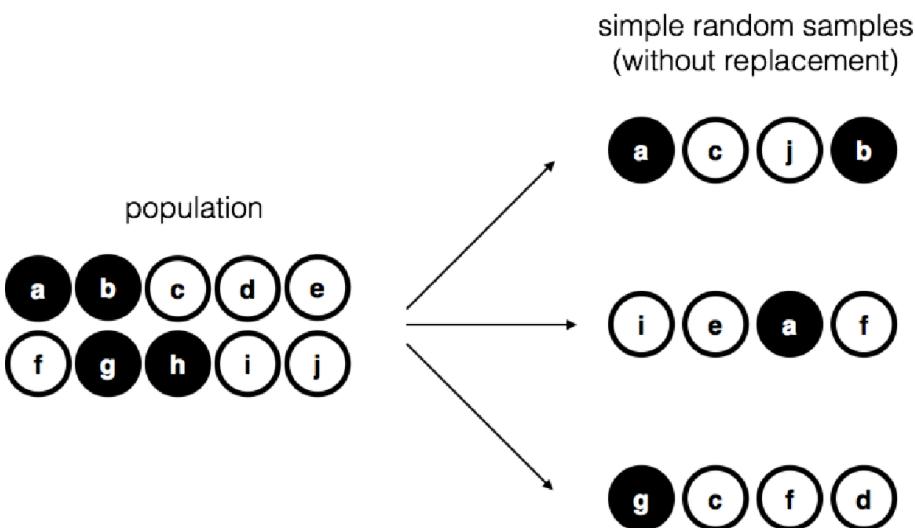


Figure 4.2: Simple random sampling without replacement from a finite population

Irrespective of how I define the population, the critical point is that the sample is a subset of the population, and our goal is to use our knowledge of the sample to draw inferences about the properties of the population. The relationship between the two depends on the *procedure* by which the sample was selected. This procedure is referred to as a ***sampling method***, and it is important to understand why it matters.

To keep things simple, let's imagine that we have a bag containing 10 chips. Each chip has a unique letter printed on it, so we can distinguish between the 10 chips. The chips come in two colours, black and white. This set of chips is the population of interest, and it is depicted graphically on the left of Figure 4.2.

As you can see from looking at the picture, there are 4 black chips and 6 white chips, but of course in real life we wouldn't know that unless we looked in the bag. Now imagine you run the following "experiment": you shake up the bag, close your eyes, and pull out 4 chips without putting any of them back into the bag. First out comes the a chip (black), then the c chip (white), then j (white) and then finally b (black). If you wanted, you could then put all the chips back in the bag and repeat the experiment, as depicted on the right hand side of Figure 4.2. Each time you get different results, but the procedure is identical in each case. The fact that the same procedure can lead to different results each time, we refer to it as a *random* process.² However, because we shook the bag before pulling any chips out, it seems reasonable to think that every chip has the same chance of being selected. A procedure in which every member of the population has the same chance of being selected is called a ***simple random sample***. The fact that we did *not* put the chips back in the bag after pulling them out means that you can't observe the same thing twice, and in such cases the observations are said to have been sampled ***without replacement***.

To help make sure you understand the importance of the sampling procedure, consider an alternative way in which the experiment could have been run. Suppose that my 5-year old son had opened the bag, and decided to pull out four black chips without putting any of them back in the bag. This *biased* sampling scheme is depicted in Figure 4.3. Now consider the evidentiary value of seeing 4 black chips and 0 white chips. Clearly, it depends on the sampling scheme, does it not? If you know that the sampling scheme is biased to select only black chips, then a sample that consists of only black chips doesn't tell you very much about the population! For this reason, statisticians really like it when a data set can be considered a simple random sample, because it makes the data analysis *much* easier.

A third procedure is worth mentioning. This time around we close our eyes, shake the bag, and pull out a chip. This time, however, we record the observation and then put the chip back in the bag. Again we close our eyes, shake the bag, and pull out a chip. We then repeat this procedure until we have 4 chips. Data sets generated in this way are still simple random samples, but because we put the chips back in the bag immediately after drawing them it is referred to as a sample ***with replacement***. The difference between this situation and the first one is that it is possible to observe the same population member multiple times, as illustrated in Figure 4.4.

In my experience, most psychology experiments tend to be sampling without replacement, because the same person is not allowed to participate in the experiment twice. However, most statistical theory is based on the assumption that the data arise from a simple random sample *with* replacement. In real life, this

²The proper mathematical definition of randomness is extraordinarily technical, and way beyond the scope of this book. We'll be non-technical here and say that a process has an element of randomness to it whenever it is possible to repeat the process and get different answers each time.

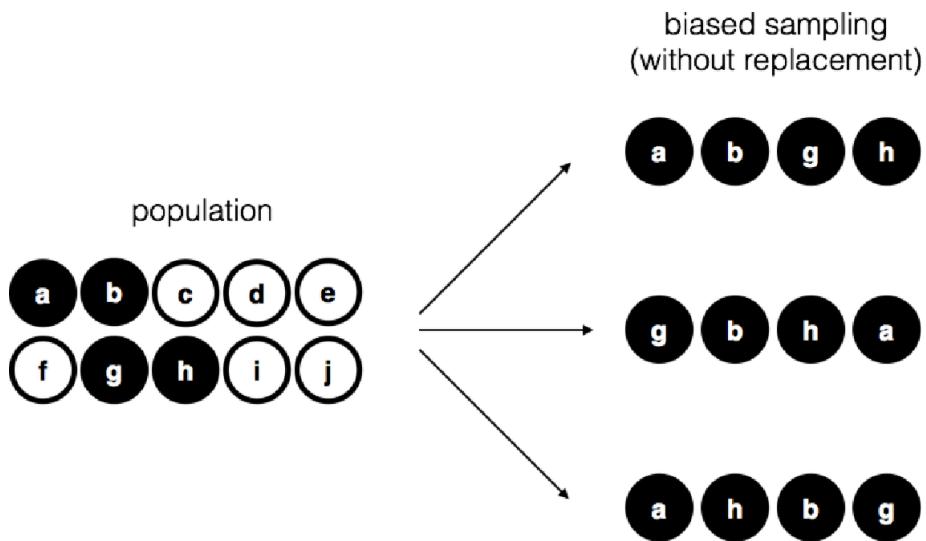
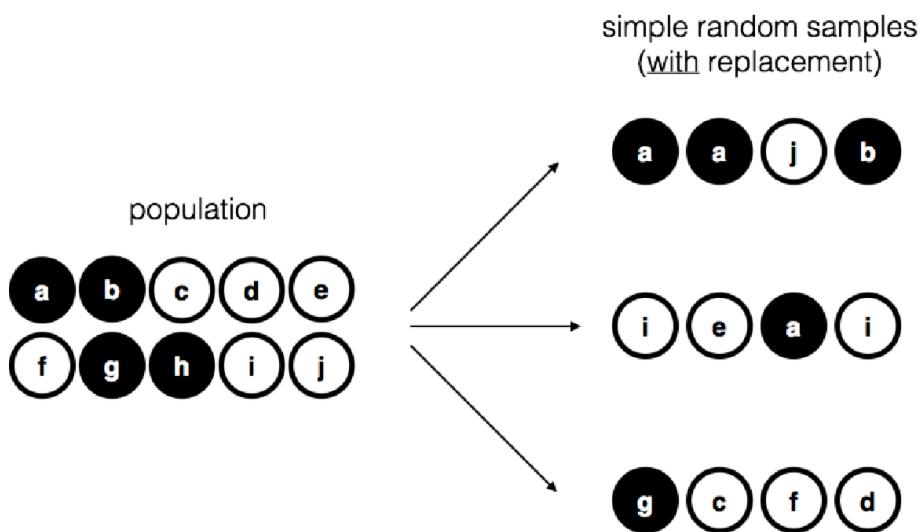


Figure 4.3: Biased sampling without replacement from a finite population

Figure 4.4: Simple random sampling *with* replacement from a finite population

very rarely matters. If the population of interest is large (e.g., has more than 10 entities!) the difference between sampling with- and without- replacement is too small to be concerned with. The difference between simple random samples and biased samples, on the other hand, is not such an easy thing to dismiss.

4.5.3 Most samples are not simple random samples

As you can see from looking at the list of possible populations that I showed above, it is almost impossible to obtain a simple random sample from most populations of interest. When I run experiments, I'd consider it a minor miracle if my participants turned out to be a random sampling of the undergraduate psychology students at Adelaide university, even though this is by far the narrowest population that I might want to generalise to. A thorough discussion of other types of sampling schemes is beyond the scope of this book, but to give you a sense of what's out there I'll list a few of the more important ones:

- *Stratified sampling.* Suppose your population is (or can be) divided into several different subpopulations, or *strata*. Perhaps you're running a study at several different sites, for example. Instead of trying to sample randomly from the population as a whole, you instead try to collect a separate random sample from each of the strata. Stratified sampling is sometimes easier to do than simple random sampling, especially when the population is already divided into the distinct strata. It can also be more efficient than simple random sampling, especially when some of the subpopulations are rare. For instance, when studying schizophrenia it would be much better to divide the population into two³ strata (schizophrenic and not-schizophrenic), and then sample an equal number of people from each group. If you selected people randomly, you would get so few schizophrenic people in the sample that your study would be useless. This specific kind of stratified sampling is referred to as *oversampling* because it makes a deliberate attempt to over-represent rare groups.
- *Snowball sampling* is a technique that is especially useful when sampling from a “hidden” or hard to access population, and is especially common in social sciences. For instance, suppose the researchers want to conduct an opinion poll among transgender people. The research team might only have contact details for a few trans folks, so the survey starts by asking them to participate (stage 1). At the end of the survey, the participants are asked to provide contact details for other people who might want to participate. In stage 2, those new contacts are surveyed. The process continues until the researchers have sufficient data. The big advantage to snowball sampling is that it gets you data in situations that might otherwise be impossible to get any. On the statistical side, the main

³Nothing in life is that simple: there's not an obvious division of people into binary categories like “schizophrenic” and “not schizophrenic”. But this isn't a clinical psychology text, so please forgive me a few simplifications here and there.

disadvantage is that the sample is highly non-random, and non-random in ways that are difficult to address. On the real life side, the disadvantage is that the procedure can be unethical if not handled well, because hidden populations are often hidden for a reason. I chose transgender people as an example here to highlight this: if you weren't careful you might end up outing people who don't want to be outed (very, very bad form), and even if you don't make that mistake it can still be intrusive to use people's social networks to study them. It's certainly very hard to get people's informed consent *before* contacting them, yet in many cases the simple act of contacting them and saying "hey we want to study you" can be hurtful. Social networks are complex things, and just because you can use them to get data doesn't always mean you should.

- *Convenience sampling* is more or less what it sounds like. The samples are chosen in a way that is convenient to the researcher, and not selected at random from the population of interest. Snowball sampling is one type of convenience sampling, but there are many others. A common example in psychology are studies that rely on undergraduate psychology students. These samples are generally non-random in two respects: firstly, reliance on undergraduate psychology students automatically means that your data are restricted to a single subpopulation. Secondly, the students usually get to pick which studies they participate in, so the sample is a self selected subset of psychology students not a randomly selected subset. In real life, most studies are convenience samples of one form or another. This is sometimes a severe limitation, but not always.

Dave here, adding two more sampling methods:

- *Systematic sampling*: Starting from a random point, select every Nth participant.
- *Cluster sampling*: Divide population into clusters or units (such as schools), take a random sample of the clusters (i.e., randomly select a school) and then measure all the participants within the cluster (i.e., measure every student in the school).

4.5.4 How much does it matter if you don't have a simple random sample?

Okay, so real world data collection tends not to involve nice simple random samples. Does that matter? A little thought should make it clear to you that it *can* matter if your data are not a simple random sample: just think about the difference between Figures 4.2 and 4.3. However, it's not quite as bad as it sounds. Some types of biased samples are entirely unproblematic. For instance, when using a stratified sampling technique you actually *know* what the bias is because you created it deliberately, often to *increase* the effectiveness of your

study, and there are statistical techniques that you can use to adjust for the biases you've introduced (not covered in this book!). So in those situations it's not a problem.

More generally though, it's important to remember that random sampling is a means to an end, not the end in itself. Let's assume you've relied on a convenience sample, and as such you can assume it's biased. A bias in your sampling method is only a problem if it causes you to draw the wrong conclusions. When viewed from that perspective, I'd argue that we don't need the sample to be randomly generated in *every* respect: we only need it to be random with respect to the psychologically-relevant phenomenon of interest. Suppose I'm doing a study looking at working memory capacity. In study 1, I actually have the ability to sample randomly from all human beings currently alive, with one exception: I can only sample people born on a Monday. In study 2, I am able to sample randomly from the Australian population. I want to generalise my results to the population of all living humans. Which study is better? The answer, obviously, is study 1. Why? Because we have no reason to think that being "born on a Monday" has any interesting relationship to working memory capacity. In contrast, I can think of several reasons why "being Australian" might matter. Australia is a wealthy, industrialised country with a very well-developed education system. People growing up in that system will have had life experiences much more similar to the experiences of the people who designed the tests for working memory capacity. This shared experience might easily translate into similar beliefs about how to "take a test", a shared assumption about how psychological experimentation works, and so on. These things might actually matter. For instance, "test taking" style might have taught the Australian participants how to direct their attention exclusively on fairly abstract test materials relative to people that haven't grown up in a similar environment; leading to a misleading picture of what working memory capacity is.

There are two points hidden in this discussion. Firstly, when designing your own studies, it's important to think about what population you care about, and try hard to sample in a way that is appropriate to that population. In practice, you're usually forced to put up with a "sample of convenience" (e.g., psychology lecturers sample psychology students because that's the least expensive way to collect data, and our coffers aren't exactly overflowing with gold), but if so you should at least spend some time thinking about what the dangers of this practice might be.

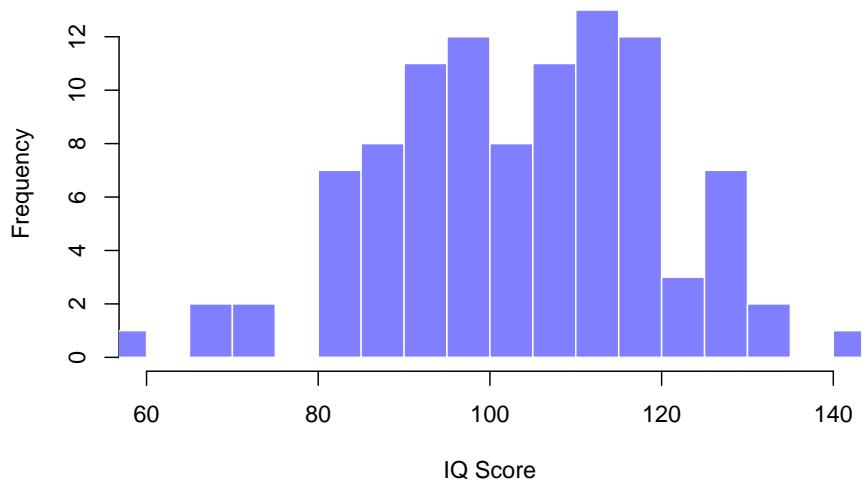
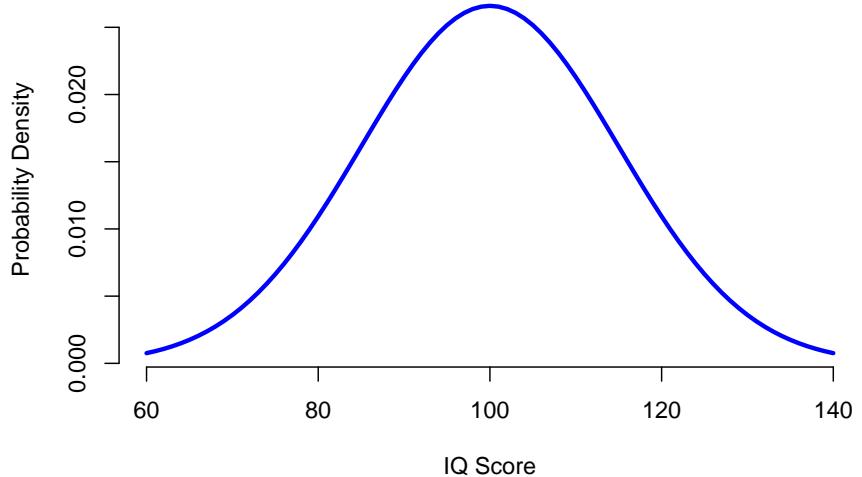
Secondly, if you're going to criticise someone else's study because they've used a sample of convenience rather than laboriously sampling randomly from the entire human population, at least have the courtesy to offer a specific theory as to *how* this might have distorted the results. Remember, everyone in science is aware of this issue, and does what they can to alleviate it. Merely pointing out that "the study only included people from group BLAH" is entirely unhelpful, and borders on being insulting to the researchers, who are *of course* aware of the issue. They just don't happen to be in possession of the infinite supply of time

and money required to construct the perfect sample. In short, if you want to offer a responsible critique of the sampling process, then be *helpful*. Rehashing the blindingly obvious truisms that I've been rambling on about in this section isn't helpful.

4.5.5 Population parameters and sample statistics

Okay. Setting aside the thorny methodological issues associated with obtaining a random sample and my rather unfortunate tendency to rant about lazy methodological criticism, let's consider a slightly different issue. Up to this point we have been talking about populations the way a scientist might. To a psychologist, a population might be a group of people. To an ecologist, a population might be a group of bears. In most cases the populations that scientists care about are concrete things that actually exist in the real world. Statisticians, however, are a funny lot. On the one hand, they *are* interested in real world data and real science in the same way that scientists are. On the other hand, they also operate in the realm of pure abstraction in the way that mathematicians do. As a consequence, statistical theory tends to be a bit abstract in how a population is defined. In much the same way that psychological researchers operationalise our abstract theoretical ideas in terms of concrete measurements (Section 1.2, statisticians operationalise the concept of a “population” in terms of mathematical objects that they know how to work with. They’re called probability distributions.

The idea is quite simple. Let's say we're talking about IQ scores. To a psychologist, the population of interest is a group of actual humans who have IQ scores. A statistician “simplifies” this by operationally defining the population as the probability distribution depicted in Figure ???. IQ tests are designed so that the average IQ is 100, the standard deviation of IQ scores is 15, and the distribution of IQ scores is normal. These values are referred to as the ***population parameters*** because they are characteristics of the entire population. That is, we say that the population mean μ is 100, and the population standard deviation σ is 15.



```
## [1] "n= 100 mean= 103.521042167666 sd= 16.0541514183256"
```

```
## [1] "n= 10000 mean= 99.8766302966644 sd= 14.6307623833259"
```

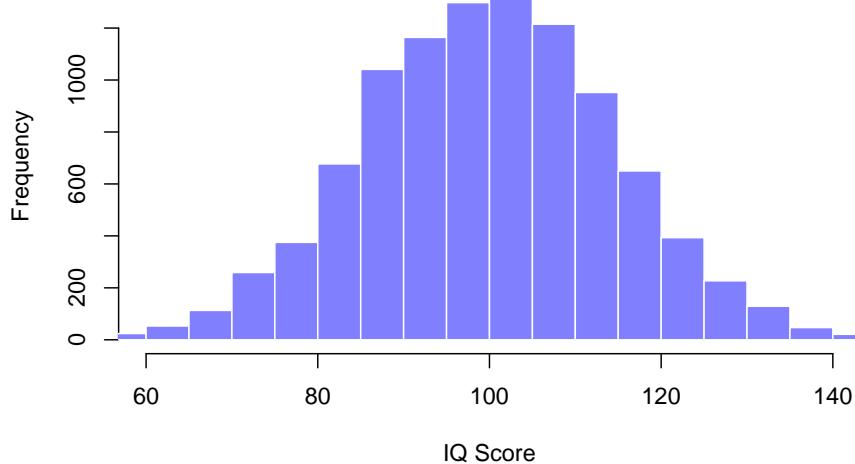


Figure 4.5: The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and panel c we have a sample of 10,000 observations.

Now suppose I run an experiment. I select 100 people at random and administer an IQ test, giving me a simple random sample from the population. My sample would consist of a collection of numbers like this:

```
106 101 98 80 74 ... 107 72 100
```

Each of these IQ scores is sampled from a normal distribution with mean 100 and standard deviation 15. So if I plot a histogram of the sample, I get something like the one shown in Figure ??b. As you can see, the histogram is *roughly* the right shape, but it's a very crude approximation to the true population distribution shown in Figure ??a. When I calculate the mean of my sample, I get a number that is fairly close to the population mean 100 but not identical. In this case, it turns out that the people in my sample have a mean IQ of 98.5, and the standard deviation of their IQ scores is 15.9. These ***sample statistics*** are properties of my data set, and although they are fairly similar to the true population values, they are not the same. In general, sample statistics are the things you can calculate from your data set, and the population parameters are the things you want to learn about. Later on in this chapter I'll talk about how you can estimate population parameters using your sample statistics (Section 4.8) and how to work out how confident you are in your estimates (Section 4.9) but before we get to that there's a few more ideas in sampling theory that you need to know about.

4.6 The law of large numbers

Text by Navarro (2018)

In the previous section I showed you the results of one fictitious IQ experiment with a sample size of $N = 100$. The results were somewhat encouraging: the true population mean is 100, and the sample mean of 98.5 is a pretty reasonable approximation to it. In many scientific studies that level of precision is perfectly acceptable, but in other situations you need to be a lot more precise. If we want our sample statistics to be much closer to the population parameters, what can we do about it?

The obvious answer is to collect more data. Suppose that we ran a much larger experiment, this time measuring the IQs of 10,000 people. We can simulate the results of this experiment using R. The `rnorm()` function generates random numbers sampled from a normal distribution. For an experiment with a sample size of `n = 10000`, and a population with `mean = 100` and `sd = 15`, R produces our fake IQ data using these commands:

```
IQ <- rnorm(n = 10000, mean = 100, sd = 15) # generate IQ scores
IQ <- round(IQ) # IQs are whole numbers!
print(head(IQ))
```

```
## [1] 60 93 110 98 124 80
```

I can compute the mean IQ using the command `mean(IQ)` and the standard deviation using the command `sd(IQ)`, and I can draw a histogram using `hist()`. The histogram of this much larger sample is shown in Figure ?? c. Even a moment's inspections makes clear that the larger sample is a much better approximation to the true population distribution than the smaller one. This is reflected in the sample statistics: the mean IQ for the larger sample turns out to be 99.9, and the standard deviation is 15.1. These values are now very close to the true population.

I feel a bit silly saying this, but the thing I want you to take away from this is that large samples generally give you better information. I feel silly saying it because it's so bloody obvious that it shouldn't need to be said. In fact, it's such an obvious point that when Jacob Bernoulli – one of the founders of probability theory – formalised this idea back in 1713, he was kind of a jerk about it. Here's how he described the fact that we all share this intuition:

For even the most stupid of men, by some instinct of nature, by himself and without any instruction (which is a remarkable thing), is convinced that the more observations have been made, the less danger there is of wandering from one's goal Stigler (1986)

Okay, so the passage comes across as a bit condescending (not to mention sexist), but his main point is correct: it really does feel obvious that more data will give you better answers. The question is, why is this so? Not surprisingly, this intuition that we all share turns out to be correct, and statisticians refer to it as the **law of large numbers**. The law of large numbers is a mathematical law that applies to many different sample statistics, but the simplest way to think about it is as a law about averages. The sample mean is the most obvious example of a statistic that relies on averaging (because that's what the mean is... an average), so let's look at that. When applied to the sample mean, what the law of large numbers states is that as the sample gets larger, the sample mean tends to get closer to the true population mean. Or, to say it a little bit more precisely, as the sample size “approaches” infinity (written as $N \rightarrow \infty$) the sample mean approaches the population mean ($\bar{X} \rightarrow \mu$).⁴

I don't intend to subject you to a proof that the law of large numbers is true, but it's one of the most important tools for statistical theory. The law of large numbers is the thing we can use to justify our belief that collecting more and

⁴Technically, the law of large numbers pertains to any sample statistic that can be described as an average of independent quantities. That's certainly true for the sample mean. However, it's also possible to write many other sample statistics as averages of one form or another. The variance of a sample, for instance, can be rewritten as a kind of average and so is subject to the law of large numbers. The minimum value of a sample, however, cannot be written as an average of anything and is therefore not governed by the law of large numbers.

more data will eventually lead us to the truth. For any particular data set, the sample statistics that we calculate from it will be wrong, but the law of large numbers tells us that if we keep collecting more data those sample statistics will tend to get closer and closer to the true population parameters.

4.7 Sampling distributions and the central limit theorem

Text by Navarro (2018)

The law of large numbers is a very powerful tool, but it's not going to be good enough to answer all our questions. Among other things, all it gives us is a “long run guarantee”. In the long run, if we were somehow able to collect an infinite amount of data, then the law of large numbers guarantees that our sample statistics will be correct. But as John Maynard Keynes famously argued in economics, a long run guarantee is of little use in real life:

[The] long run is a misleading guide to current affairs. In the long run we are all dead. Economists set themselves too easy, too useless a task, if in tempestuous seasons they can only tell us, that when the storm is long past, the ocean is flat again. Keynes (1923)

As in economics, so too in psychology and statistics. It is not enough to know that we will *eventually* arrive at the right answer when calculating the sample mean. Knowing that an infinitely large data set will tell me the exact value of the population mean is cold comfort when my *actual* data set has a sample size of $N = 100$. In real life, then, we must know something about the behaviour of the sample mean when it is calculated from a more modest data set!

4.7.1 Sampling distribution of the mean

With this in mind, let's abandon the idea that our studies will have sample sizes of 10000, and consider a very modest experiment indeed. This time around we'll sample $N = 5$ people and measure their IQ scores. As before, I can simulate this experiment in R using the `rnorm()` function:

```
> IQ.1 <- round( rnorm(n=5, mean=100, sd=15 ) )
> IQ.1
[1] 90 82 94 99 110
```

The mean IQ in this sample turns out to be exactly 95. Not surprisingly, this is much less accurate than the previous experiment. Now imagine that I decided to

replicate the experiment. That is, I repeat the procedure as closely as possible: I randomly sample 5 new people and measure their IQ. Again, R allows me to simulate the results of this procedure:

```
> IQ.2 <- round( rnorm(n=5, mean=100, sd=15 ) )
> IQ.2
[1] 78 88 111 111 117
```

This time around, the mean IQ in my sample is 101. If I repeat the experiment 10 times I obtain the results shown in Table ??, and as you can see the sample mean varies from one replication to the next.

NANA	Person.1	Person.2	Person.3	Person.4	Person.5	Sample.Mean	caption
Replication 1	90	82	94	99	110	95.0	Ten replications of the IQ
Replication 2	78	88	111	111	117	101.0	Ten replications of the IQ
Replication 3	111	122	91	98	86	101.6	Ten replications of the IQ
Replication 4	98	96	119	99	107	103.8	Ten replications of the IQ
Replication 5	105	113	103	103	98	104.4	Ten replications of the IQ
Replication 6	81	89	93	85	114	92.4	Ten replications of the IQ
Replication 7	100	93	108	98	133	106.4	Ten replications of the IQ
Replication 8	107	100	105	117	85	102.8	Ten replications of the IQ
Replication 9	86	119	108	73	116	100.4	Ten replications of the IQ
Replication 10	95	126	112	120	76	105.8	Ten replications of the IQ

Now suppose that I decided to keep going in this fashion, replicating this “five IQ scores” experiment over and over again. Every time I replicate the experiment I write down the sample mean. Over time, I’d be amassing a new data set, in which every experiment generates a single data point. The first 10 observations from my data set are the sample means listed in Table ??, so my data set starts out like this:

95.0 101.0 101.6 103.8 104.4 ...

What if I continued like this for 10,000 replications, and then drew a histogram? Using the magical powers of R that’s exactly what I did, and you can see the results in Figure 4.6. As this picture illustrates, the average of 5 IQ scores is usually between 90 and 110. But more importantly, what it highlights is that if we replicate an experiment over and over again, what we end up with is a *distribution* of sample means! This distribution has a special name in statistics: it’s called the **sampling distribution of the mean**.

Sampling distributions are another important theoretical idea in statistics, and they’re crucial for understanding the behaviour of small samples. For instance, when I ran the very first “five IQ scores” experiment, the sample mean turned out to be 95. What the sampling distribution in Figure 4.6 tells us, though, is that the “five IQ scores” experiment is not very accurate. If I repeat the

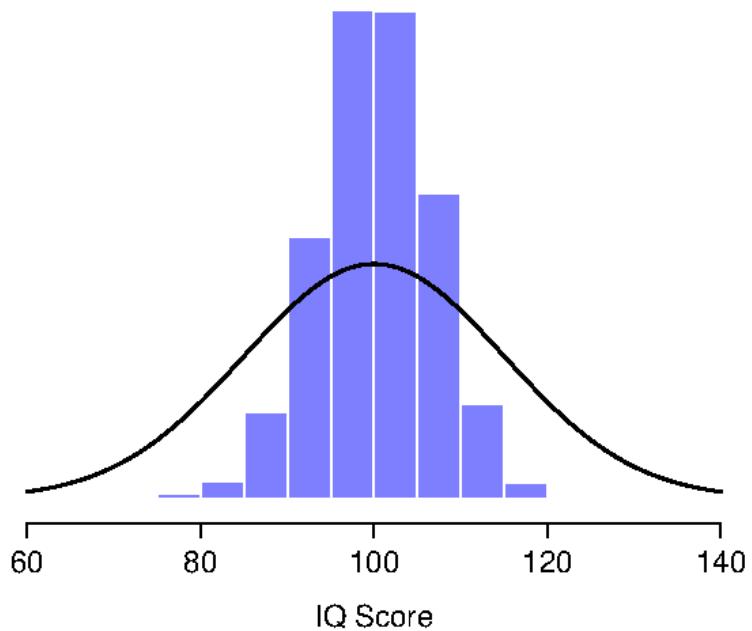


Figure 4.6: The sampling distribution of the mean for the “five IQ scores experiment”. If you sample 5 people at random and calculate their *average* IQ, you’ll almost certainly get a number between 80 and 120, even though there are quite a lot of individuals who have IQs above 120 or below 80. For comparison, the black line plots the population distribution of IQ scores.

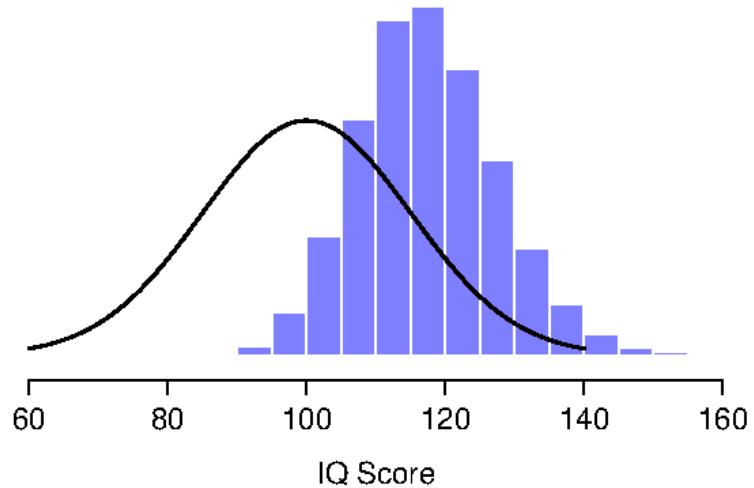


Figure 4.7: The sampling distribution of the *maximum* for the “five IQ scores experiment”. If you sample 5 people at random and select the one with the highest IQ score, you’ll probably see someone with an IQ between 100 and 140.

experiment, the sampling distribution tells me that I can expect to see a sample mean anywhere between 80 and 120.

With an explanation of sampling distributions out of the way, Dave will now explain some additional detail behind this important concept.

4.7.2 Sample size and population size

Text by David Schuster

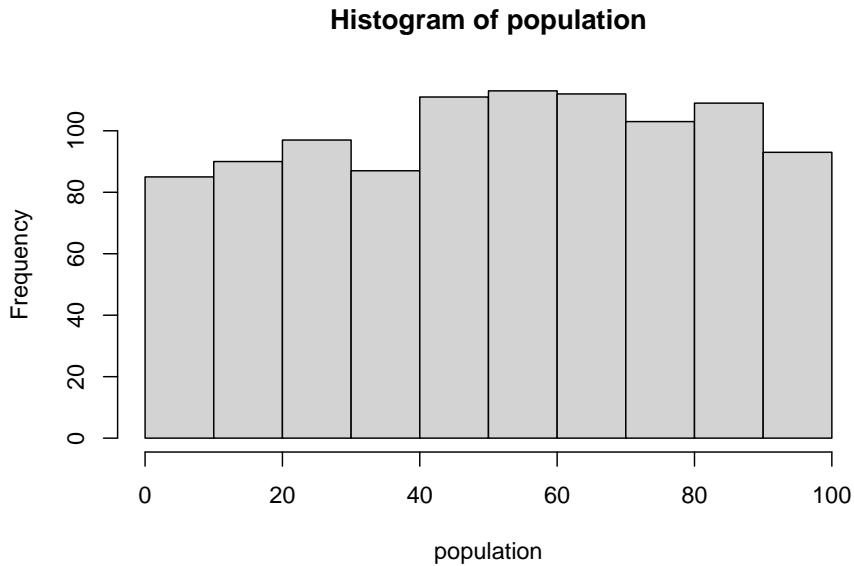
The size of a distribution is the number of units it contains. In mathematics, sample size is typically represented as n and population size is typically represented as N . In APA-style writing, however, sample size is represented as N , and n is used to represent a subsample (a part of a sample, such as the units in one condition). There does not seem to be a recommended symbol for population size in APA style. Why not? Many times, the population size is unknown.

4.7.3 Sampling error

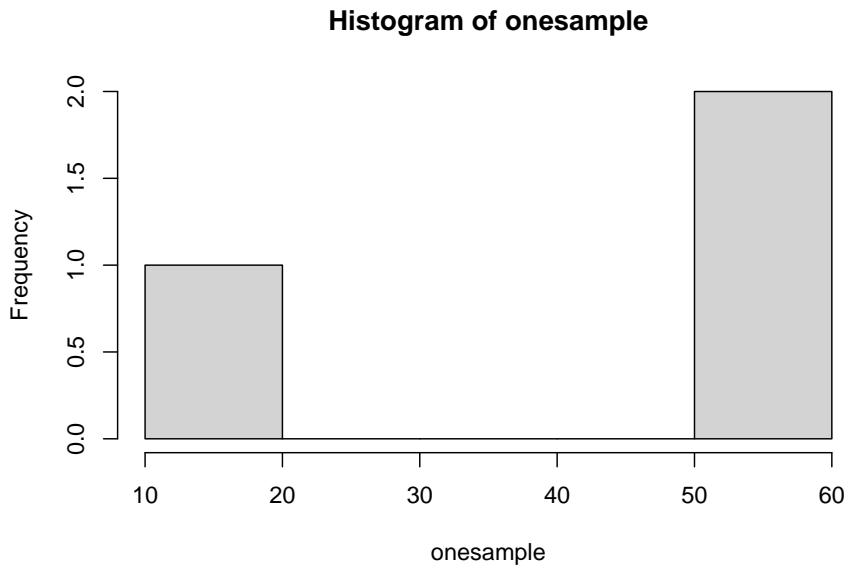
Text by David Schuster

Sampling error is the mismatch between a sample statistic and a population parameter. As we have seen, taking a single random sample does not guarantee perfect representation of a population. Let's generate a population distribution, take a single random sample, and then compare the population and sample distributions:

```
population <- runif(1000, 1, 100) # Generate a variable called 'population' with 1000 values
mean(population) # Display the mean of the population distribution
## [1] 51.67367
hist(population) # Display a histogram of the population distribution
```



```
onesample <- sample(population, size = 3, replace = TRUE) # Take a random sample of size 3 from the population
mean(onesample) # Display the mean of the sample distribution
## [1] 43.44863
hist(onesample) # Display a histogram of the sample distribution
```



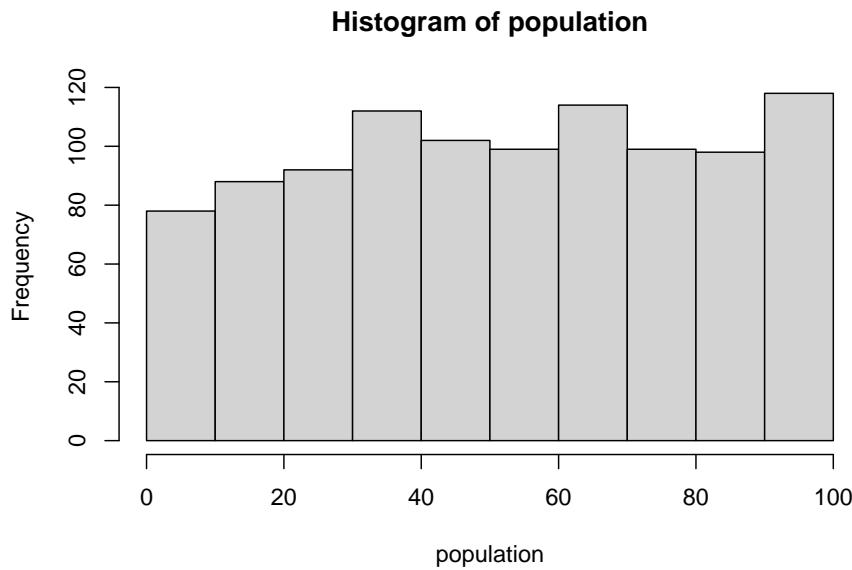
```
mean(onesample) - mean(population) # The difference between the sample mean and the population mean
## [1] -8.225038
```

This required a few R functions you may not have seen before. `runif()` randomly generates uniform distributions (all values have equal probability) of any specified size and between any specified values. `sample()` takes a random sample from a specified distribution.

What was the difference between the sample mean and the population mean? If random sampling was perfect, it would be zero. Although these values were randomly generated, I am fairly confident the difference was not zero. Next, how do the *shapes* of the two distributions compare?

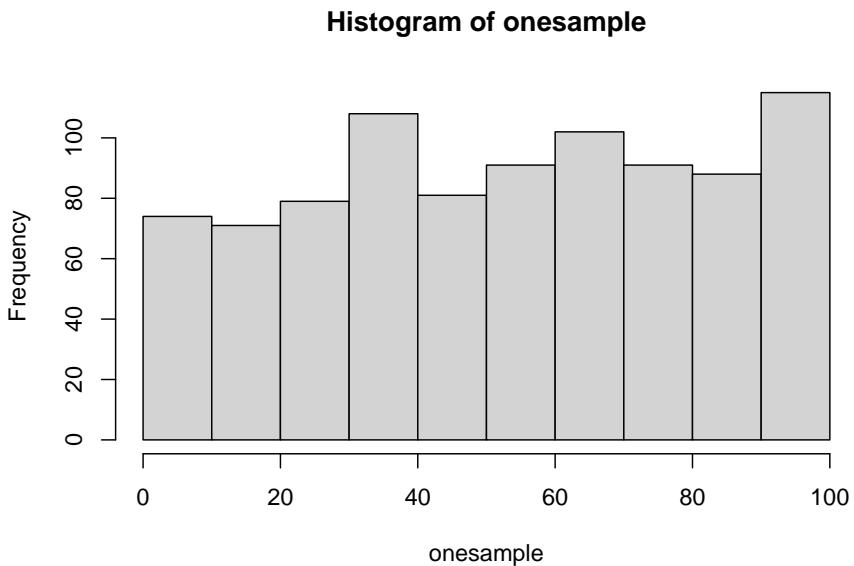
Okay, the sample is not perfect. But sampling error is not dichotomous. Sometimes we can observe larger or smaller sampling error. What causes sampling error to be larger or smaller? First, we'll change *sample size*:

```
population <- runif(1000, 1, 100) # Generate a variable called 'population' with 1000 values
mean(population) # Display the mean of the population distribution
## [1] 52.22362
hist(population) # Display a histogram of the population distribution
```



```
onesample <- sample(population, size = 900, replace = TRUE) # Take a random sample of 900 values
mean(onesample) # Display the mean of the sample distribution
```

```
## [1] 52.92439
hist(onesample) # Display a histogram of the sample distribution
```



```
mean(onesample) - mean(population) # The difference between the sample mean and the population mean
## [1] 0.7007694
```

Did increasing the sample size to 900 increase or decrease sampling error? It decreased it. That hints that using a larger sample size can give us less sampling error. What do you think would happen if we put sample size back to 3 and, instead, changed the population so that all values were between 1 and 2?

Samples have variability because populations have variability; the sample mean ultimately depends on who gets selected to be in the sample. The larger the sample, the smaller the sampling error. The greater the variability in the population, the larger the sampling error.

Next, we will expand this discussion to consider what would happen if we construct a sampling distribution. We will take one sample, then take another and another. Our units will be sample means instead of scores.

4.7.4 Another sampling distribution

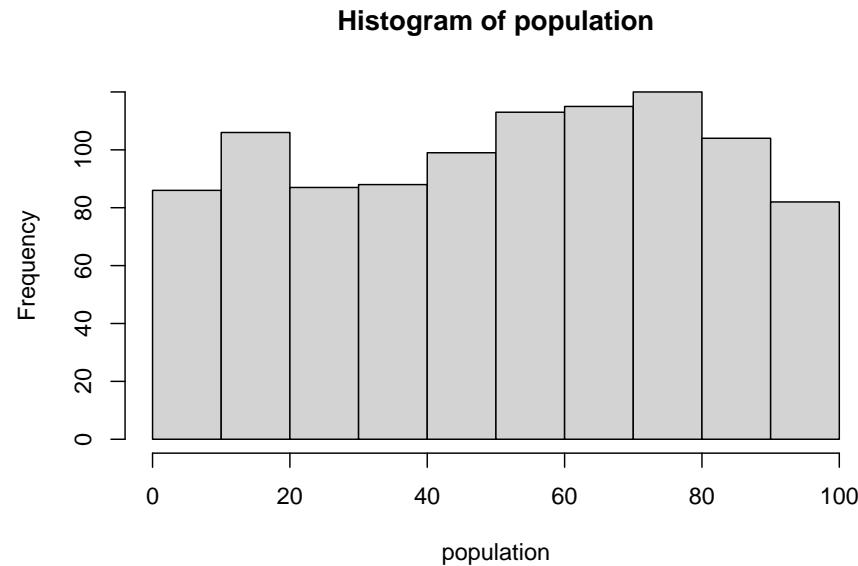
Text by David Schuster

A **sampling distribution** is a distribution of sample means.

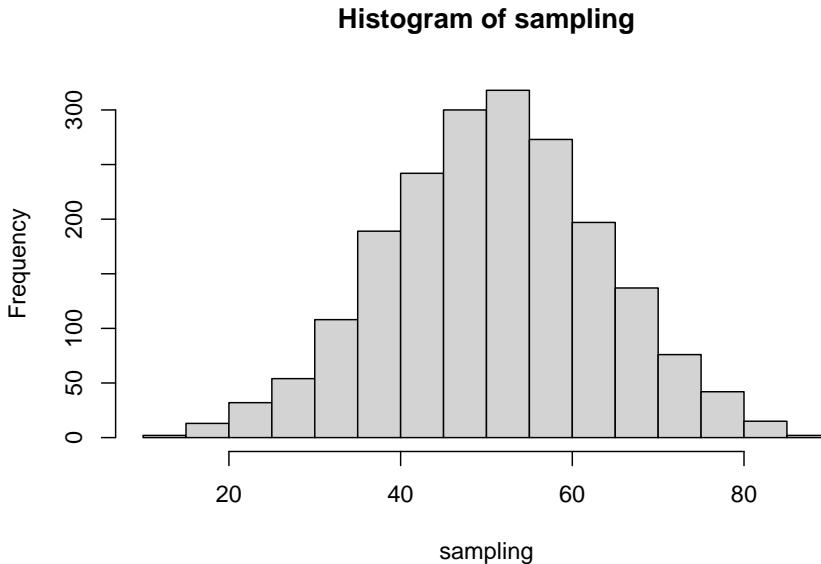
If you take repeated samples, you can plot the mean of each sample. A collection of sample means forms a sampling distribution of the mean. Sampling distributions are made of many samples.

We will modify our prior example slightly. This time, we will create a loop (technically, a for...next loop where we tell R how many times we want a command repeated) to repeat taking the sample mean. Loops are common in computer programming languages, but we don't use them very often when we do statistics in R. All you really need to know is that the loop causes the computer to repeat the commands inside the brackets.

```
population <- runif(1000, 1, 100) # Generate a variable called 'population' with 1000 values
mean(population) # Display the mean of the population distribution
## [1] 51.08084
hist(population) # Display a histogram of the population distribution
```



```
sampling = rep(NA, 2000) # create a variable with 2000 values of NA
for(i in 1:2000){
  onesample <- sample(population, size = 5, replace = TRUE) # take a random sample of 5
  sampling[i] = mean(onesample) # add the sample's mean to the sampling distribution
}
mean(sampling) # Display the mean of the SAMPLING distribution
## [1] 50.92024
hist(sampling) # Display a histogram of the SAMPLING distribution
```



```
mean(sampling) - mean(population) # The difference between the SAMPLING distribution mean and the population mean
## [1] -0.1605968
```

Okay, some *really* interesting things happened in this last example. Before we continue, make sure you understand the steps what we've taken to create a sampling distribution:

1. We started with a population distribution. The shape of the population distribution is not important (did you notice that all the populations were close to a uniform distribution? None of the populations were normally distributed). While we're at it, the population *size* is also not important. This example would have worked with a population of 50 or a population of 300,000.
2. We took a random sample from the population, with replacement. We found the mean of our random sample. We stored the mean in a variable called `sampling`.
3. We repeated Step 2 many times. Following this, we had a list of 2000 sample means stored in a variable called `sampling`. This is our sampling distribution. **Sampling distributions are made of sample means.** Put another way, the units of a sampling distribution are sample means.

What did you notice when you look at the histogram of the sampling distribution? It's normally distributed! We started with a non-normal population

and ended up with a normally distributed sampling distribution. This is one of the outcomes specified by the central limit theorem.

What did you notice about the difference between the mean of the sampling distribution and the mean of the population? It's small. It is probably the smallest value of all the examples in this section. This is another outcome specified by the central limit theorem. As we collect more and more sample means, the mean of the sampling distribution will approach the mean of the population.

It has taken us a lot of steps and several examples to get here.

4.7.5 Defining the central limit theorem

Text by David Schuster

The central limit theorem (CLT) says that sampling distributions have special properties.

The CLT says that: (1) assuming two things, (2) if you do a series of steps, then (3) you will obtain an outcome. The outcome has implications for us.

- The two **assumptions** are a random sample and a variable that is continuous.
- The **steps** are to take repeated random samples of the population and calculate the mean of each of those samples. Construct a sampling distribution from the sample means.
- The **outcome** is that the histogram of the sample means is normally distributed. We call this the sampling distribution of the mean. It will always be normally distributed under the CLT, as long as we have a sufficiently large sample size.
- This frequency distribution, like all frequency distributions, has a standard deviation called the standard error of the mean.

4.7.6 More on standard error

Text by David Schuster

Sampling distributions have a mean and standard deviation, just like any other distribution we have seen. However, the standard deviation of a sampling distribution has a special name: the standard error.

Standard error is calculated using this formula: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

In words: divide the standard deviation of the population by the square root of the sample size. Let's run the example one more time, this time calculating the standard error two ways:

```

population <- runif(1000, 1, 100) # Generate a variable called 'population' with 1000 values chosen uniformly between 1 and 100

sampling = rep(NA, 2000) # create a variable with 2000 values of NA to store our sampling distribution
for(i in 1:2000){
  onesample <- sample(population, size = 5, replace = TRUE) # take a random sample of size 5 from population
  sampling[i] = mean(onesample) # add the sample's mean to the sampling distribution
}
sampling_sd <- sqrt(var(sampling) * (length(sampling)-1)/length(sampling)) # Calculate the standard deviation of the sampling distribution
sampling_sd # Display the standard deviation of our observed (generated) sampling distribution

## [1] 12.85289

pop_sd <- sqrt(var(population) * (length(population)-1)/length(population)) # find population standard deviation
standard_error <- pop_sd / sqrt(5) # Use the central limit theorem to calculate standard error using the formula
print(standard_error) # Display the standard error

## [1] 12.8597

standard_error - sampling_sd # Display the difference between the calculated standard error and the sampling standard deviation

## [1] 0.006811136

```

They are pretty close. Why do we need the standard error formula if we could just find the standard deviation of the sampling distribution? Well, we typically do not work with the sampling distribution directly. We simply understand that it exists. Creating a sampling distribution requires the population distribution to be available to us, and this isn't usually the case when we are doing research. Further, the central limit theorem specifies taking an *unlimited* number of samples in order for the sampling distribution mean to equal the population mean. It also requires a sample size of $N \geq 30$ when the population is not normally distributed. We have violated that rule (we used a uniform population distribution and we set our sample size as low as 5), but the numbers still came out pretty close.

Notice that if we assume that the central limit theorem applies, we already know the shape, mean, and standard deviation of a sampling distribution without having to construct it. This is one key to inferential statistics, and, specifically, **parametric statistics**. Parametric statistics that are methods that are based on known (or assumed) probability distributions. The sampling distribution of the mean is one such example.

4.7.7 The sampling distribution tells us about the probability of sample means

Where this gets useful is using the sampling distribution to make statements about the probability of obtaining a single sample mean. In many research contexts, we work with a single sample distribution. We do not have access to the population distribution nor the sampling distribution. But, we can use the central limit theorem to imagine what the sampling distribution looks like (it's normal with its mean equal to the population mean and a standard error based on population standard deviation and sample size). Because the sampling distribution is made of sample means, it tells us about what we can expect if we take one single random sample from a population.

To summarize, the central limit theorem allows us to say useful things for research:

- A single random sample will have a mean that approximates the population mean. We can use samples in place of having to measure every member of the population.
- Each time we take a random sample and calculate the mean, we are most likely to get the population mean.
- Our sample means will vary. We can predict how much they vary by calculating the standard error.
- It is possible to take a random sample and calculate the mean only to get a sample mean that is far away from the population mean, but this is unlikely to happen.
- A larger sample size reduces the standard error of the mean. Larger sample sizes give us better estimates of the mean.

4.7.8 Sampling distributions exist for any sample statistic!

Text by Navarro (2018)

One thing to keep in mind when thinking about sampling distributions is that *any* sample statistic you might care to calculate has a sampling distribution. For example, suppose that each time I replicated the “five IQ scores” experiment I wrote down the largest IQ score in the experiment. This would give me a data set that started out like this:

110 117 122 119 113 ...

Doing this over and over again would give me a very different sampling distribution, namely the *sampling distribution of the maximum*. The sampling distribution of the maximum of 5 IQ scores is shown in Figure 4.7. Not surprisingly, if you pick 5 people at random and then find the person with the highest IQ score, they're going to have an above average IQ. Most of the time you'll end up with someone whose IQ is measured in the 100 to 140 range.

4.7.9 The central limit theorem

An illustration of the how sampling distribution of the mean depends on sample size. In each panel, I generated 10,000 samples of IQ data, and calculated the mean IQ observed within each of these data sets. The histograms in these plots show the distribution of these means (i.e., the sampling distribution of the mean). Each individual IQ score was drawn from a normal distribution with mean 100 and standard deviation 15, which is shown as the solid black line).

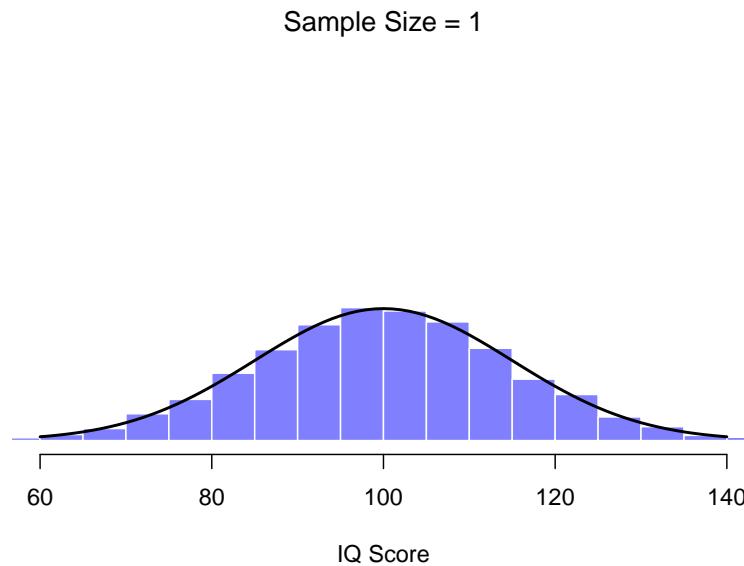


Figure 4.8: Each data set contained only a single observation, so the mean of each sample is just one person's IQ score. As a consequence, the sampling distribution of the mean is of course identical to the population distribution of IQ scores.

At this point I hope you have a pretty good sense of what sampling distributions are, and in particular what the sampling distribution of the mean is. In this section I want to talk about how the sampling distribution of the mean changes

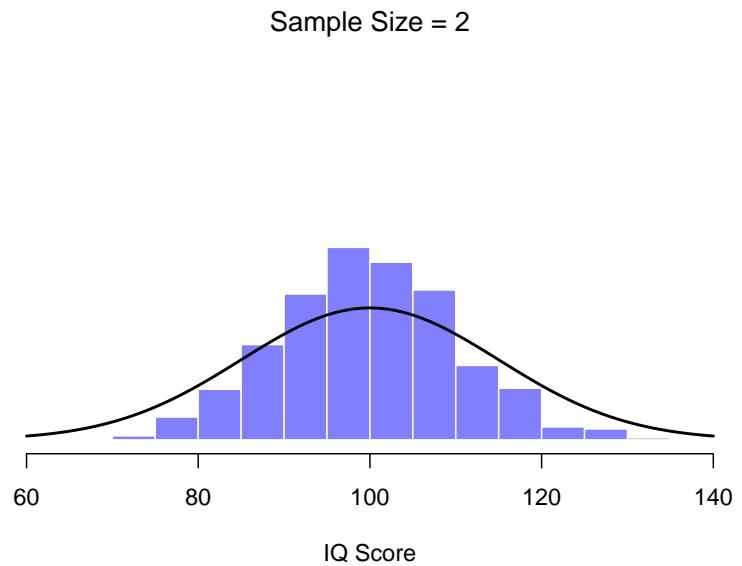


Figure 4.9: When we raise the sample size to 2, the mean of any one sample tends to be closer to the population mean than a one person's IQ score, and so the histogram (i.e., the sampling distribution) is a bit narrower than the population distribution.

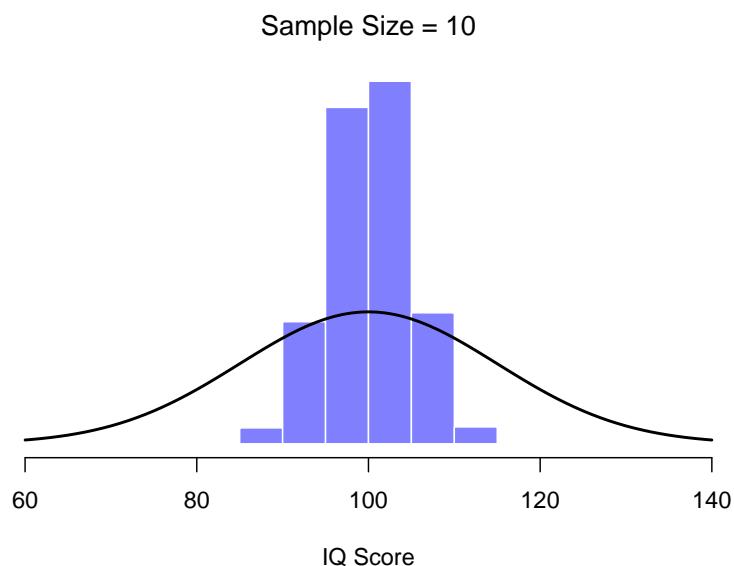


Figure 4.10: By the time we raise the sample size to 10, we can see that the distribution of sample means tend to be fairly tightly clustered around the true population mean.

as a function of sample size. Intuitively, you already know part of the answer: if you only have a few observations, the sample mean is likely to be quite inaccurate: if you replicate a small experiment and recalculate the mean you'll get a very different answer. In other words, the sampling distribution is quite wide. If you replicate a large experiment and recalculate the sample mean you'll probably get the same answer you got last time, so the sampling distribution will be very narrow. You can see this visually in Figures 4.8, 4.9 and 4.10: the bigger the sample size, the narrower the sampling distribution gets. We can quantify this effect by calculating the standard deviation of the sampling distribution, which is referred to as the ***standard error***. The standard error of a statistic is often denoted SE, and since we're usually interested in the standard error of the sample *mean*, we often use the acronym SEM. As you can see just by looking at the picture, as the sample size N increases, the SEM decreases.

Okay, so that's one part of the story. However, there's something I've been glossing over so far. All my examples up to this point have been based on the "IQ scores" experiments, and because IQ scores are roughly normally distributed, I've assumed that the population distribution is normal. What if it isn't normal? What happens to the sampling distribution of the mean? The remarkable thing is this: no matter what shape your population distribution is, as N increases the sampling distribution of the mean starts to look more like a normal distribution. To give you a sense of this, I ran some simulations using R. To do this, I started with the "ramped" distribution shown in the histogram in Figure ???. As you can see by comparing the triangular shaped histogram to the bell curve plotted by the black line, the population distribution doesn't look very much like a normal distribution at all. Next, I used R to simulate the results of a large number of experiments. In each experiment I took $N = 2$ samples from this distribution, and then calculated the sample mean. Figure ?? plots the histogram of these sample means (i.e., the sampling distribution of the mean for $N = 2$). This time, the histogram produces a \cap -shaped distribution: it's still not normal, but it's a lot closer to the black line than the population distribution in Figure ???. When I increase the sample size to $N = 4$, the sampling distribution of the mean is very close to normal (Figure ??, and by the time we reach a sample size of $N = 8$ it's almost perfectly normal. In other words, as long as your sample size isn't tiny, the sampling distribution of the mean will be approximately normal no matter what your population distribution looks like!

```
# needed for printing
width <- 6
height <- 6

# parameters of the beta
a <- 2
b <- 1

# mean and standard deviation of the beta
```

```

s <- sqrt( a*b / (a+b)^2 / (a+b+1) )
m <- a / (a+b)

# define function to draw a plot
plotOne <- function(n,N=50000) {

  # generate N random sample means of size n
  X <- matrix(rbeta(n*N,a,b),n,N)
  X <- colMeans(X)

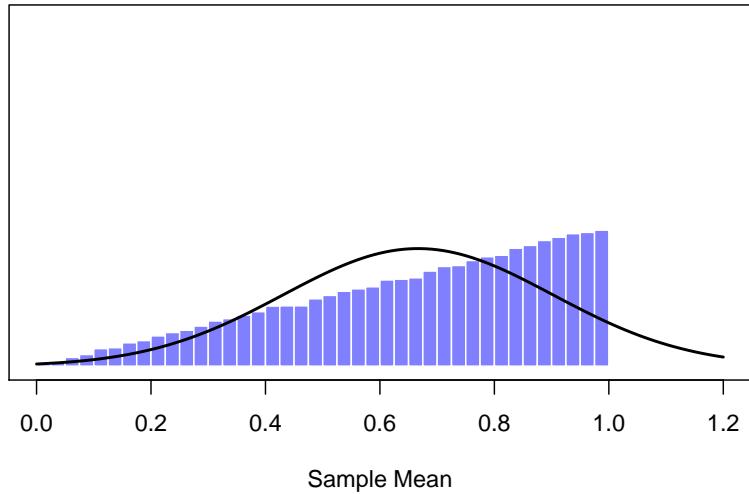
  # plot the data
  hist( X, breaks=seq(0,1,.025), border="white", freq=FALSE,
        col=ifelse(colour,emphColLight,emphGrey),
        xlab="Sample Mean", ylab="", xlim=c(0,1.2),
        main=paste("Sample Size =",n), axes=FALSE,
        font.main=1, ylim=c(0,5)
  )
  box()
  axis(1)
  #axis(2)

  # plot the theoretical distribution
  lines( x <- seq(0,1.2,.01), dnorm(x,m,s/sqrt(n)),
         lwd=2, col="black", type="l"
  )
}

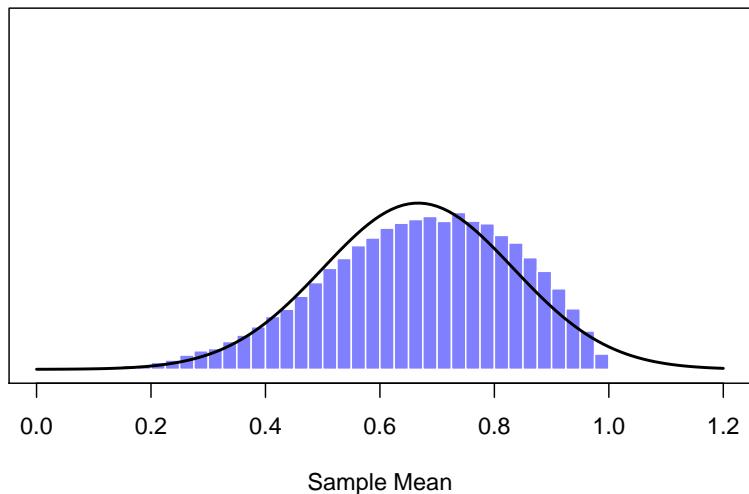
for( i in c(1,2,4,8)) {
  plotOne(i)}

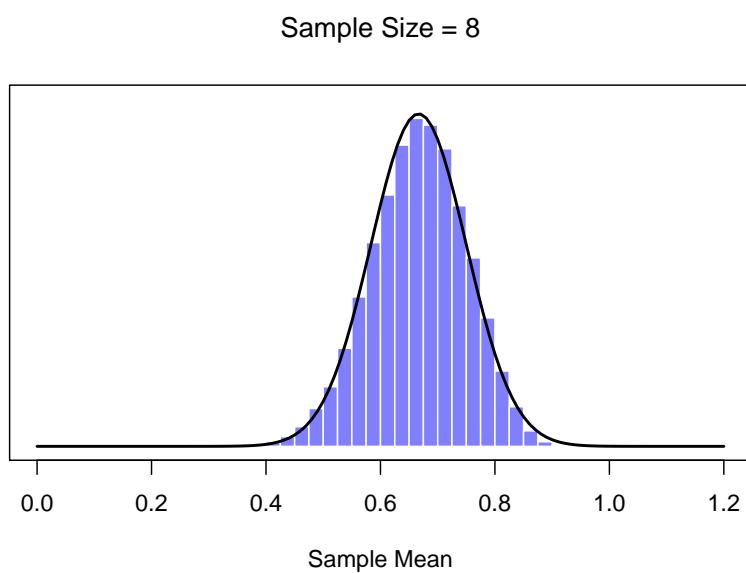
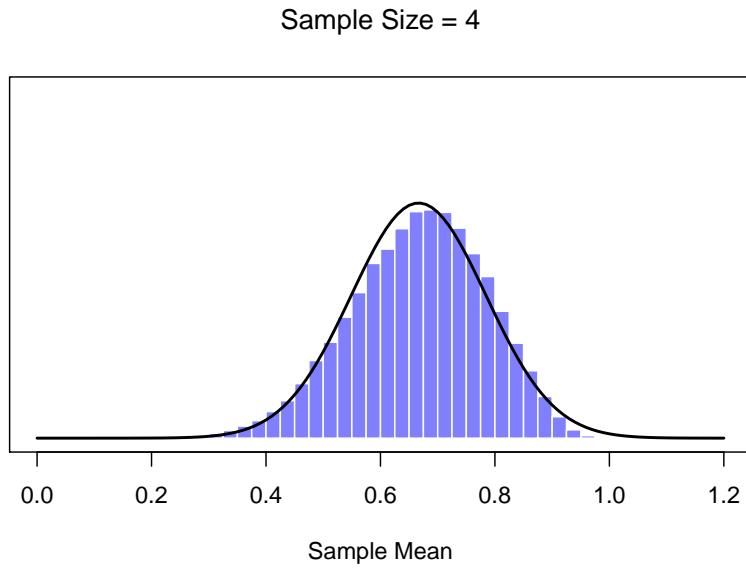
```

Sample Size = 1



Sample Size = 2





On the basis of these figures, it seems like we have evidence for all of the following claims about the sampling distribution of the mean:

- The mean of the sampling distribution is the same as the mean of the population

- The standard deviation of the sampling distribution (i.e., the standard error) gets smaller as the sample size increases
- The shape of the sampling distribution becomes normal as the sample size increases

As it happens, not only are all of these statements true, there is a very famous theorem in statistics that proves all three of them, known as the *central limit theorem*. Among other things, the central limit theorem tells us that if the population distribution has mean μ and standard deviation σ , then the sampling distribution of the mean also has mean μ , and the standard error of the mean is

$$\text{SEM} = \frac{\sigma}{\sqrt{N}}$$

Because we divide the population standard deviation σ by the square root of the sample size N , the SEM gets smaller as the sample size increases. It also tells us that the shape of the sampling distribution becomes normal.⁵

This result is useful for all sorts of things. It tells us why large experiments are more reliable than small ones, and because it gives us an explicit formula for the standard error it tells us *how much* more reliable a large experiment is. It tells us why the normal distribution is, well, *normal*. In real experiments, many of the things that we want to measure are actually averages of lots of different quantities (e.g., arguably, “general” intelligence as measured by IQ is an average of a large number of “specific” skills and abilities), and when that happens, the averaged quantity should follow a normal distribution. Because of this mathematical law, the normal distribution pops up over and over again in real data.

4.8 Estimating population parameters

Text by Navarro (2018)

In all the IQ examples in the previous sections, we actually knew the population parameters ahead of time. As every undergraduate gets taught in their very first lecture on the measurement of intelligence, IQ scores are *defined* to have mean 100 and standard deviation 15. However, this is a bit of a lie. How do we know that IQ scores have a true population mean of 100? Well, we know this because the people who designed the tests have administered them to very large samples, and have then “rigged” the scoring rules so that their sample has mean

⁵As usual, I’m being a bit sloppy here. The central limit theorem is a bit more general than this section implies. Like most introductory stats texts, I’ve discussed one situation where the central limit theorem holds: when you’re taking an average across lots of independent events drawn from the same distribution. However, the central limit theorem is much broader than this. There’s a whole class of things called “*U*-statistics” for instance, all of which satisfy the central limit theorem and therefore become normally distributed for large sample sizes. The mean is one such statistic, but it’s not the only one.

100. That's not a bad thing of course: it's an important part of designing a psychological measurement. However, it's important to keep in mind that this theoretical mean of 100 only attaches to the population that the test designers used to design the tests. Good test designers will actually go to some lengths to provide "test norms" that can apply to lots of different populations (e.g., different age groups, nationalities etc).

This is very handy, but of course almost every research project of interest involves looking at a different population of people to those used in the test norms. For instance, suppose you wanted to measure the effect of low level lead poisoning on cognitive functioning in Port Pirie, a South Australian industrial town with a lead smelter. Perhaps you decide that you want to compare IQ scores among people in Port Pirie to a comparable sample in Whyalla, a South Australian industrial town with a steel refinery.⁶ Regardless of which town you're thinking about, it doesn't make a lot of sense simply to *assume* that the true population mean IQ is 100. No-one has, to my knowledge, produced sensible norming data that can automatically be applied to South Australian industrial towns. We're going to have to *estimate* the population parameters from a sample of data. So how do we do this?

4.8.1 Estimating the population mean

Suppose we go to Port Pirie and 100 of the locals are kind enough to sit through an IQ test. The average IQ score among these people turns out to be $\bar{X} = 98.5$. So what is the true mean IQ for the entire population of Port Pirie? Obviously, we don't know the answer to that question. It could be 97.2, but it could also be 103.5. Our sampling isn't exhaustive so we cannot give a definitive answer. Nevertheless if I was forced at gunpoint to give a "best guess" I'd have to say 98.5. That's the essence of statistical estimation: giving a best guess.

In this example, estimating the unknown poulation parameter is straightforward. I calculate the sample mean, and I use that as my *estimate of the*

⁶Please note that if you were *actually* interested in this question, you would need to be a *lot* more careful than I'm being here. You *can't* just compare IQ scores in Whyalla to Port Pirie and assume that any differences are due to lead poisoning. Even if it were true that the only differences between the two towns corresponded to the different refineries (and it isn't, not by a long shot), you need to account for the fact that people already *believe* that lead pollution causes cognitive deficits: if you recall back to Chapter 1.6, this means that there are different demand effects for the Port Pirie sample than for the Whyalla sample. In other words, you might end up with an illusory group difference in your data, caused by the fact that people *think* that there is a real difference. I find it pretty implausible to think that the locals wouldn't be well aware of what you were trying to do if a bunch of researchers turned up in Port Pirie with lab coats and IQ tests, and even less plausible to think that a lot of people would be pretty resentful of you for doing it. Those people won't be as co-operative in the tests. Other people in Port Pirie might be *more* motivated to do well because they don't want their home town to look bad. The motivational effects that would apply in Whyalla are likely to be weaker, because people don't have any concept of "iron ore poisoning" in the same way that they have a concept for "lead poisoning". Psychology is *hard*.

population mean. It's pretty simple, and in the next section I'll explain the statistical justification for this intuitive answer. However, for the moment what I want to do is make sure you recognise that the sample statistic and the estimate of the population parameter are conceptually different things. A sample statistic is a description of your data, whereas the estimate is a guess about the population. With that in mind, statisticians often different notation to refer to them. For instance, if true population mean is denoted μ , then we would use $\hat{\mu}$ to refer to our estimate of the population mean. In contrast, the sample mean is denoted \bar{X} or sometimes m . However, in simple random samples, the estimate of the population mean is identical to the sample mean: if I observe a sample mean of $\bar{X} = 98.5$, then my estimate of the population mean is also $\hat{\mu} = 98.5$. To help keep the notation clear, here's a handy table:

```
knitr::kable(data.frame(stringsAsFactors=FALSE,
    Symbol = c("$\\bar{X}$", "$\\mu$", "$\\hat{\\mu}$"),
    What.is.it = c("Sample mean", "True population mean",
                  "Estimate of the population mean"),
    Do.we.know.what.it.is = c("Yes calculated from the raw data",
                               "Almost never known for sure",
                               "Yes identical to the sample mean")))

```

Symbol	What.is.it	Do.we.know.what.it.is
\$\bar{X}\$	Sample mean	Yes calculated from the raw data
μ	True population mean	Almost never known for sure
$\hat{\mu}$	Estimate of the population mean	Yes identical to the sample mean

4.8.2 Estimating the population standard deviation

So far, estimation seems pretty simple, and you might be wondering why I forced you to read through all that stuff about sampling theory. In the case of the mean, our estimate of the population parameter (i.e. $\hat{\mu}$) turned out to be identical to the corresponding sample statistic (i.e. \bar{X}). However, that's not always true. To see this, let's have a think about how to construct an *estimate of the population standard deviation*, which we'll denote $\hat{\sigma}$. What shall we use as our estimate in this case? Your first thought might be that we could do the same thing we did when estimating the mean, and just use the sample statistic as our estimate. That's almost the right thing to do, but not quite.

Here's why. Suppose I have a sample that contains a single observation. For this example, it helps to consider a sample where you have no intuitions at all about what the true population values might be, so let's use something completely fictitious. Suppose the observation in question measures the *cromulence* of my shoes. It turns out that my shoes have a cromulence of 20. So here's my sample:

This is a perfectly legitimate sample, even if it does have a sample size of $N = 1$. It has a sample mean of 20, and because every observation in this sample is equal to the sample mean (obviously!) it has a sample standard deviation of 0. As a description of the *sample* this seems quite right: the sample contains a single observation and therefore there is no variation observed within the sample. A sample standard deviation of $s = 0$ is the right answer here. But as an estimate of the *population* standard deviation, it feels completely insane, right? Admittedly, you and I don't know anything at all about what "cromulence" is, but we know something about data: the only reason that we don't see any variability in the *sample* is that the sample is too small to display any variation! So, if you have a sample size of $N = 1$, it *feels* like the right answer is just to say "no idea at all".

Notice that you *don't* have the same intuition when it comes to the sample mean and the population mean. If forced to make a best guess about the population mean, it doesn't feel completely insane to guess that the population mean is 20. Sure, you probably wouldn't feel very confident in that guess, because you have only the one observation to work with, but it's still the best guess you can make.

Let's extend this example a little. Suppose I now make a second observation. My data set now has $N = 2$ observations of the cromulence of shoes, and the complete sample now looks like this:

20, 22

This time around, our sample is *just* large enough for us to be able to observe some variability: two observations is the bare minimum number needed for any variability to be observed! For our new data set, the sample mean is $\bar{X} = 21$, and the sample standard deviation is $s = 1$. What intuitions do we have about the population? Again, as far as the population mean goes, the best guess we can possibly make is the sample mean: if forced to guess, we'd probably guess that the population mean cromulence is 21. What about the standard deviation? This is a little more complicated. The sample standard deviation is only based on two observations, and if you're at all like me you probably have the intuition that, with only two observations, we haven't given the population "enough of a chance" to reveal its true variability to us. It's not just that we suspect that the estimate is *wrong*: after all, with only two observations we expect it to be wrong to some degree. The worry is that the error is *systematic*. Specifically, we suspect that the sample standard deviation is likely to be smaller than the population standard deviation.

This intuition feels right, but it would be nice to demonstrate this somehow. There are in fact mathematical proofs that confirm this intuition, but unless you have the right mathematical background they don't help very much. Instead, what I'll do is use R to simulate the results of some experiments. With that in mind, let's return to our IQ studies. Suppose the true population mean IQ

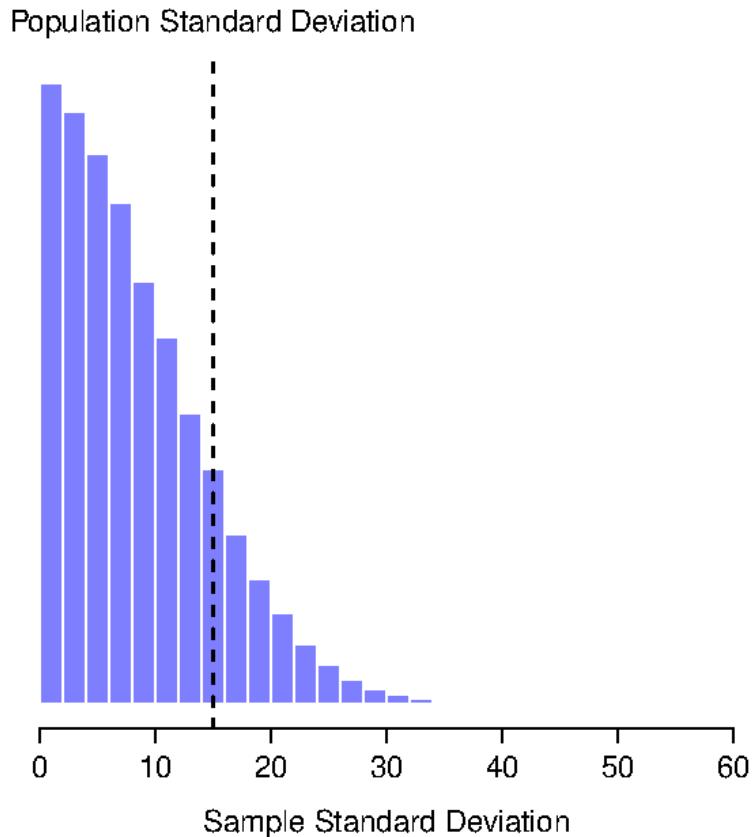


Figure 4.11: The sampling distribution of the sample standard deviation for a “two IQ scores” experiment. The true population standard deviation is 15 (dashed line), but as you can see from the histogram, the vast majority of experiments will produce a much smaller sample standard deviation than this. On average, this experiment would produce a sample standard deviation of only 8.5, well below the true value! In other words, the sample standard deviation is a *biased* estimate of the population standard deviation.

is 100 and the standard deviation is 15. I can use the `rnorm()` function to generate the results of an experiment in which I measure $N = 2$ IQ scores, and calculate the sample standard deviation. If I do this over and over again, and plot a histogram of these sample standard deviations, what I have is the *sampling distribution of the standard deviation*. I've plotted this distribution in Figure 4.11. Even though the true population standard deviation is 15, the average of the *sample* standard deviations is only 8.5. Notice that this is a very different result to what we found in Figure 4.9 when we plotted the sampling distribution of the mean. If you look at that sampling distribution, what you see is that the population mean is 100, and the average of the sample means is also 100.

Now let's extend the simulation. Instead of restricting ourselves to the situation where we have a sample size of $N = 2$, let's repeat the exercise for sample sizes from 1 to 10. If we plot the average sample mean and average sample standard deviation as a function of sample size, you get the results shown in Figure 4.12. On the left hand side (panel a), I've plotted the average sample mean and on the right hand side (panel b), I've plotted the average standard deviation. The two plots are quite different: *on average*, the average sample mean is equal to the population mean. It is an ***unbiased estimator***, which is essentially the reason why your best estimate for the population mean is the sample mean.⁷ The plot on the right is quite different: *on average*, the sample standard deviation s is *smaller* than the population standard deviation σ . It is a ***biased estimator***. In other words, if we want to make a “best guess” $\hat{\sigma}$ about the value of the population standard deviation σ , we should make sure our guess is a little bit larger than the sample standard deviation s .

The fix to this systematic bias turns out to be very simple. Here's how it works. Before tackling the standard deviation, let's look at the variance. If you recall from Section 3.4, the sample variance is defined to be the average of the squared deviations from the sample mean. That is:

$$s^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$$

The sample variance s^2 is a biased estimator of the population variance σ^2 . But as it turns out, we only need to make a tiny tweak to transform this into an unbiased estimator. All we have to do is divide by $N - 1$ rather than by N . If we do that, we obtain the following formula:

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$$

⁷I should note that I'm hiding something here. Unbiasedness is a desirable characteristic for an estimator, but there are other things that matter besides bias. However, it's beyond the scope of this book to discuss this in any detail. I just want to draw your attention to the fact that there's some hidden complexity here.

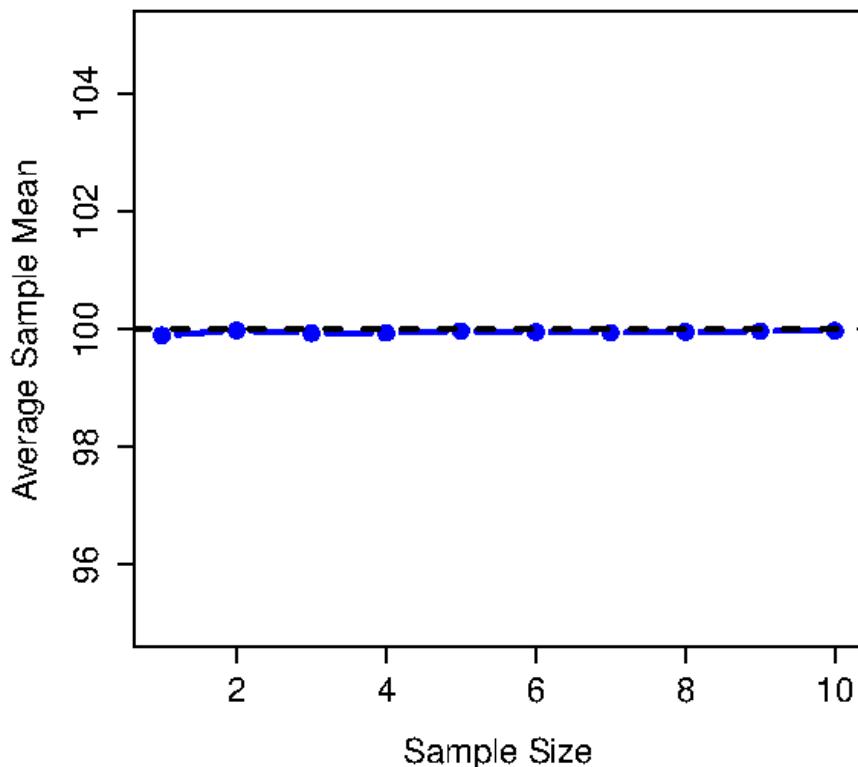


Figure 4.12: An illustration of the fact that the sample mean is an unbiased estimator of the population mean (panel a), but the sample standard deviation is a biased estimator of the population standard deviation (panel b). To generate the figure, I generated 10,000 simulated data sets with 1 observation each, 10,000 more with 2 observations, and so on up to a sample size of 10. Each data set consisted of fake IQ data: that is, the data were normally distributed with a true population mean of 100 and standard deviation 15. *On average*, the sample means turn out to be 100, regardless of sample size (panel a). However, the sample standard deviations turn out to be systematically too small (panel b), especially for small sample sizes.

This is an unbiased estimator of the population variance σ^2 . Moreover, this finally answers the question we raised in Section 3.4. Why did R give us slightly different answers when we used the `var()` function? Because the `var()` function calculates $\hat{\sigma}^2$ not s^2 , that's why. A similar story applies for the standard deviation. If we divide by $N - 1$ rather than N , our estimate of the population standard deviation becomes:

$$\hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2}$$

and when we use R's built in standard deviation function `sd()`, what it's doing is calculating $\hat{\sigma}$, not s .⁸

One final point: in practice, a lot of people tend to refer to $\hat{\sigma}$ (i.e., the formula where we divide by $N - 1$) as the *sample* standard deviation. Technically, this is incorrect: the *sample* standard deviation should be equal to s (i.e., the formula where we divide by N). These aren't the same thing, either conceptually or numerically. One is a property of the sample, the other is an estimated characteristic of the population. However, in almost every real life application, what we actually care about is the estimate of the population parameter, and so people always report $\hat{\sigma}$ rather than s . This is the right number to report, of course, it's that people tend to get a little bit imprecise about terminology when they write it up, because "sample standard deviation" is shorter than "estimated population standard deviation". It's no big deal, and in practice I do the same thing everyone else does. Nevertheless, I think it's important to keep the two *concepts* separate: it's never a good idea to confuse "known properties of your sample" with "guesses about the population from which it came". The moment you start thinking that s and $\hat{\sigma}$ are the same thing, you start doing exactly that.

To finish this section off, here's another couple of tables to help keep things clear:

```
knitr::kable(data.frame(stringsAsFactors=FALSE,
  Symbol = c("$s$",
             "$\\sigma$",
             "$\\hat{\\sigma}$",
             "$s^2$",
             "$\\sigma^2$",
             "$\\hat{\\sigma}^2$"),
  What.is.it = c("Sample standard deviation",
                "Population standard deviation",
                "Estimate of the population standard deviation",
                "Sample variance"))
```

⁸Okay, I'm hiding something else here. In a bizarre and counterintuitive twist, since $\hat{\sigma}^2$ is an unbiased estimator of σ^2 , you'd assume that taking the square root would be fine, and $\hat{\sigma}$ would be an unbiased estimator of σ . Right? Weirdly, it's not. There's actually a subtle, tiny bias in $\hat{\sigma}$. This is just bizarre: $\hat{\sigma}^2$ is an unbiased estimate of the population variance σ^2 , but when you take the square root, it turns out that $\hat{\sigma}$ is a biased estimator of the population standard deviation σ . Weird, weird, weird, right? So, why is $\hat{\sigma}$ biased? The technical answer is "because non-linear transformations (e.g., the square root) don't commute with expectation", but that just sounds like gibberish to everyone who hasn't taken a course in mathematical statistics. Fortunately, it doesn't matter for practical purposes. The bias is small, and in real life everyone uses $\hat{\sigma}$ and it works just fine. Sometimes mathematics is just annoying.

```

    "Population variance",
    "Estimate of the population variance"),
Do.we.know.what.it.is = c("Yes - calculated from the raw data",
    "Almost never known for sure",
    "Yes - but not the same as the sample standard deviation",
    "Yes - calculated from the raw data",
    "Almost never known for sure",
    "Yes - but not the same as the sample variance")
))
```

Symbol	What.is.it	Do.we.know.what.it.is
s_s	Sample standard deviation	Yes - calculated from the raw data
σ	Population standard deviation	Almost never known for sure
$\hat{\sigma}$	Estimate of the population standard deviation	Yes - but not the same as the sample standard deviation
s^2	Sample variance	Yes - calculated from the raw data
σ^2	Population variance	Almost never known for sure
$\hat{\sigma}^2$	Estimate of the population variance	Yes - but not the same as the sample variance

4.9 Estimating a confidence interval

Text by Navarro (2018)

Statistics means never having to say you're certain – Unknown origin⁹ but I've never found the original source.

Up to this point in this chapter, I've outlined the basics of sampling theory which statisticians rely on to make guesses about population parameters on the basis of a sample of data. As this discussion illustrates, one of the reasons we need all this sampling theory is that every data set leaves us with a some of uncertainty, so our estimates are never going to be perfectly accurate. The thing that has been missing from this discussion is an attempt to *quantify* the amount of uncertainty that attaches to our estimate. It's not enough to be able guess that, say, the mean IQ of undergraduate psychology students is 115 (yes, I just made that number up). We also want to be able to say something that expresses the degree of certainty that we have in our guess. For example, it would be nice to be able to say that there is a 95% chance that the true mean lies between 109 and 121. The name for this is a **confidence interval** for the mean.

Armed with an understanding of sampling distributions, constructing a confidence interval for the mean is actually pretty easy. Here's how it works. Suppose the true population mean is μ and the standard deviation is σ . I've just finished running my study that has N participants, and the mean IQ among those

⁹This quote appears on a great many t-shirts and websites, and even gets a mention in a few academic papers (e.g., \url{http://www.amstat.org/publications/jse/v10n3/friedman.html})

participants is \bar{X} . We know from our discussion of the central limit theorem (Section 4.7.9) that the sampling distribution of the mean is approximately normal. We also know from our discussion of the normal distribution Section ?? that there is a 95% chance that a normally-distributed quantity will fall within two standard deviations of the true mean. To be more precise, we can use the `qnorm()` function to compute the 2.5th and 97.5th percentiles of the normal distribution

```
qnorm( p = c(.025, .975) )
```

```
## [1] -1.959964 1.959964
```

Okay, so I lied earlier on. The more correct answer is that 95% chance that a normally-distributed quantity will fall within 1.96 standard deviations of the true mean. Next, recall that the standard deviation of the sampling distribution is referred to as the standard error, and the standard error of the mean is written as SEM. When we put all these pieces together, we learn that there is a 95% probability that the sample mean X that we have actually observed lies within 1.96 standard errors of the population mean. Mathematically, we write this as:

$$\mu - (1.96 \times \text{SEM}) \leq \bar{X} \leq \mu + (1.96 \times \text{SEM})$$

where the SEM is equal to σ/\sqrt{N} , and we can be 95% confident that this is true. However, that's not answering the question that we're actually interested in. The equation above tells us what we should expect about the sample mean, given that we know what the population parameters are. What we *want* is to have this work the other way around: we want to know what we should believe about the population parameters, given that we have observed a particular sample. However, it's not too difficult to do this. Using a little high school algebra, a sneaky way to rewrite our equation is like this:

$$\bar{X} - (1.96 \times \text{SEM}) \leq \mu \leq \bar{X} + (1.96 \times \text{SEM})$$

What this is telling us is that the range of values has a 95% probability of containing the population mean μ . We refer to this range as a **95% confidence interval**, denoted CI_{95} . In short, as long as N is sufficiently large – large enough for us to believe that the sampling distribution of the mean is normal – then we can write this as our formula for the 95% confidence interval:

$$\text{CI}_{95} = \bar{X} \pm \left(1.96 \times \frac{\sigma}{\sqrt{N}} \right)$$

Of course, there's nothing special about the number 1.96: it just happens to be the multiplier you need to use if you want a 95% confidence interval. If I'd wanted a 70% confidence interval, I could have used the `qnorm()` function to calculate the 15th and 85th quantiles:

```
qnorm( p = c(.15, .85) )
```

```
## [1] -1.036433 1.036433
```

and so the formula for CI_{70} would be the same as the formula for CI_{95} except that we'd use 1.04 as our magic number rather than 1.96.

4.9.1 A slight mistake in the formula

As usual, I lied. The formula that I've given above for the 95% confidence interval is approximately correct, but I glossed over an important detail in the discussion. Notice my formula requires you to use the standard error of the mean, SEM, which in turn requires you to use the true population standard deviation σ . Yet, in Section 4.8 I stressed the fact that we don't actually *know* the true population parameters. Because we don't know the true value of σ , we have to use an estimate of the population standard deviation $\hat{\sigma}$ instead. This is pretty straightforward to do, but this has the consequence that we need to use the quantiles of the t -distribution rather than the normal distribution to calculate our magic number; and the answer depends on the sample size. When N is very large, we get pretty much the same value using `qt()` that we would if we used `qnorm()`...

```
N <- 10000  # suppose our sample size is 10,000
qt( p = .975, df = N-1)  # calculate the 97.5th quantile of the t-dist
```

```
## [1] 1.960201
```

But when N is small, we get a much bigger number when we use the t distribution:

```
N <- 10  # suppose our sample size is 10
qt( p = .975, df = N-1)  # calculate the 97.5th quantile of the t-dist
```

```
## [1] 2.262157
```

There's nothing too mysterious about what's happening here. Bigger values mean that the confidence interval is wider, indicating that we're more uncertain about what the true value of μ actually is. When we use the t distribution instead of the normal distribution, we get bigger numbers, indicating that we have more uncertainty. And why do we have that extra uncertainty? Well, because our estimate of the population standard deviation $\hat{\sigma}$ might be wrong! If it's wrong, it implies that we're a bit less sure about what our sampling distribution of the mean actually looks like... and this uncertainty ends up getting reflected in a wider confidence interval.

4.9.2 Interpreting a confidence interval

The hardest thing about confidence intervals is understanding what they *mean*. Whenever people first encounter confidence intervals, the first instinct is almost always to say that “there is a 95% probability that the true mean lies inside the confidence interval”. It’s simple, and it seems to capture the common sense idea of what it means to say that I am “95% confident”. Unfortunately, it’s not quite right. The intuitive definition relies very heavily on your own personal *beliefs* about the value of the population mean. I say that I am 95% confident because those are my beliefs. In everyday life that’s perfectly okay, but if you remember back to Section 4.4, you’ll notice that talking about personal belief and confidence is a Bayesian idea. Personally (speaking as a Bayesian) I have no problem with the idea that the phrase “95% probability” is allowed to refer to a personal belief. However, confidence intervals are *not* Bayesian tools. Like everything else in this chapter, confidence intervals are *frequentist* tools, and if you are going to use frequentist methods then it’s not appropriate to attach a Bayesian interpretation to them. If you use frequentist methods, you must adopt frequentist interpretations!

Okay, so if that’s not the right answer, what is? Remember what we said about frequentist probability: the only way we are allowed to make “probability statements” is to talk about a sequence of events, and to count up the frequencies of different kinds of events. From that perspective, the interpretation of a 95% confidence interval must have something to do with replication. Specifically: if we replicated the experiment over and over again and computed a 95% confidence interval for each replication, then 95% of those *intervals* would contain the true mean. More generally, 95% of all confidence intervals constructed using this procedure should contain the true population mean. This idea is illustrated in Figure 4.13, which shows 50 confidence intervals constructed for a “measure 10 IQ scores” experiment (top panel) and another 50 confidence intervals for a “measure 25 IQ scores” experiment (bottom panel). A bit fortuitously, across the 100 replications that I simulated, it turned out that exactly 95 of them contained the true mean.

The critical difference here is that the Bayesian claim makes a probability statement about the population mean (i.e., it refers to our uncertainty about the population mean), which is not allowed under the frequentist interpretation of probability because you can’t “replicate” a population! In the frequentist claim, the population mean is fixed and no probabilistic claims can be made about it. Confidence intervals, however, are repeatable so we can replicate experiments. Therefore a frequentist is allowed to talk about the probability that the *confidence interval* (a random variable) contains the true mean; but is not allowed to talk about the probability that the *true population mean* (not a repeatable event) falls within the confidence interval.

I know that this seems a little pedantic, but it does matter. It matters because the difference in interpretation leads to a difference in the mathematics. There

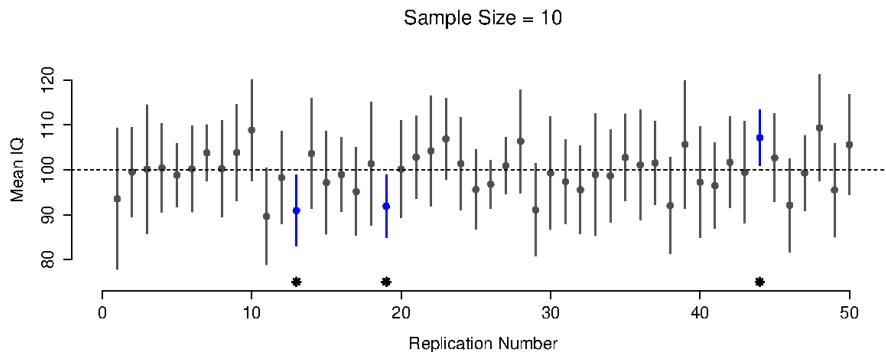


Figure 4.13: 95% confidence intervals. The top (panel a) shows 50 simulated replications of an experiment in which we measure the IQs of 10 people. The dot marks the location of the sample mean, and the line shows the 95% confidence interval. In total 47 of the 50 confidence intervals do contain the true mean (i.e., 100), but the three intervals marked with asterisks do not. The lower graph (panel b) shows a similar simulation, but this time we simulate replications of an experiment that measures the IQs of 25 people.

is a Bayesian alternative to confidence intervals, known as *credible intervals*. In most situations credible intervals are quite similar to confidence intervals, but in other cases they are drastically different. As promised, though, I'll talk more about the Bayesian perspective in Chapter ??.

4.9.3 Calculating confidence intervals in R

As far as I can tell, the core packages in R don't include a simple function for calculating confidence intervals for the mean. They *do* include a lot of complicated, extremely powerful functions that can be used to calculate confidence intervals associated with lots of different things, such as the `confint()` function that we'll use in Chapter 8. But I figure that when you're first learning statistics, it might be useful to start with something simpler. As a consequence, the `lsr` package includes a function called `ciMean()` which you can use to calculate your confidence intervals. There are two arguments that you might want to specify:¹⁰

- `x`. This should be a numeric vector containing the data.

¹⁰As of the current writing, these are the only arguments to the function. However, I am planning to add a bit more functionality to `ciMean()`. However, regardless of what those future changes might look like, the `x` and `conf` arguments will remain the same, and the commands used in this book will still work.

- `conf`. This should be a number, specifying the confidence level. By default, `conf = .95`, since 95% confidence intervals are the de facto standard in psychology.

So, for example, if I load the `afl24.Rdata` file, calculate the confidence interval associated with the mean attendance:

```
> ciMean( x = afl$attendance )
  2.5%    97.5%
31597.32 32593.12
```

Hopefully that's fairly clear.

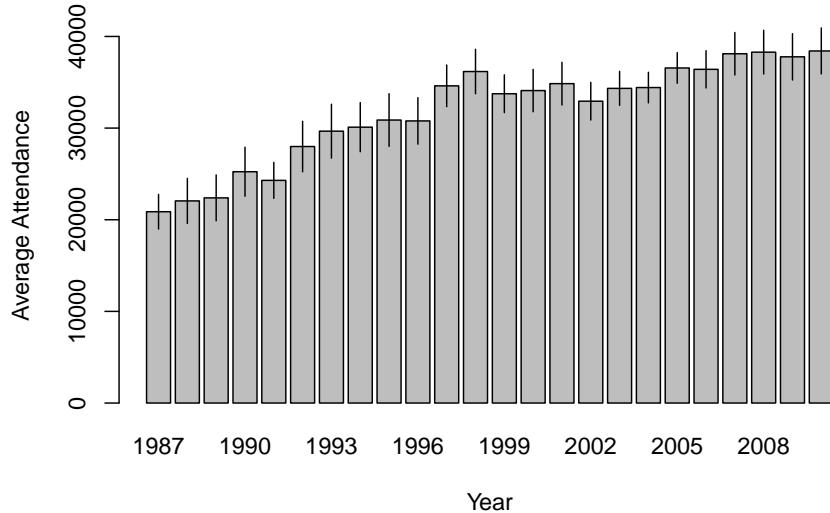
4.9.4 Plotting confidence intervals in R

There's several different ways you can draw graphs that show confidence intervals as error bars. I'll show three versions here, but this certainly doesn't exhaust the possibilities. In doing so, what I'm assuming is that you want to draw is a plot showing the means and confidence intervals for one variable, broken down by different levels of a second variable. For instance, in our `afl` data that we discussed earlier, we might be interested in plotting the average `attendance` by `year`. I'll do this using two different functions, `bargraph.CI()` and `lineplot.CI()` (both of which are in the `sciplot` package). Assuming that you've installed these packages on your system (see Section 2.17 if you've forgotten how to do this), you'll need to load them. You'll also need to load the `lsr` package, because we'll make use of the `ciMean()` function to actually calculate the confidence intervals

```
load( file.path(projecthome, "data/afl24.Rdata" ) ) # contains the "afl" data frame
library( sciplot )      # bargraph.CI() and lineplot.CI() functions
library( lsr )          # ciMean() function
```

Here's how to plot the means and confidence intervals drawn using `bargraph.CI()`.

```
bargraph.CI( x.factor = year,           # grouping variable
             response = attendance,     # outcome variable
             data = afl,               # data frame with the variables
             ci.fun= ciMean,           # name of the function to calculate CIs
             xlab = "Year",            # x-axis label
             ylab = "Average Attendance" # y-axis label
           )
```



We can use the same arguments when calling the `lineplot.CI()` function:

```
lineplot.CI( x.factor = year,           # grouping variable
             response = attendance,    # outcome variable
             data = afl,               # data frame with the variables
             ci.fun= ciMean,          # name of the function to calculate CIs
             xlab = "Year",            # x-axis label
             ylab = "Average Attendance" # y-axis label
)
```

4.10 Summary

Text by Navarro (2018)

In this chapter I've covered two main topics. The first half of the chapter talks about sampling theory, and the second half talks about how we can use sampling theory to construct estimates of the population parameters. The section breakdown looks like this:

- Basic ideas about samples, sampling and populations (Section 4.5)
- Statistical theory of sampling: the law of large numbers (Section 4.6), sampling distributions and the central limit theorem (Section 4.7).
- Estimating means and standard deviations (Section 4.8)
- Estimating a confidence interval (Section 4.9)

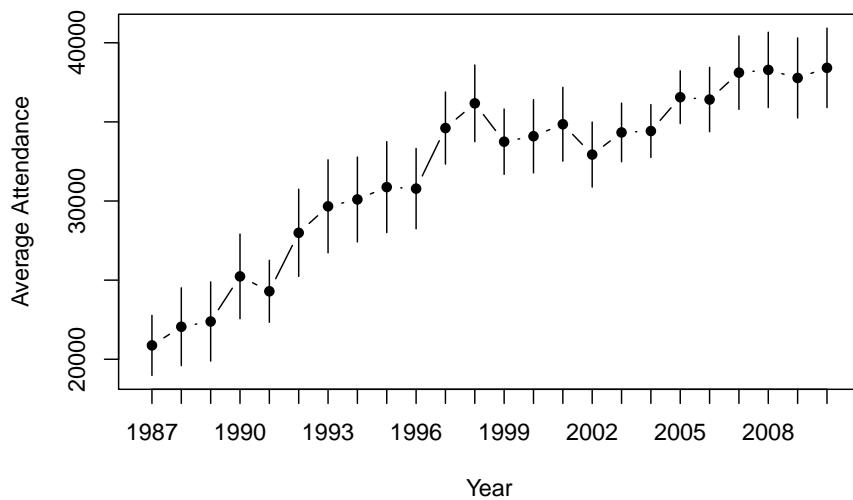


Figure 4.14: Means and 95% confidence intervals for AFL attendance, plotted separately for each year from 1987 to 2010. This graph was drawn using the `lineplot.CI()` function.

As always, there's a lot of topics related to sampling and estimation that aren't covered in this chapter, but for an introductory psychology class this is fairly comprehensive I think. For most applied researchers you won't need much more theory than this. One big question that I haven't touched on in this chapter is what you do when you don't have a simple random sample. There is a lot of statistical theory you can draw on to handle this situation, but it's well beyond the scope of this book.

Chapter 5

Hypothesis testing

Text by Navarro (2018)

5.1 Videos

- Video: Statistical power and errors
- Video: The one-sample Z-test

5.2 Introduction

The process of induction is the process of assuming the simplest law that can be made to harmonize with our experience. This process, however, has no logical foundation but only a psychological one. It is clear that there are no grounds for believing that the simplest course of events will really happen. It is an hypothesis that the sun will rise tomorrow: and this means that we do not know whether it will rise.

– Ludwig Wittgenstein¹

In the last chapter, I discussed the ideas behind estimation, which is one of the two “big ideas” in inferential statistics. It’s now time to turn out attention to the other big idea, which is *hypothesis testing*. In its most abstract form, hypothesis testing really a very simple idea: the researcher has some theory about the world, and wants to determine whether or not the data actually support that theory. However, the details are messy, and most people find the theory of

¹The quote comes from Wittgenstein’s (1922) text, *Tractatus Logico-Philosophicus*.

hypothesis testing to be the most frustrating part of statistics. The structure of the chapter is as follows. Firstly, I'll describe how hypothesis testing works, in a fair amount of detail, using a simple running example to show you how a hypothesis test is "built". I'll try to avoid being too dogmatic while doing so, and focus instead on the underlying logic of the testing procedure.² Afterwards, I'll spend a bit of time talking about the various dogmas, rules and heresies that surround the theory of hypothesis testing.

5.3 A menagerie of hypotheses

Eventually we all succumb. For me, that day will arrive once I'm finally promoted to full professor. Safely ensconced in my ivory tower, happily protected by tenure, I will finally be able to take leave of my senses (so to speak), and indulge in that most thoroughly unproductive line of psychological research: the search for extrasensory perception (ESP).³

Let's suppose that this glorious day has come. My first study is a simple one, in which I seek to test whether clairvoyance exists. Each participant sits down at a table, and is shown a card by an experimenter. The card is black on one side and white on the other. The experimenter takes the card away, and places it on a table in an adjacent room. The card is placed black side up or white side up completely at random, with the randomisation occurring only after the experimenter has left the room with the participant. A second experimenter comes in and asks the participant which side of the card is now facing upwards. It's purely a one-shot experiment. Each person sees only one card, and gives only one answer; and at no stage is the participant actually in contact with someone who knows the right answer. My data set, therefore, is very simple. I have asked the question of N people, and some number X of these people have given the correct response. To make things concrete, let's suppose that I have tested $N = 100$ people, and $X = 62$ of these got the answer right... a surprisingly large number, sure, but is it large enough for me to feel safe in claiming I've found evidence for ESP? This is the situation where hypothesis testing comes in useful. However, before we talk about how to *test* hypotheses, we need to be clear about what we mean by hypotheses.

²A technical note. The description below differs subtly from the standard description given in a lot of introductory texts. The orthodox theory of null hypothesis testing emerged from the work of Sir Ronald Fisher and Jerzy Neyman in the early 20th century; but Fisher and Neyman actually had very different views about how it should work. The standard treatment of hypothesis testing that most texts use is a hybrid of the two approaches. The treatment here is a little more Neyman-style than the orthodox view, especially as regards the meaning of the p value.

³My apologies to anyone who actually believes in this stuff, but on my reading of the literature on ESP, it's just not reasonable to think this is real. To be fair, though, some of the studies are rigorously designed; so it's actually an interesting area for thinking about psychological research design. And of course it's a free country, so you can spend your own time and effort proving me wrong if you like, but I wouldn't think that's a terribly practical use of your intellect.

5.3.1 Research hypotheses versus statistical hypotheses

The first distinction that you need to keep clear in your mind is between research hypotheses and statistical hypotheses. In my ESP study, my overall scientific goal is to demonstrate that clairvoyance exists. In this situation, I have a clear research goal: I am hoping to discover evidence for ESP. In other situations I might actually be a lot more neutral than that, so I might say that my research goal is to determine whether or not clairvoyance exists. Regardless of how I want to portray myself, the basic point that I'm trying to convey here is that a research hypothesis involves making a substantive, testable scientific claim... if you are a psychologist, then your research hypotheses are fundamentally *about* psychological constructs. Any of the following would count as **research hypotheses**:

- *Listening to music reduces your ability to pay attention to other things.* This is a claim about the causal relationship between two psychologically meaningful concepts (listening to music and paying attention to things), so it's a perfectly reasonable research hypothesis.
- *Intelligence is related to personality.* Like the last one, this is a relational claim about two psychological constructs (intelligence and personality), but the claim is weaker: it is not causal.
- *Intelligence is* speed of information processing. This hypothesis has a quite different character: it's not actually a relational claim at all. It's an ontological claim about the fundamental character of intelligence (and I'm pretty sure it's wrong). It's worth expanding on this one actually: It's usually easier to think about how to construct experiments to test research hypotheses of the form “does X affect Y?” than it is to address claims like “what is X?” And in practice, what usually happens is that you find ways of testing relational claims that follow from your ontological ones. For instance, if I believe that intelligence is* speed of information processing in the brain, my experiments will often involve looking for relationships between measures of intelligence and measures of speed.* As a consequence, most everyday research questions do tend to be relational in nature, but they're almost always motivated by deeper ontological questions about the state of nature.

Notice that in practice, my research hypotheses could overlap a lot. My ultimate goal in the ESP experiment might be to test an ontological claim like “ESP exists”, but I might operationally restrict myself to a narrower hypothesis like “Some people can ‘see’ objects in a clairvoyant fashion”. That said, there are some things that really don’t count as proper research hypotheses in any meaningful sense:

- *Love is a battlefield.* This is too vague to be testable. While it’s okay for a research hypothesis to have a degree of vagueness to it, it has to be possible

to operationalise your constructs. Maybe I'm just not creative enough to see it, but I can't see how this can be converted into any concrete research design. If that's true, then this isn't a scientific research hypothesis, it's a pop song. That doesn't mean it's not interesting – a lot of deep questions that humans have fall into this category. Maybe one day science will be able to construct testable theories of love, or to test to see if God exists, and so on; but right now we can't, and I wouldn't bet on ever seeing a satisfying scientific approach to either.

- *The first rule of tautology club is the first rule of tautology club.* This is not a substantive claim of any kind. It's true by definition. No conceivable state of nature could possibly be inconsistent with this claim. As such, we say that this is an unfalsifiable hypothesis, and as such it is outside the domain of science. Whatever else you do in science, your claims must have the possibility of being wrong.
- *More people in my experiment will say “yes” than “no”.* This one fails as a research hypothesis because it's a claim about the data set, not about the psychology (unless of course your actual research question is whether people have some kind of “yes” bias!). As we'll see shortly, this hypothesis is starting to sound more like a statistical hypothesis than a research hypothesis.

As you can see, research hypotheses can be somewhat messy at times; and ultimately they are *scientific* claims. **Statistical hypotheses** are neither of these two things. Statistical hypotheses must be mathematically precise, and they must correspond to specific claims about the characteristics of the data generating mechanism (i.e., the “population”). Even so, the intent is that statistical hypotheses bear a clear relationship to the substantive research hypotheses that you care about! For instance, in my ESP study my research hypothesis is that some people are able to see through walls or whatever. What I want to do is to “map” this onto a statement about how the data were generated. So let's think about what that statement would be. The quantity that I'm interested in within the experiment is $P(\text{"correct"})$, the true-but-unknown probability with which the participants in my experiment answer the question correctly. Let's use the Greek letter θ (theta) to refer to this probability. Here are four different statistical hypotheses:

- If ESP doesn't exist and if my experiment is well designed, then my participants are just guessing. So I should expect them to get it right half of the time and so my statistical hypothesis is that the true probability of choosing correctly is $\theta = 0.5$.
- Alternatively, suppose ESP does exist and participants can see the card. If that's true, people will perform better than chance. The statistical hypothesis would be that $\theta > 0.5$.
- A third possibility is that ESP does exist, but the colours are all reversed and people don't realise it (okay, that's wacky, but you never know...). If

that's how it works then you'd expect people's performance to be *below* chance. This would correspond to a statistical hypothesis that $\theta < 0.5$.

- Finally, suppose ESP exists, but I have no idea whether people are seeing the right colour or the wrong one. In that case, the only claim I could make about the data would be that the probability of making the correct answer is *not* equal to 50. This corresponds to the statistical hypothesis that $\theta \neq 0.5$.

All of these are legitimate examples of a statistical hypothesis because they are statements about a population parameter and are meaningfully related to my experiment.

What this discussion makes clear, I hope, is that when attempting to construct a statistical hypothesis test the researcher actually has two quite distinct hypotheses to consider. First, he or she has a research hypothesis (a claim about psychology), and this corresponds to a statistical hypothesis (a claim about the data generating population). In my ESP example, these might be

Dan.s.research.hypothesis	Dan.s.statistical.hypothesis
ESP.exists	$\$\\theta \\neq 0.5\$$

And the key thing to recognise is this: *a statistical hypothesis test is a test of the statistical hypothesis, not the research hypothesis*. If your study is badly designed, then the link between your research hypothesis and your statistical hypothesis is broken. To give a silly example, suppose that my ESP study was conducted in a situation where the participant can actually see the card reflected in a window; if that happens, I would be able to find very strong evidence that $\theta \neq 0.5$, but this would tell us nothing about whether "ESP exists".

5.3.2 Null hypotheses and alternative hypotheses

So far, so good. I have a research hypothesis that corresponds to what I want to believe about the world, and I can map it onto a statistical hypothesis that corresponds to what I want to believe about how the data were generated. It's at this point that things get somewhat counterintuitive for a lot of people. Because what I'm about to do is invent a new statistical hypothesis (the "null" hypothesis, H_0) that corresponds to the exact opposite of what I want to believe, and then focus exclusively on that, almost to the neglect of the thing I'm actually interested in (which is now called the "alternative" hypothesis, H_1). In our ESP example, the null hypothesis is that $\theta = 0.5$, since that's what we'd expect if ESP *didn't* exist. My hope, of course, is that ESP is totally real, and so the *alternative* to this null hypothesis is $\theta \neq 0.5$. In essence, what we're doing here is dividing up the possible values of θ into two groups: those values that I really hope aren't true (the null), and those values that I'd be happy with if they turn out to be right (the alternative). Having done so, the important thing to recognise is that the goal of a hypothesis test is *not* to show that the alternative

hypothesis is (probably) true; the goal is to show that the null hypothesis is (probably) false. Most people find this pretty weird.

The best way to think about it, in my experience, is to imagine that a hypothesis test is a criminal trial⁴... *the trial of the null hypothesis*. The null hypothesis is the defendant, the researcher is the prosecutor, and the statistical test itself is the judge. Just like a criminal trial, there is a presumption of innocence: the null hypothesis is *deemed* to be true unless you, the researcher, can prove beyond a reasonable doubt that it is false. You are free to design your experiment however you like (within reason, obviously!), and your goal when doing so is to maximise the chance that the data will yield a conviction... for the crime of being false. The catch is that the statistical test sets the rules of the trial, and those rules are designed to protect the null hypothesis – specifically to ensure that if the null hypothesis is actually true, the chances of a false conviction are guaranteed to be low. This is pretty important: after all, the null hypothesis doesn't get a lawyer. And given that the researcher is trying desperately to prove it to be false, *someone* has to protect it.

5.4 Two types of errors

Before going into details about how a statistical test is constructed, it's useful to understand the philosophy behind it. I hinted at it when pointing out the similarity between a null hypothesis test and a criminal trial, but I should now be explicit. Ideally, we would like to construct our test so that we never make any errors. Unfortunately, since the world is messy, this is never possible. Sometimes you're just really unlucky: for instance, suppose you flip a coin 10 times in a row and it comes up heads all 10 times. That feels like very strong evidence that the coin is biased (and it is!), but of course there's a 1 in 1024 chance that this would happen even if the coin was totally fair. In other words, in real life we *always* have to accept that there's a chance that we did the wrong thing. As a consequence, the goal behind statistical hypothesis testing is not to *eliminate* errors, but to *minimise* them.

At this point, we need to be a bit more precise about what we mean by “errors”. Firstly, let's state the obvious: it is either the case that the null hypothesis is true, or it is false; and our test will either reject the null hypothesis or retain it.⁵ So, as the table below illustrates, after we run the test and make our choice, one of four things might have happened:

⁴This analogy only works if you're from an adversarial legal system like UK/US/Australia. As I understand these things, the French inquisitorial system is quite different.

⁵An aside regarding the language you use to talk about hypothesis testing. Firstly, one thing you really want to avoid is the word “prove”: a statistical test really doesn't *prove* that a hypothesis is true or false. Proof implies certainty, and as the saying goes, statistics means never having to say you're certain. On that point almost everyone would agree. However, beyond that there's a fair amount of confusion. Some people argue that you're only allowed to make statements like “rejected the null”, “failed to reject the null”, or possibly “retained the null”. According to this line of thinking, you can't say things like “accept the alternative”

	retain H_0	reject H_0
H_0 is true	correct decision	error (type I)
H_0 is false	error (type II)	correct decision

As a consequence there are actually *two* different types of error here. If we reject a null hypothesis that is actually true, then we have made a ***type I error***. On the other hand, if we retain the null hypothesis when it is in fact false, then we have made a ***type II error***.

Remember how I said that statistical testing was kind of like a criminal trial? Well, I meant it. A criminal trial requires that you establish “beyond a reasonable doubt” that the defendant did it. All of the evidentiary rules are (in theory, at least) designed to ensure that there’s (almost) no chance of wrongfully convicting an innocent defendant. The trial is designed to protect the rights of a defendant: as the English jurist William Blackstone famously said, it is “better that ten guilty persons escape than that one innocent suffer.” In other words, a criminal trial doesn’t treat the two types of error in the same way~... punishing the innocent is deemed to be much worse than letting the guilty go free. A statistical test is pretty much the same: the single most important design principle of the test is to *control* the probability of a type I error, to keep it below some fixed probability. This probability, which is denoted α , is called the ***significance level*** of the test (or sometimes, the *size* of the test). And I’ll say it again, because it is so central to the whole set-up~... a hypothesis test is said to have significance level α if the type I error rate is no larger than α .

So, what about the type II error rate? Well, we’d also like to keep those under control too, and we denote this probability by β . However, it’s much more common to refer to the ***power*** of the test, which is the probability with which we reject a null hypothesis when it really is false, which is $1 - \beta$. To help keep this straight, here’s the same table again, but with the relevant numbers added:

	retain H_0	reject H_0
H_0 is true	$1 - \alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1 - \beta$ (power of the test)

A “powerful” hypothesis test is one that has a small value of β , while still keeping α fixed at some (small) desired level. By convention, scientists make use of three different α levels: .05, .01 and .001. Notice the asymmetry here~... the tests are designed to *ensure* that the α level is kept small, but there’s no corresponding guarantee regarding β . We’d certainly *like* the type II error rate to be small, and we try to design tests that keep it small, but this is very much secondary to the

or “accept the null”. Personally I think this is too strong: in my opinion, this conflates null hypothesis testing with Karl Popper’s falsificationist view of the scientific process. While there are similarities between falsificationism and null hypothesis testing, they aren’t equivalent. However, while I personally think it’s fine to talk about accepting a hypothesis (on the proviso that “acceptance” doesn’t actually mean that it’s necessarily true, especially in the case of the null hypothesis), many people will disagree. And more to the point, you should be aware that this particular weirdness exists, so that you’re not caught unawares by it when writing up your own results.

overwhelming need to control the type I error rate. As Blackstone might have said if he were a statistician, it is “better to retain 10 false null hypotheses than to reject a single true one”. To be honest, I don’t know that I agree with this philosophy – there are situations where I think it makes sense, and situations where I think it doesn’t – but that’s neither here nor there. It’s how the tests are built.

5.5 Test statistics and sampling distributions

At this point we need to start talking specifics about how a hypothesis test is constructed. To that end, let’s return to the ESP example. Let’s ignore the actual data that we obtained, for the moment, and think about the structure of the experiment. Regardless of what the actual numbers are, the *form* of the data is that X out of N people correctly identified the colour of the hidden card. Moreover, let’s suppose for the moment that the null hypothesis really is true: ESP doesn’t exist, and the true probability that anyone picks the correct colour is exactly $\theta = 0.5$. What would we *expect* the data to look like? Well, obviously, we’d expect the proportion of people who make the correct response to be pretty close to 50%. Or, to phrase this in more mathematical terms, we’d say that X/N is approximately 0.5. Of course, we wouldn’t expect this fraction to be *exactly* 0.5: if, for example we tested $N = 100$ people, and $X = 53$ of them got the question right, we’d probably be forced to concede that the data are quite consistent with the null hypothesis. On the other hand, if $X = 99$ of our participants got the question right, then we’d feel pretty confident that the null hypothesis is wrong. Similarly, if only $X = 3$ people got the answer right, we’d be similarly confident that the null was wrong. Let’s be a little more technical about this: we have a quantity X that we can calculate by looking at our data; after looking at the value of X , we make a decision about whether to believe that the null hypothesis is correct, or to reject the null hypothesis in favour of the alternative. The name for this thing that we calculate to guide our choices is a *test statistic*.

Having chosen a test statistic, the next step is to state precisely which values of the test statistic would cause us to reject the null hypothesis, and which values would cause us to keep it. In order to do so, we need to determine what the *sampling distribution of the test statistic* would be if the null hypothesis were actually true (we talked about sampling distributions earlier in Section 4.7.1). Why do we need this? Because this distribution tells us exactly what values of X our null hypothesis would lead us to expect. And therefore, we can use this distribution as a tool for assessing how closely the null hypothesis agrees with our data.

How do we actually determine the sampling distribution of the test statistic? For a lot of hypothesis tests this step is actually quite complicated, and later on in the book you’ll see me being slightly evasive about it for some of the tests

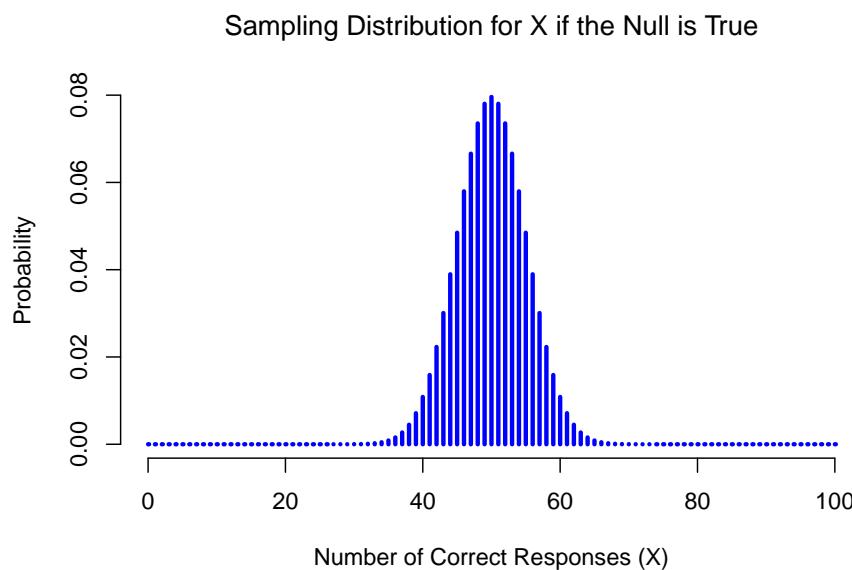


Figure 5.1: The sampling distribution for our test statistic X when the null hypothesis is true. For our ESP scenario, this is a binomial distribution. Not surprisingly, since the null hypothesis says that the probability of a correct response is $\theta = .5$, the sampling distribution says that the most likely value is 50 (out of 100) correct responses. Most of the probability mass lies between 40 and 60.

(some of them I don't even understand myself). However, sometimes it's very easy. And, fortunately for us, our ESP example provides us with one of the easiest cases. Our population parameter θ is just the overall probability that people respond correctly when asked the question, and our test statistic X is the *count* of the number of people who did so, out of a sample size of N . There is a distribution that describes exactly that, called the binomial distribution. So, to use the notation and terminology of the binomial distribution, we would say that the null hypothesis predicts that X is binomially distributed, which is written

$$X \sim \text{Binomial}(\theta, N)$$

Since the null hypothesis states that $\theta = 0.5$ and our experiment has $N = 100$ people, we have the sampling distribution we need. This sampling distribution is plotted in Figure 5.1. No surprises really: the null hypothesis says that $X = 50$ is the most likely outcome, and it says that we're almost certain to see somewhere between 40 and 60 correct responses.

5.6 Making decisions

Okay, we're very close to being finished. We've constructed a test statistic (X), and we chose this test statistic in such a way that we're pretty confident that if X is close to $N/2$ then we should retain the null, and if not we should reject it. The question that remains is this: exactly which values of the test statistic should we associate with the null hypothesis, and which exactly values go with the alternative hypothesis? In my ESP study, for example, I've observed a value of $X = 62$. What decision should I make? Should I choose to believe the null hypothesis, or the alternative hypothesis?

5.6.1 Critical regions and critical values

To answer this question, we need to introduce the concept of a *critical region* for the test statistic X . The critical region of the test corresponds to those values of X that would lead us to reject null hypothesis (which is why the critical region is also sometimes called the rejection region). How do we find this critical region? Well, let's consider what we know:

- X should be very big or very small in order to reject the null hypothesis.
- If the null hypothesis is true, the sampling distribution of X is $\text{Binomial}(0.5, N)$.
- If $\alpha = .05$, the critical region must cover 5% of this sampling distribution.

It's important to make sure you understand this last point: the critical region corresponds to those values of X for which we would reject the null hypothesis,

and the sampling distribution in question describes the probability that we would obtain a particular value of X if the null hypothesis were actually true. Now, let's suppose that we chose a critical region that covers 20% of the sampling distribution, and suppose that the null hypothesis is actually true. What would be the probability of incorrectly rejecting the null? The answer is of course 20%. And therefore, we would have built a test that had an α level of 0.2. If we want $\alpha = .05$, the critical region is only *allowed* to cover 5% of the sampling distribution of our test statistic.

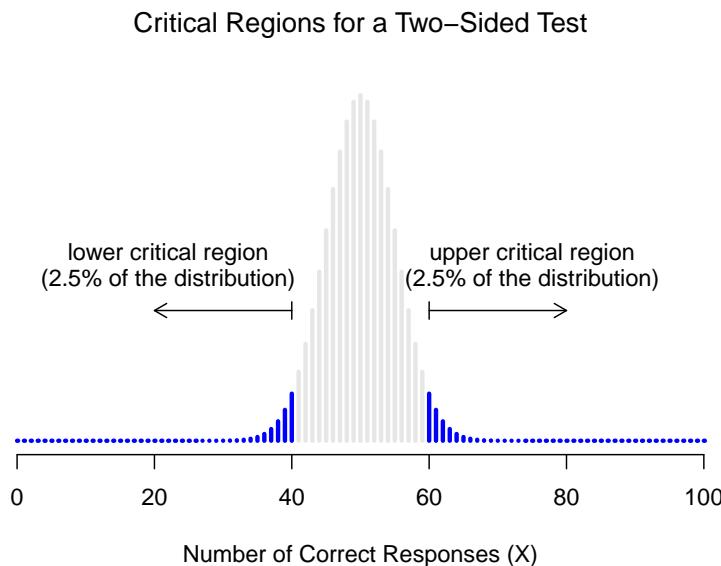


Figure 5.2: The critical region associated with the hypothesis test for the ESP study, for a hypothesis test with a significance level of $\alpha = .05$. The plot itself shows the sampling distribution of X under the null hypothesis: the grey bars correspond to those values of X for which we would retain the null hypothesis. The black bars show the critical region: those values of X for which we would reject the null. Because the alternative hypothesis is two sided (i.e., allows both $\theta < .5$ and $\theta > .5$), the critical region covers both tails of the distribution. To ensure an α level of .05, we need to ensure that each of the two regions encompasses 2.5% of the sampling distribution.

As it turns out, those three things uniquely solve the problem: our critical region consists of the most *extreme values*, known as the **tails** of the distribution. This is illustrated in Figure 5.2. As it turns out, if we want $\alpha = .05$, then our critical regions correspond to $X \leq 40$ and $X \geq 60$.⁶ That is, if the number of people saying “true” is between 41 and 59, then we should retain the null hypothesis.

⁶Strictly speaking, the test I just constructed has $\alpha = .057$, which is a bit too generous.

If the number is between 0 to 40 or between 60 to 100, then we should reject the null hypothesis. The numbers 40 and 60 are often referred to as the *critical values*, since they define the edges of the critical region.

At this point, our hypothesis test is essentially complete: (1) we choose an α level (e.g., $\alpha = .05$), (2) come up with some test statistic (e.g., X) that does a good job (in some meaningful sense) of comparing H_0 to H_1 , (3) figure out the sampling distribution of the test statistic on the assumption that the null hypothesis is true (in this case, binomial) and then (4) calculate the critical region that produces an appropriate α level (0-40 and 60-100). All that we have to do now is calculate the value of the test statistic for the real data (e.g., $X = 62$) and then compare it to the critical values to make our decision. Since 62 is greater than the critical value of 60, we would reject the null hypothesis. Or, to phrase it slightly differently, we say that the test has produced a *significant* result.

5.6.2 A note on statistical “significance”

Like other occult techniques of divination, the statistical method has a private jargon deliberately contrived to obscure its methods from non-practitioners.

– Attributed to G. O. Ashley⁷

A very brief digression is in order at this point, regarding the word “significant”. The concept of statistical significance is actually a very simple one, but has a very unfortunate name. If the data allow us to reject the null hypothesis, we say that “the result is *statistically significant*”, which is often shortened to “the result is significant”. This terminology is rather old, and dates back to a time when “significant” just meant something like “indicated”, rather than its modern meaning, which is much closer to “important”. As a result, a lot of modern readers get very confused when they start learning statistics, because they think that a “significant result” must be an important one. It doesn’t mean that at all. All that “statistically significant” means is that the data allowed us to reject a null hypothesis. Whether or not the result is actually important in the real world is a very different question, and depends on all sorts of other things.

However, if I’d chosen 39 and 61 to be the boundaries for the critical region, then the critical region only covers 3.5% of the distribution. I figured that it makes more sense to use 40 and 60 as my critical values, and be willing to tolerate a 5.7% type I error rate, since that’s as close as I can get to a value of $\alpha = .05$.

⁷The internet seems fairly convinced that Ashley said this, though I can’t for the life of me find anyone willing to give a source for the claim.

5.6.3 The difference between one sided and two sided tests

There's one more thing I want to point out about the hypothesis test that I've just constructed. If we take a moment to think about the statistical hypotheses I've been using,

$$\begin{aligned} H_0 : \theta &= .5 \\ H_1 : \theta &\neq .5 \end{aligned}$$

we notice that the alternative hypothesis covers *both* the possibility that $\theta < .5$ and the possibility that $\theta > .5$. This makes sense if I really think that ESP could produce better-than-chance performance *or* worse-than-chance performance (and there are some people who think that). In statistical language, this is an example of a ***two-sided test***. It's called this because the alternative hypothesis covers the area on both "sides" of the null hypothesis, and as a consequence the critical region of the test covers both tails of the sampling distribution (2.5% on either side if $\alpha = .05$), as illustrated earlier in Figure 5.2.

However, that's not the only possibility. It might be the case, for example, that I'm only willing to believe in ESP if it produces better than chance performance. If so, then my alternative hypothesis would only covers the possibility that $\theta > .5$, and as a consequence the null hypothesis now becomes $\theta \leq .5$:

$$\begin{aligned} H_0 : \theta &\leq .5 \\ H_1 : \theta &> .5 \end{aligned}$$

When this happens, we have what's called a ***one-sided test***, and when this happens the critical region only covers one tail of the sampling distribution. This is illustrated in Figure 5.3.

5.7 The p value of a test

In one sense, our hypothesis test is complete; we've constructed a test statistic, figured out its sampling distribution if the null hypothesis is true, and then constructed the critical region for the test. Nevertheless, I've actually omitted the most important number of all: ***the p value***. It is to this topic that we now turn. There are two somewhat different ways of interpreting a p value, one proposed by Sir Ronald Fisher and the other by Jerzy Neyman. Both versions are legitimate, though they reflect very different ways of thinking about hypothesis tests. Most introductory textbooks tend to give Fisher's version only, but I think that's a bit of a shame. To my mind, Neyman's version is cleaner, and actually better reflects the logic of the null hypothesis test. You might disagree though, so I've included both. I'll start with Neyman's version...

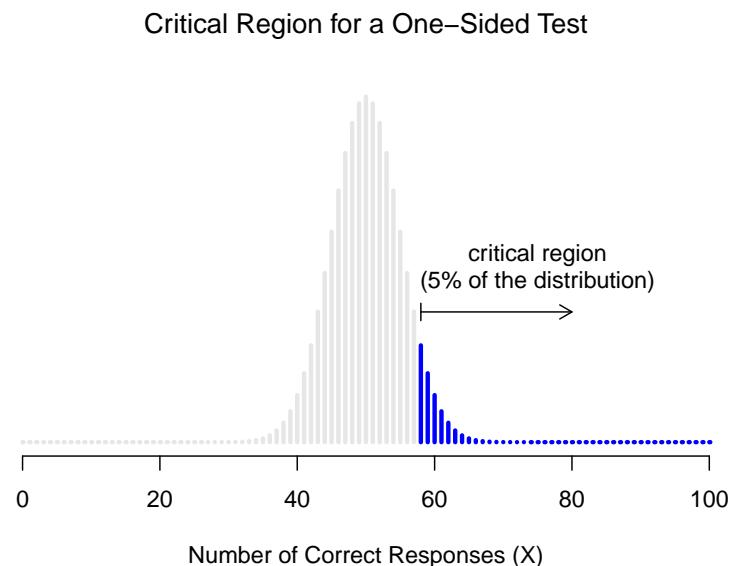


Figure 5.3: The critical region for a one sided test. In this case, the alternative hypothesis is that $\theta > .05$, so we would only reject the null hypothesis for large values of X . As a consequence, the critical region only covers the upper tail of the sampling distribution; specifically the upper 5% of the distribution. Contrast this to the two-sided version earlier)

5.7.1 A softer view of decision making

One problem with the hypothesis testing procedure that I've described is that it makes no distinction at all between a result this "barely significant" and those that are "highly significant". For instance, in my ESP study the data I obtained only just fell inside the critical region - so I did get a significant effect, but was a pretty near thing. In contrast, suppose that I'd run a study in which $X = 97$ out of my $N = 100$ participants got the answer right. This would obviously be significant too, but by a much larger margin; there's really no ambiguity about this at all. The procedure that I described makes no distinction between the two. If I adopt the standard convention of allowing $\alpha = .05$ as my acceptable Type I error rate, then both of these are significant results.

This is where the p value comes in handy. To understand how it works, let's suppose that we ran lots of hypothesis tests on the same data set: but with a different value of α in each case. When we do that for my original ESP data, what we'd get is something like this

Value of α	Reject the null?
0.05	Yes
0.04	Yes
0.03	Yes
0.02	No
0.01	No

When we test ESP data ($X = 62$ successes out of $N = 100$ observations) using α levels of $.03$ and above, we'd always find ourselves rejecting the null hypothesis. For α levels of $.02$ and below, we always end up retaining the null hypothesis. Therefore, somewhere between $.02$ and $.03$ there must be a smallest value of α that would allow us to reject the null hypothesis for this data. This is the p value; as it turns out the ESP data has $p = .021$. In short:

p is defined to be the smallest Type I error rate (α) that you have to be willing to tolerate if you want to reject the null hypothesis.

If it turns out that p describes an error rate that you find intolerable, then you must retain the null. If you're comfortable with an error rate equal to p , then it's okay to reject the null hypothesis in favour of your preferred alternative.

In effect, p is a summary of all the possible hypothesis tests that you could have run, taken across all possible α values. And as a consequence it has the effect of "softening" our decision process. For those tests in which $p \leq \alpha$ you would have rejected the null hypothesis, whereas for those tests in which $p > \alpha$ you would have retained the null. In my ESP study I obtained $X = 62$, and as a consequence I've ended up with $p = .021$. So the error rate I have to tolerate is 2.1%. In contrast, suppose my experiment had yielded $X = 97$. What happens to my p value now? This time it's shrunk to $p = 1.36 \times 10^{-25}$, which is a tiny,

tiny⁸ Type I error rate. For this second case I would be able to reject the null hypothesis with a lot more confidence, because I only have to be “willing” to tolerate a type I error rate of about 1 in 10 trillion trillion in order to justify my decision to reject.

5.7.2 The probability of extreme data

The second definition of the p -value comes from Sir Ronald Fisher, and it's actually this one that you tend to see in most introductory statistics textbooks. Notice how, when I constructed the critical region, it corresponded to the *tails* (i.e., extreme values) of the sampling distribution? That's not a coincidence: almost all "good" tests have this characteristic (good in the sense of minimising our type II error rate, β). The reason for that is that a good critical region almost always corresponds to those values of the test statistic that are least likely to be observed if the null hypothesis is true. If this rule is true, then we can define the p -value as the probability that we would have observed a test statistic that is at least as extreme as the one we actually did get. In other words, if the data are extremely implausible according to the null hypothesis, then the null hypothesis is probably wrong.

5.7.3 A common mistake

Okay, so you can see that there are two rather different but legitimate ways to interpret the p value, one based on Neyman's approach to hypothesis testing and the other based on Fisher's. Unfortunately, there is a third explanation that people sometimes give, especially when they're first learning statistics, and it is *absolutely and completely wrong*. This mistaken approach is to refer to the p value as "the probability that the null hypothesis is true". It's an intuitively appealing way to think, but it's wrong in two key respects: (1) null hypothesis testing is a frequentist tool, and the frequentist approach to probability does *not* allow you to assign probabilities to the null hypothesis... according to this view of probability, the null hypothesis is either true or it is not; it cannot have a "5% chance" of being true. (2) even within the Bayesian approach, which does let you assign probabilities to hypotheses, the p value would not correspond to the probability that the null is true; this interpretation is entirely inconsistent with the mathematics of how the p value is calculated. Put bluntly, despite the intuitive appeal of thinking this way, there is *no* justification for interpreting a p value this way. Never do it.

5.8 Reporting the results of a hypothesis test

When writing up the results of a hypothesis test, there's usually several pieces of information that you need to report, but it varies a fair bit from test to test. Throughout the rest of the book I'll spend a little time talking about how to report the results of different tests, so that you can get a feel for how it's usually done. However, regardless of what test you're doing, the one thing that you always have to do is say something about the p value, and whether or not the outcome was significant.

The fact that you have to do this is unsurprising; it's the whole point of doing the test. What might be surprising is the fact that there is some contention over exactly how you're supposed to do it. Leaving aside those people who completely disagree with the entire framework underpinning null hypothesis testing, there's a certain amount of tension that exists regarding whether or not to report the exact p value that you obtained, or if you should state only that $p < \alpha$ for a significance level that you chose in advance (e.g., $p < .05$).

5.8.1 The issue

To see why this is an issue, the key thing to recognise is that p values are *terribly* convenient. In practice, the fact that we can compute a p value means that we don't actually have to specify any α level at all in order to run the test. Instead, what you can do is calculate your p value and interpret it directly: if you get $p = .062$, then it means that you'd have to be willing to tolerate a Type I error rate of 6.2% to justify rejecting the null. If you personally find 6.2% intolerable, then you retain the null. Therefore, the argument goes, why don't we just report the actual p value and let the reader make up their own minds about what an acceptable Type I error rate is? This approach has the big advantage of "softening" the decision making process – in fact, if you accept the Neyman definition of the p value, that's the whole point of the p value. We no longer have a fixed significance level of $\alpha = .05$ as a bright line separating "accept" from "reject" decisions; and this removes the rather pathological problem of being forced to treat $p = .051$ in a fundamentally different way to $p = .049$.

This flexibility is both the advantage and the disadvantage to the p value. The reason why a lot of people don't like the idea of reporting an exact p value is that it gives the researcher a bit *too much* freedom. In particular, it lets you change your mind about what error tolerance you're willing to put up with *after* you look at the data. For instance, consider my ESP experiment. Suppose I ran my test, and ended up with a p value of .09. Should I accept or reject? Now, to be honest, I haven't yet bothered to think about what level of Type I error I'm "really" willing to accept. I don't have an opinion on that topic. But I *do* have an opinion about whether or not ESP exists, and I *definitely* have an opinion about whether my research should be published in a reputable scientific journal. And amazingly, now that I've looked at the data I'm starting to think that a

Table 5.1: A commonly adopted convention for reporting p values: in many places it is conventional to report one of four different things (e.g., $p < .05$) as shown below. I've included the "significance stars" notation (i.e., a * indicates $p < .05$) because you sometimes see this notation produced by statistical software. It's also worth noting that some people will write *n.s.* (not significant) rather than $p > .05$.

Usual notation	Signif. stars	Signif. stars
$p>.05$	NA	The test wasn't significant
$p<.05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$
$p<.01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$
$p<.001$	***	The test was significant at all levels

9% error rate isn't so bad, especially when compared to how annoying it would be to have to admit to the world that my experiment has failed. So, to avoid looking like I just made it up after the fact, I now say that my α is .1: a 10% type I error rate isn't too bad, and at that level my test is significant! I win.

In other words, the worry here is that I might have the best of intentions, and be the most honest of people, but the temptation to just "shade" things a little bit here and there is really, really strong. As anyone who has ever run an experiment can attest, it's a long and difficult process, and you often get *very* attached to your hypotheses. It's hard to let go and admit the experiment didn't find what you wanted it to find. And that's the danger here. If we use the "raw" p -value, people will start interpreting the data in terms of what they *want* to believe, not what the data are actually saying... and if we allow that, well, why are we bothering to do science at all? Why not let everyone believe whatever they like about anything, regardless of what the facts are? Okay, that's a bit extreme, but that's where the worry comes from. According to this view, you really *must* specify your α value in advance, and then only report whether the test was significant or not. It's the only way to keep ourselves honest.

5.8.2 Two proposed solutions

In practice, it's pretty rare for a researcher to specify a single α level ahead of time. Instead, the convention is that scientists rely on three standard significance levels: .05, .01 and .001. When reporting your results, you indicate which (if any) of these significance levels allow you to reject the null hypothesis. This is summarised in Table 5.1. This allows us to soften the decision rule a little bit, since $p < .01$ implies that the data meet a stronger evidentiary standard than $p < .05$ would. Nevertheless, since these levels are fixed in advance by convention, it does prevent people choosing their α level after looking at the data.

Nevertheless, quite a lot of people still prefer to report exact p values. To many

people, the advantage of allowing the reader to make up their own mind about how to interpret $p = .06$ outweighs any disadvantages. In practice, however, even among those researchers who prefer exact p values it is quite common to just write $p < .001$ instead of reporting an exact value for small p . This is in part because a lot of software doesn't actually print out the p value when it's that small (e.g., SPSS just writes $p = .000$ whenever $p < .001$), and in part because a very small p value can be kind of misleading. The human mind sees a number like $.0000000001$ and it's hard to suppress the gut feeling that the evidence in favour of the alternative hypothesis is a near certainty. In practice however, this is usually wrong. Life is a big, messy, complicated thing: and every statistical test ever invented relies on simplifications, approximations and assumptions. As a consequence, it's probably not reasonable to walk away from *any* statistical analysis with a feeling of confidence stronger than $p < .001$ implies. In other words, $p < .001$ is really code for “as far as *this test* is concerned, the evidence is overwhelming.”

In light of all this, you might be wondering exactly what you should do. There's a fair bit of contradictory advice on the topic, with some people arguing that you should report the exact p value, and other people arguing that you should use the tiered approach illustrated in Table 5.1. As a result, the best advice I can give is to suggest that you look at papers/reports written in your field and see what the convention seems to be. If there doesn't seem to be any consistent pattern, then use whichever method you prefer.

5.9 Running the hypothesis test in practice

At this point some of you might be wondering if this is a “real” hypothesis test, or just a toy example that I made up. It's real. In the previous discussion I built the test from first principles, thinking that it was the simplest possible problem that you might ever encounter in real life. However, this test already exists: it's called the *binomial test*, and it's implemented by an R function called `binom.test()`. To test the null hypothesis that the response probability is one-half $p = .5$,⁹ using data in which $x = 62$ of $n = 100$ people made the correct response, here's how to do it in R:

```
binom.test( x=62, n=100, p=.5 )
```

```
##  
## Exact binomial test  
##  
## data: 62 and 100
```

⁹Note that the `p` here has nothing to do with a p value. The `p` argument in the `binom.test()` function corresponds to the probability of making a correct response, according to the null hypothesis. In other words, it's the θ value.

```

## number of successes = 62, number of trials = 100, p-value =
## 0.02098
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
## 0.5174607 0.7152325
## sample estimates:
## probability of success
## 0.62

```

Right now, this output looks pretty unfamiliar to you, but you can see that it's telling you more or less the right things. Specifically, the p -value of 0.02 is less than the usual choice of $\alpha = .05$, so you can reject the null. We'll talk a lot more about how to read this sort of output as we go along; and after a while you'll hopefully find it quite easy to read and understand. For now, however, I just wanted to make the point that R contains a whole lot of functions corresponding to different kinds of hypothesis test. And while I'll usually spend quite a lot of time explaining the logic behind how the tests are built, every time I discuss a hypothesis test the discussion will end with me showing you a fairly simple R command that you can use to run the test in practice.

5.10 Effect size, sample size and power

In previous sections I've emphasised the fact that the major design principle behind statistical hypothesis testing is that we try to control our Type I error rate. When we fix $\alpha = .05$ we are attempting to ensure that only 5% of true null hypotheses are incorrectly rejected. However, this doesn't mean that we don't care about Type II errors. In fact, from the researcher's perspective, the error of failing to reject the null when it is actually false is an extremely annoying one. With that in mind, a secondary goal of hypothesis testing is to try to minimise β , the Type II error rate, although we don't usually *talk* in terms of minimising Type II errors. Instead, we talk about maximising the *power* of the test. Since power is defined as $1 - \beta$, this is the same thing.

5.10.1 The power function

Let's take a moment to think about what a Type II error actually is. A Type II error occurs when the alternative hypothesis is true, but we are nevertheless unable to reject the null hypothesis. Ideally, we'd be able to calculate a single number β that tells us the Type II error rate, in the same way that we can set $\alpha = .05$ for the Type I error rate. Unfortunately, this is a lot trickier to do. To see this, notice that in my ESP study the alternative hypothesis actually corresponds to lots of possible values of θ . In fact, the alternative hypothesis corresponds to every value of θ *except* 0.5. Let's suppose that the

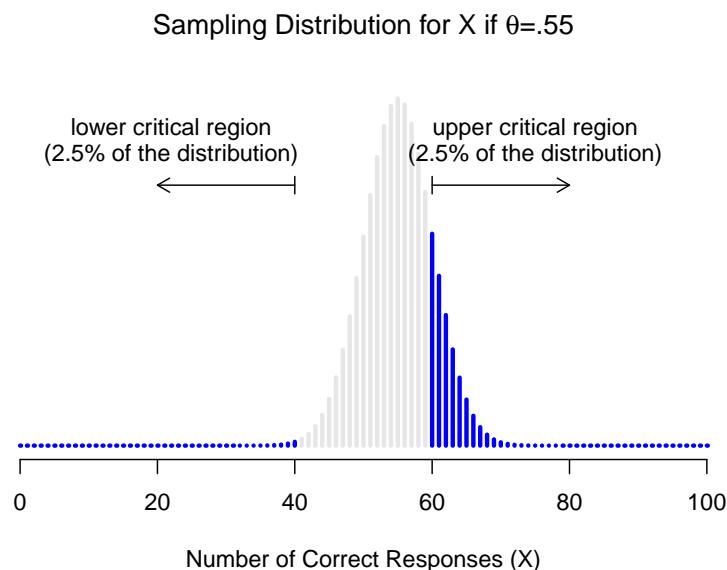


Figure 5.4: Sampling distribution under the *alternative hypothesis*, for a population parameter value of $\theta = 0.55$. A reasonable proportion of the distribution lies in the rejection region.

true probability of someone choosing the correct response is 55% (i.e., $\theta = .55$). If so, then the *true* sampling distribution for X is not the same one that the null hypothesis predicts: the most likely value for X is now 55 out of 100. Not only that, the whole sampling distribution has now shifted, as shown in Figure 5.4. The critical regions, of course, do not change: by definition, the critical regions are based on what the null hypothesis predicts. What we're seeing in this figure is the fact that when the null hypothesis is wrong, a much larger proportion of the sampling distribution falls in the critical region. And of course that's what should happen: the probability of rejecting the null hypothesis is larger when the null hypothesis is actually false! However $\theta = .55$ is not the only possibility consistent with the alternative hypothesis. Let's instead suppose that the true value of θ is actually 0.7. What happens to the sampling distribution when this occurs? The answer, shown in Figure 5.5, is that almost the entirety of the sampling distribution has now moved into the critical region. Therefore, if $\theta = 0.7$ the probability of us correctly rejecting the null hypothesis (i.e., the power of the test) is much larger than if $\theta = 0.55$. In short, while $\theta = .55$ and $\theta = .70$ are both part of the alternative hypothesis, the Type II error rate is different.

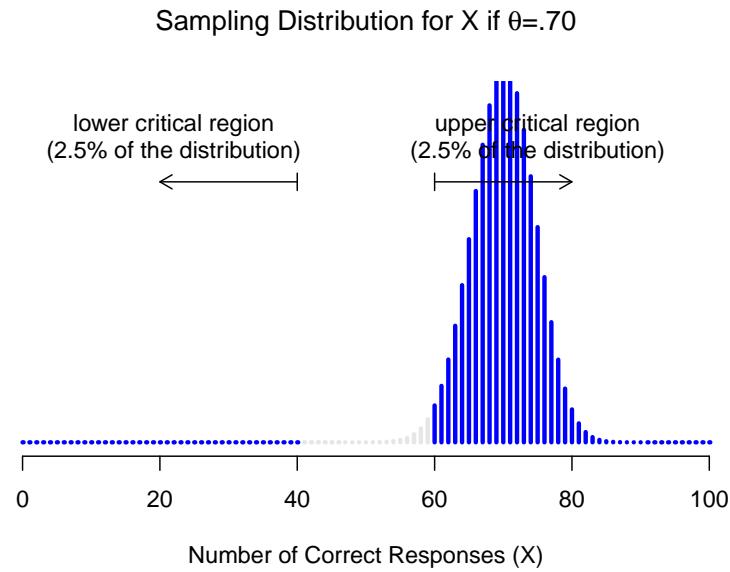


Figure 5.5: Sampling distribution under the *alternative* hypothesis, for a population parameter value of $\theta = 0.70$. Almost all of the distribution lies in the rejection region.

What all this means is that the power of a test (i.e., $1 - \beta$) depends on the true value of θ . To illustrate this, I've calculated the expected probability of

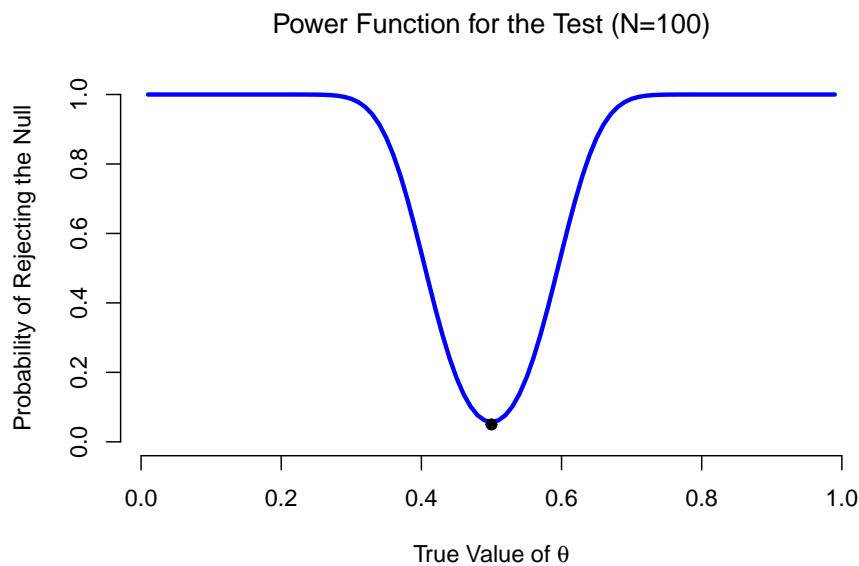


Figure 5.6: The probability that we will reject the null hypothesis, plotted as a function of the true value of θ . Obviously, the test is more powerful (greater chance of correct rejection) if the true value of θ is very different from the value that the null hypothesis specifies (i.e., $\theta = .5$). Notice that when θ actually is equal to .5 (plotted as a black dot), the null hypothesis is in fact true: rejecting the null hypothesis in this instance would be a Type I error.

rejecting the null hypothesis for all values of θ , and plotted it in Figure 5.6. This plot describes what is usually called the ***power function*** of the test. It's a nice summary of how good the test is, because it actually tells you the power ($1 - \beta$) for all possible values of θ . As you can see, when the true value of θ is very close to 0.5, the power of the test drops very sharply, but when it is further away, the power is large.

5.10.2 Effect size

Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned with mice when there are tigers abroad

– George Box 1976

The plot shown in Figure 5.6 captures a fairly basic point about hypothesis testing. If the true state of the world is very different from what the null hypothesis predicts, then your power will be very high; but if the true state of the world is similar to the null (but not identical) then the power of the test is going to be very low. Therefore, it's useful to be able to have some way of quantifying how "similar" the true state of the world is to the null hypothesis. A statistic that does this is called a measure of ***effect size*** (e.g. Cohen, 1988; Ellis, 2010). Effect size is defined slightly differently in different contexts,¹⁰ (and so this section just talks in general terms) but the qualitative idea that it tries to capture is always the same: how big is the difference between the *true* population parameters, and the parameter values that are assumed by the null hypothesis? In our ESP example, if we let $\theta_0 = 0.5$ denote the value assumed by the null hypothesis, and let θ denote the true value, then a simple measure of effect size could be something like the difference between the true value and null (i.e., $\theta - \theta_0$), or possibly just the magnitude of this difference, $\text{abs}(\theta - \theta_0)$.

	big effect size	small effect size
significant result	difference is real, and of practical importance	difference is real, but might not be
non-significant result	no effect observed	no effect observed

Why calculate effect size? Let's assume that you've run your experiment, collected the data, and gotten a significant effect when you ran your hypothesis test. Isn't it enough just to say that you've gotten a significant effect? Surely that's the *point* of hypothesis testing? Well, sort of. Yes, the point of doing a hypothesis test is to try to demonstrate that the null hypothesis is wrong, but that's hardly the only thing we're interested in. If the null hypothesis claimed that $\theta = .5$, and we show that it's wrong, we've only really told half of the story. Rejecting the null hypothesis implies that we believe that $\theta \neq .5$, but there's a

¹⁰There's an R package called `compute.es` that can be used for calculating a very broad range of effect size measures; but for the purposes of the current book we won't need it: all of the effect size measures that I'll talk about here have functions in the `lsr` package

big difference between $\theta = .51$ and $\theta = .8$. If we find that $\theta = .8$, then not only have we found that the null hypothesis is wrong, it appears to be *very* wrong. On the other hand, suppose we've successfully rejected the null hypothesis, but it looks like the true value of θ is only $.51$ (this would only be possible with a large study). Sure, the null hypothesis is wrong, but it's not at all clear that we actually *care*, because the effect size is so small. In the context of my ESP study we might still care, since any demonstration of real psychic powers would actually be pretty cool¹¹, but in other contexts a 1% difference isn't very interesting, even if it is a real difference. For instance, suppose we're looking at differences in high school exam scores between males and females, and it turns out that the female scores are 1% higher on average than the males. If I've got data from thousands of students, then this difference will almost certainly be *statistically significant*, but regardless of how small the p value is it's just not very interesting. You'd hardly want to go around proclaiming a crisis in boys education on the basis of such a tiny difference would you? It's for this reason that it is becoming more standard (slowly, but surely) to report some kind of standard measure of effect size along with the results of the hypothesis test. The hypothesis test itself tells you whether you should believe that the effect you have observed is real (i.e., not just due to chance); the effect size tells you whether or not you should care.

5.10.3 Increasing the power of your study

Not surprisingly, scientists are fairly obsessed with maximising the power of their experiments. We want our experiments to work, and so we want to maximise the chance of rejecting the null hypothesis if it is false (and of course we usually want to believe that it is false!) As we've seen, one factor that influences power is the effect size. So the first thing you can do to increase your power is to increase the effect size. In practice, what this means is that you want to design your study in such a way that the effect size gets magnified. For instance, in my ESP study I might believe that psychic powers work best in a quiet, darkened room; with fewer distractions to cloud the mind. Therefore I would try to conduct my experiments in just such an environment: if I can strengthen people's ESP abilities somehow, then the true value of θ will go up¹² and therefore my effect size will be larger. In short, clever experimental design is one way to boost power; because it can alter the effect size.

Unfortunately, it's often the case that even with the best of experimental de-

¹¹Although in practice a very small effect size is worrying, because even very minor methodological flaws might be responsible for the effect; and in practice no experiment is perfect, so there are always methodological issues to worry about.

¹²Notice that the true population parameter θ doesn't necessarily correspond to an immutable fact of nature. In this context θ is just the true probability that people would correctly guess the colour of the card in the other room. As such the population parameter can be influenced by all sorts of things. Of course, this is all on the assumption that ESP actually exists!

signs you may have only a small effect. Perhaps, for example, ESP really does exist, but even under the best of conditions it's very very weak. Under those circumstances, your best bet for increasing power is to increase the sample size. In general, the more observations that you have available, the more likely it is that you can discriminate between two hypotheses. If I ran my ESP experiment with 10 participants, and 7 of them correctly guessed the colour of the hidden card, you wouldn't be terribly impressed. But if I ran it with 10,000 participants and 7,000 of them got the answer right, you would be much more likely to think I had discovered something. In other words, power increases with the sample size. This is illustrated in Figure 5.7, which shows the power of the test for a true parameter of $\theta = 0.7$, for all sample sizes N from 1 to 100, where I'm assuming that the null hypothesis predicts that $\theta_0 = 0.5$.

```
## [1] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.11837800
## [7] 0.08257300 0.05771362 0.19643626 0.14945203 0.11303734 0.25302172
## [13] 0.20255096 0.16086106 0.29695959 0.24588947 0.38879291 0.33269435
## [19] 0.28223844 0.41641377 0.36272868 0.31341925 0.43996501 0.38859619
## [25] 0.51186665 0.46049782 0.41129777 0.52752694 0.47870819 0.58881596
## [31] 0.54162450 0.49507894 0.59933871 0.55446069 0.65155826 0.60907715
## [37] 0.69828554 0.65867614 0.61815357 0.70325017 0.66542910 0.74296156
## [43] 0.70807163 0.77808343 0.74621569 0.71275488 0.78009449 0.74946571
## [49] 0.81000236 0.78219322 0.83626633 0.81119597 0.78435605 0.83676444
## [55] 0.81250680 0.85920268 0.83741123 0.87881491 0.85934395 0.83818214
## [61] 0.87858194 0.85962510 0.89539581 0.87849413 0.91004390 0.89503851
## [67] 0.92276845 0.90949768 0.89480727 0.92209753 0.90907263 0.93304809
## [73] 0.92153987 0.94254237 0.93240638 0.92108426 0.94185449 0.93185881
## [79] 0.95005094 0.94125189 0.95714694 0.94942195 0.96327866 0.95651332
## [85] 0.94886329 0.96265653 0.95594208 0.96796884 0.96208909 0.97255504
## [91] 0.96741721 0.97650832 0.97202770 0.97991117 0.97601093 0.97153910
## [97] 0.97944717 0.97554675 0.98240749 0.97901142
```

Because power is important, whenever you're contemplating running an experiment it would be pretty useful to know how much power you're likely to have. It's never possible to know for sure, since you can't possibly know what your effect size is. However, it's often (well, sometimes) possible to guess how big it should be. If so, you can guess what sample size you need! This idea is called **power analysis**, and if it's feasible to do it, then it's very helpful, since it can tell you something about whether you have enough time or money to be able to run the experiment successfully. It's increasingly common to see people arguing that power analysis should be a required part of experimental design, so it's worth knowing about.

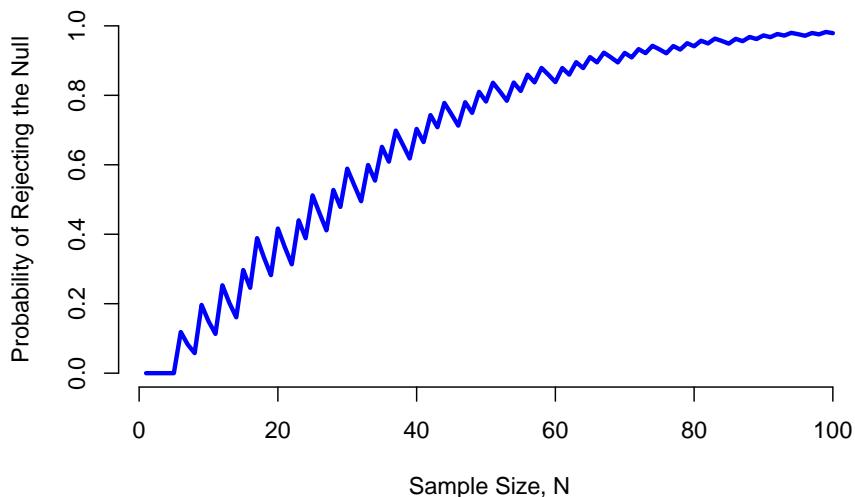


Figure 5.7: The power of our test, plotted as a function of the sample size N . In this case, the true value of θ is 0.7, but the null hypothesis is that $\theta = 0.5$. Overall, larger N means greater power. (The small zig-zags in this function occur because of some odd interactions between θ , α and the fact that the binomial distribution is discrete; it doesn't matter for any serious purpose)

5.11 Some issues to consider

What I've described to you in this chapter is the orthodox framework for null hypothesis significance testing (NHST). Understanding how NHST works is an absolute necessity, since it has been the dominant approach to inferential statistics ever since it came to prominence in the early 20th century. It's what the vast majority of working scientists rely on for their data analysis, so even if you hate it you need to know it. However, the approach is not without problems. There are a number of quirks in the framework, historical oddities in how it came to be, theoretical disputes over whether or not the framework is right, and a lot of practical traps for the unwary. I'm not going to go into a lot of detail on this topic, but I think it's worth briefly discussing a few of these issues.

5.11.1 Neyman versus Fisher

The first thing you should be aware of is that orthodox NHST is actually a mash-up of two rather different approaches to hypothesis testing, one proposed by Sir Ronald Fisher and the other proposed by Jerzy Neyman (for a historical summary see Lehmann, 2011). The history is messy because Fisher and Neyman were real people whose opinions changed over time, and at no point did either of them offer "the definitive statement" of how we should interpret their work many decades later. That said, here's a quick summary of what I take these two approaches to be.

First, let's talk about Fisher's approach. As far as I can tell, Fisher assumed that you only had the one hypothesis (the null), and what you want to do is find out if the null hypothesis is inconsistent with the data. From his perspective, what you should do is check to see if the data are "sufficiently unlikely" according to the null. In fact, if you remember back to our earlier discussion, that's how Fisher defines the *p*-value. According to Fisher, if the null hypothesis provided a very poor account of the data, you could safely reject it. But, since you don't have any other hypotheses to compare it to, there's no way of "accepting the alternative" because you don't necessarily have an explicitly stated alternative. That's more or less all that there was to it.

In contrast, Neyman thought that the point of hypothesis testing was as a guide to action, and his approach was somewhat more formal than Fisher's. His view was that there are multiple things that you could *do* (accept the null or accept the alternative) and the point of the test was to tell you which one the data support. From this perspective, it is critical to specify your alternative hypothesis properly. If you don't know what the alternative hypothesis is, then you don't know how powerful the test is, or even which action makes sense. His framework genuinely requires a competition between different hypotheses. For Neyman, the *p* value didn't directly measure the probability of the data (or data more extreme) under the null, it was more of an abstract description about

which “possible tests” were telling you to accept the null, and which “possible tests” were telling you to accept the alternative.

As you can see, what we have today is an odd mishmash of the two. We talk about having both a null hypothesis and an alternative (Neyman), but usually¹³ define the p value in terms of extreme data (Fisher), but we still have α values (Neyman). Some of the statistical tests have explicitly specified alternatives (Neyman) but others are quite vague about it (Fisher). And, according to some people at least, we’re not allowed to talk about accepting the alternative (Fisher). It’s a mess: but I hope this at least explains why it’s a mess.

5.11.2 Bayesians versus frequentists

Earlier on in this chapter I was quite emphatic about the fact that you *cannot* interpret the p value as the probability that the null hypothesis is true. NHST is fundamentally a frequentist tool (see Chapter 4.2) and as such it does not allow you to assign probabilities to hypotheses: the null hypothesis is either true or it is not. The Bayesian approach to statistics interprets probability as a degree of belief, so it’s totally okay to say that there is a 10% chance that the null hypothesis is true: that’s just a reflection of the degree of confidence that you have in this hypothesis. **You aren’t allowed to do this within the frequentist approach.** Remember, if you’re a frequentist, a probability can only be defined in terms of what happens after a large number of independent replications (i.e., a long run frequency). If this is your interpretation of probability, talking about the “probability” that the null hypothesis is true is complete gibberish: a null hypothesis is either true or it is false. There’s no way you can talk about a long run frequency for this statement. To talk about “the probability of the null hypothesis” is as meaningless as “the colour of freedom”. It doesn’t have one!

Most importantly, this *isn’t* a purely ideological matter. If you decide that you are a Bayesian and that you’re okay with making probability statements about hypotheses, you have to follow the Bayesian rules for calculating those probabilities. For now what I want to point out to you is the p value is a *terrible* approximation to the probability that H_0 is true. If what you want to know is the probability of the null, then the p value is not what you’re looking for!

5.11.3 Traps

As you can see, the theory behind hypothesis testing is a mess, and even now there are arguments in statistics about how it “should” work. However, disagreements among statisticians are not our real concern here. Our real concern

¹³Although this book describes both Neyman’s and Fisher’s definition of the p value, most don’t. Most introductory textbooks will only give you the Fisher version.

is practical data analysis. And while the “orthodox” approach to null hypothesis significance testing has many drawbacks, even an unrepentant Bayesian like myself would agree that they can be useful if used responsibly. Most of the time they give sensible answers, and you can use them to learn interesting things. Setting aside the various ideologies and historical confusions that we’ve discussed, the fact remains that the biggest danger in all of statistics is *thoughtlessness*. I don’t mean stupidity, here: I literally mean thoughtlessness. The rush to interpret a result without spending time thinking through what each test actually says about the data, and checking whether that’s consistent with how you’ve interpreted it. That’s where the biggest trap lies.

To give an example of this, consider the following example (see Gelman and Stern, 2006). Suppose I’m running my ESP study, and I’ve decided to analyse the data separately for the male participants and the female participants. Of the male participants, 33 out of 50 guessed the colour of the card correctly. This is a significant effect ($p = .03$). Of the female participants, 29 out of 50 guessed correctly. This is not a significant effect ($p = .32$). Upon observing this, it is extremely tempting for people to start wondering why there is a difference between males and females in terms of their psychic abilities. However, this is wrong. If you think about it, we haven’t *actually* run a test that explicitly compares males to females. All we have done is compare males to chance (binomial test was significant) and compared females to chance (binomial test was non significant). If we want to argue that there is a real difference between the males and the females, we should probably run a test of the null hypothesis that there is no difference! We can do that using a different hypothesis test,¹⁴ but when we do that it turns out that we have no evidence that males and females are significantly different ($p = .54$). Now do you think that there’s anything fundamentally different between the two groups? Of course not. What’s happened here is that the data from both groups (male and female) are pretty borderline: by pure chance, one of them happened to end up on the magic side of the $p = .05$ line, and the other one didn’t. That doesn’t actually imply that males and females are different. This mistake is so common that you should always be wary of it: the difference between significant and not-significant is *not* evidence of a real difference – if you want to say that there’s a difference between two groups, then you have to test for that difference!

The example above is just that: an example. I’ve singled it out because it’s such a common one, but the bigger picture is that data analysis can be tricky to get right. Think about *what* it is you want to test, *why* you want to test it, and whether or not the answers that your test gives could possibly make any sense in the real world.

¹⁴In this case, the Pearson chi-square test of independence (Chapter ??; `chisq.test()` in R) is what we use; see also the `prop.test()` function.

5.12 Summary

Null hypothesis testing is one of the most ubiquitous elements to statistical theory. The vast majority of scientific papers report the results of some hypothesis test or another. As a consequence it is almost impossible to get by in science without having at least a cursory understanding of what a p -value means, making this one of the most important chapters in the book. As usual, I'll end the chapter with a quick recap of the key ideas that we've talked about:

- Research hypotheses and statistical hypotheses. Null and alternative hypotheses. (Section 5.3).
- Type 1 and Type 2 errors (Section 5.4)
- Test statistics and sampling distributions (Section 5.5)
- Hypothesis testing as a decision making process (Section 5.6)
- p -values as “soft” decisions (Section 5.7)
- Writing up the results of a hypothesis test (Section 5.8)
- Effect size and power (Section 5.10)
- A few issues to consider regarding hypothesis testing (Section 5.11)

Chapter 6

Issues in Hypothesis Testing

Text by David Schuster

6.1 Videos

- Video: Introduction to power in significance tests
- Video: P-Hacking
- Video: Publication bias
- Video: Overview of Scholarly Publishing

6.2 Introduction

Dave here. In this chapter, I want to dive deeper into the issues involved in the practice of null hypothesis significance testing (NHST) as it is used in science. In the previous chapter, we introduced hypothesis testing and saw how it is supposed to work and how it should be interpreted. Here, we will see that NHST is not a neutral weighing of evidence; rather, the researcher affects NHST outcomes. Following that, we will talk about misunderstandings of NHST so that you do not carry the same misunderstandings with you from this course. Knowing common NHST misunderstandings will also help you when you inevitably encounter classmates, colleagues, and (hopefully not) journal and conference authors make the same mistakes. This chapter concludes with limitations that are inherent to NHST and some conclusions.

6.3 The researcher affects NHST outcomes

Some ways the researcher affects NHST outcomes is strategic and good because it involves more efficient science. Let's start with those more positive ways the researcher can affect NHST.

Researchers want to investigate effects that are present, detect them, and then be able to label them statistically significant. The question a researcher should ask themselves is, "assuming there really is an effect there, what are my chances of getting a significant result." We have already seen the name for that concept, it is statistical power. Statistical power only applies to effects that exist, so we want to maximize it as much as possible. A high probability of finding effects that are present is generally good; running studies where the probability of finding true effects is low is a waste of time and resources. There are many ways researchers can increase statistical power, and we will discover ways to increase power throughout the rest of this course. For now, we will consider a few general approaches:

1. **Increase sample size.** We saw the impact of N on standard error; because it is in the denominator of the standard error formula, larger sample sizes reduce standard error. If you observe the same effect size a second time with a larger sample, the p-value will be lower. Later on we will see that some tests perform better when the sample size for each condition is equal. Despite that, a larger sample size is typically better (at least until you get above a 2:1 ratio), even if it means you have to unbalance your groups. The strategy (and difficulty) of increasing your sample size depends on the research design.
2. **Increase the effect size.** Wait, how can you increase the effect size if that is what you are trying to estimate? This might be better phrased as "increase the observed effect size." In other words, design the study so that you have a better opportunity to see the effects. One way to do this is to increase the size of the treatment, creating a bigger difference between the levels of the independent variable. In a drug study, this could mean giving a stronger dose of the drug under study. As another example, consider a study of human-robot interaction where the independent variable under investigation is the use of multiple robots. If researchers hypothesize that a task is more difficult as people simultaneously use more robots, then they might observe a stronger treatment effect with conditions of 1 robot and 4 robots rather than 1 robot versus 2 robots. I am including this example to show that treatment effects apply across all manipulated IVs.
3. **Use a one-tailed test when appropriate.** Some statistical tests (including z and t) are two-tailed, meaning they support testing either one-tailed or two-tailed hypotheses. You may want to reread that, as the terminology is confusing. F is a one-tailed statistic, meaning that all F -test

(i.e., ANOVA) hypotheses are two-tailed. When I refer to the tailedness of a statistic, I mean whether the distribution has two tails on either side like the normal distribution or just one. When there are two tails present, you can choose to test in just one of them (which is a one-tailed test) or both (which is a two-tailed test). Statistics aside, the default is a two-tailed test. It will have lower statistical power than a one-tailed test whenever you can accurately predict the direction of the effect. A good way to decide is **before you run your analysis**. You should choose a one-tailed test if you have a logical reason for expecting (or caring) about effects in one direction only. In a correlation analysis, directionality means whether the relationship is positive (increasing together) or negative (one variable increases as the other decreases). In any test comparing means, directionality means saying which group will have a higher mean and which group a lower mean. If you have good reason to expect one group to be higher than the other, or if you really could care less if the results turned out in the opposite direction, then decide to run a one-tailed test before running your analysis. For example, in a study looking at the effectiveness of an after school program, a non-significant finding or a finding that the program reduced student achievement might be equally meaningful—in neither case would you recommend enrolling in the program. That said, be very careful to avoid drawing conclusions from null findings (non-significant). Would you want to tell a school to cancel a program all because your sample size was too low? We will explore that issue later in this chapter.

4. **Reduce variability in the population.** The lower the population standard deviation, the smaller the standard error. Here's an analogy that I am quite proud of (I believe I thought of it myself, but if you've seen it elsewhere, let me know). Have you ever made a cake? How about some soup from scratch? Imagine you are baking a cake and want to sample it to see if it is done cooking. You take a toothpick, stab the cake in a random spot, and see if the toothpick comes out clean, which would suggest the cake is done. Imagine you are making soup at the same time. You want to see if the soup is seasoned how you like it. You take a sip of the soup. Assuming you are taking random samples of cake and soup, which $N = 1$ sample is stronger? Are you more confident the cake is done, or are you more confident the soup is well-seasoned? Which one would be better with a second check? Probably the cake. If random sampling, you need a larger sample size to discover the doneness of a cake than the taste of soup. Where you place the toothpick matters, since the cake cooks from the outside in. Seasonings dissolve pretty quickly into soup and it matters less where in the soup you taste. Still with me? (I said I was proud of the analogy; I did not claim it was a great analogy) A cake needs to be sampled repeatedly because it is **less consistent** than a soup. This point of this elaborate example is this: The more consistent (less variability, lower standard deviation) a population, the easier it is to sample.

You usually have little control over this. However, population variability is something to keep in mind when planning your research. If you want to study reading comprehension of elementary school children, you will have less statistical power than the exact same study of reading comprehension in preschoolers.

5. **Choose a simpler design.** Increasing the number of experimental conditions, or choosing more complex designs (such as by adding multiple IVs to make a factorial ANOVA), decreases your statistical power.
6. **Increase the alpha level. Don't do this!** I am including this to complete the list. *alpha* is our threshold of evidence; by lowering it, we are merely lowering the bar for the magnitude of effects that we will accept as statistically significant. It would be like declaring oneself to pass a test by lowering the passing score.

This last (bad) strategy is a good transition to the more maladaptive ways researchers affect outcomes. These are not strategic; these are gaming the NHST system.

1. **P-hacking.** A significant result says, “There is less than an *alpha* (.05) chance I could obtain these results under the null hypothesis.” P-hacking occurs whenever we make decisions about which statistical test to run on the basis of our observed data. If I collected data on three variables, found correlations between all of them, then ran a statistical test on the strongest correlation, the single test would be p-hacked. There are many variations of p-hacking, including:

- Running many tests and reporting only the significant ones
- Running a test on a study in-progress and stopping data collection if it is significant (or continuing data collection if it is not ‘yet’ significant)
- Running a test, finding it to be non-significant, removing participants, and then rerunning the test
- Running a two-tailed test, finding it to be non-significant, then running the one-tailed version

What these examples have in common are decisions that are made in response to the significance test. The significance test should always be the outcome of a hypothesis and research design. Using the significance test to write a hypothesis or to change the research design is p-hacking, and its unethical. It is unethical because it makes the alpha level (and the p-value) meaningless. It inflates the Type I error rate.

The best way to avoid problems with p-hacking is to plan your analysis before the data are collected, including your choice of one-tailed or two-tailed test.

Even better would be to declare your study design and analysis ahead of time, publicly, so that there is no question you are not p-hacking. This practice is emerging but not widespread, and it's called **preregistration**.

Finally, this does not mean you cannot collect data without all hypotheses planned ahead of time. Collecting data without strong hypotheses is called exploratory research, distinguished from confirmatory research using NHST. If you are doing exploratory research, you can avoid any implication of p-hacking by admitting that. In the case of exploratory research, either do not use NHST, or make it clear that you are not interpreting and making claims about exploratory use of NHST.

Deliberate p-hacking is fraud, but subtle p-hacking has similar effects (it inflates the Type I error rate). Because of this, researchers need to know how to use statistics responsibly. This is an ethical issue because people (your coworkers, supervisors, editors, advisers, publishers) will largely trust your methods.

2. **There is more trust than suspicion of statistics in the professional world.** Related to the previous point, the only artifact that any audience of your research will see is usually your write-up. Despite some progress in open data and open science, it is still common for researchers to conduct their study, analyze all data, submit, revise, and publish without very little oversight of their statistical methods. Reviewers can and do comment on statistical methods that are described in a manuscript, but some practices, like p-hacking, are not immediately evident from an author's manuscript alone.
 - There are few formal checks on accuracy (data verification project) - statistical conclusion validity. Creeping Type I error rate is a big, largely undetectable problem. What happens in a creeping Type II error rate? Solution: Keep the implications and costs of TI TII in mind! Ethics, people will largely trust your numbers.
3. **The researcher is not an impartial party.** Researchers are incentivized to publish novel (i.e., non-replication) studies that are statistically significant. There are severe disincentives against research fraud, as it can end one's career. Beyond that, there are relatively weak safeguards against mistakes that are not evident in publication. This is also an important consideration outside of academia's incentive to publish. In industry, a researcher may be the only member of a project team with statistics proficiency. Business outcomes may depend on the appropriateness of statistical conclusions.

The name for this concept is **statistical conclusion validity**. Statistical conclusion validity is the truth of claims made from statistical results. Statistical conclusion validity is the absence of a Type I or Type II error.

6.4 NHST Misunderstandings

Next, we will look at some confusing parts of NHST.

1. **Your decision to retain or reject is all-or-nothing.** When making a statistical decision, you either reject or retain, based on the p-value. There is no grey area. There is no such thing as “highly significant” or “approaching significance.” These lead to misunderstandings of NHST.
2. **The fallacy of affirming the null is widespread.** Affirming the null is a tempting logical fallacy. Affirming the null is when a non-significant effect is taken as evidence of something. If the results of your drug trial are non-significant, you have not shown that your experimental drug has no effect. Rather, you have shown nothing; your results are inconclusive. Only rejecting the null allows conclusions to be made. For this reason, avoid the term “insignificant” because it really confuses non-researchers. Instead, use “not significant” or “did not reach significance.”
3. **The meaning of p is based on conditional probability.** p is not probability of the null being false. Since we really want to know if the null is true or false, it’s natural to think that p provides this information, but it does not. p is the probability of obtaining a sample statistic at least this extreme, assuming the null is true.

p is based on conditional probability, which confuses many people. p is a probability that already assumes the null hypothesis is true. Any statement about p should begin with “Assuming the null is true....” People fall into the trap of reversing the conditional probability when they think p is the probability of a hypothesis. Assuming I start with a brand-new deck of cards, what is the probability of drawing red? It’s 50%. Let’s reverse the conditional probability: Assuming I drew a red card from a second deck, what is the probability of the second deck being new? It’s not 50%; the probability of red depends on the deck being new, but the probability of the deck being new does not depend on drawing a red card. We make same mistake with NHST. P is probability of obtaining these data if the null were true. It is not the probability of the null being true if we obtained these data.

4. **Significance does not mean effect size. Significance does not mean importance.** Just because a result is significant does not mean it is important. For example, would you invest in an insomnia drug that has been shown to help people sleep for one additional minute per night, on average? NHST helps you decide but going back to the data is needed to interpret the real-world meaning of your results.

6.5 NHST Issues

All of the issues we have discussed so far could be solved if researchers used NHST properly. This last category of issues are inherent in NHST.

1. **Retaining the null is informative but provides no information** in NHST, leading to publication bias (also called the file drawer effect). At the same time, it is informative to know about non-significant findings. I do think it matters if this $p = .04$ publication was the first time this study has been run or if it resulted after 10 years of unpublished null findings. We simultaneously want to learn from null findings but are prohibited from drawing inferences about them by NHST.
2. **Outcomes in NHST are affected by the probability that the null is true**, which we never know. As researchers, we do not run studies unless we believe the null to be false. A significant finding when the null is highly unlikely is more likely to be spurious. However, NHST does not consider the probability of the null hypothesis. In fact, it is an elaborate method of avoiding discussing the probability of the null hypothesis. This means that significant results that are very novel or unexpected should be scrutinized more than studies that confirm well-established findings.
3. There is **nothing magical about NHST or $\alpha = .05$** . A 5% chance of significance assuming no effect exists has been established as a good threshold of evidence by convention. By allowing an alpha level of .05 we accept that some studies will lack statistical conclusion validity (i.e., they will be wrong). The cost of a lower alpha level is greater sample size and a greater Type II error rate (unless we do the work to overcome the loss of statistical power). The cost of a higher level is a greater Type I error rate.
4. **NHST can be too sensitive under some conditions.** This is a smaller problem and can be overcome if researchers understand that significance does not mean effect size nor importance. Much of NHST is developed for small sample sizes (up to a few hundred). If NHST is run on big data (many thousands), then the effect size needed to reject the null hypothesis becomes very low. When this happens, even the smallest mean difference or the weakest correlation will be labeled with a low p-value. Researchers simply need to be aware this can happen when working with large sample sizes. A good solution is to **always report effect sizes when reporting p-values** to add context. Would you buy an expensive course shown to significantly increase your IQ by 0.0000001%?

6.6 Conclusions

The issues presented in this chapter are not insurmountable. Hopefully a few themes have become clear:

- Turning NHST into knowledge depends on the statistics being used correctly. We should be skeptical of researchers who make claims from statistics but cannot explain their statistical methods.
- Statistical conclusion validity is our goal. We want the conclusions we make from our statistics to reflect truth in the population. Unfortunately, we will not always achieve this, and we will rarely know with any certainty. We should do what we can to maintain statistical conclusion validity. Because statistical conclusion validity is the absence of a Type I and Type II error, we can learn about these types of errors and the strategies to minimize them.
- **Because of these issues, we can never trust the results of a single study to definitively answer a research question.** To the lay public, the study results and knowledge science generates are the same. Because of the noise involved in statistics and research methods, we understand that we only generate knowledge from the data aggregated across many studies. In a way, every study is tentative. **Meta-analysis** is the process of drawing conclusions across multiple studies. We cannot conclude anything definitely from a single study because of the possibility of Type I and Type II error. This connects to the replication crisis—if we see it as a failure that we publish studies that lack validity (are not true), we are hoping for too much out of a single study and of NHST.

Chapter 7

Data Cleaning and Missing Values Analysis

Text by David Schuster

7.1 Videos

- Video: Data cleaning overview - 17 min
- Video: Missing data basics - 12 min
- Video: Dealing with missing data - 30 min
- Interpreting Q-Q plots - 6 min
- Outlier analysis - 16 min

7.2 Introduction: Dealing with the Unexpected

So far, we have largely proceeded under the assumption that our data will be tidy and predictable. We have seen that R is very particular about data formatting and syntax. At least in my experience conducting laboratory experimental research, the actual data never come out this way. Despite the experimentalist's emphasis on control in the design of research, working with human participants (and human researchers) means that there will be several opportunities for data to be messy. Participants will decline to answer some items; experimental apparatus will occasionally miss data or misreport it; and, anyone doing lengthy data entry will eventually make a typo.

In this chapter, we will see how we can address problems with our data with the goal of maximizing statistical conclusion validity. As researchers, we solve the problems with messy data through the steps of **data coding** and **data cleaning**. Data coding is the process of translating observations from the format in which they were collected into a format that is suitable for data analysis. For example, if you had a spreadsheet of observations with one score per column (i.e., data in many columns and one row), you would at least need to transpose that spreadsheet so that the observations were collected into one column (i.e., data in one column and many rows) for it to be imported into R as a data frame. If your research involves handwritten data, audio transcription, or interview data, data coding will be needed to turn qualitative observations into quantitative scores.

7.3 Data Cleaning

Data cleaning is a process of detecting and addressing errors in the data, including missing and erroneous scores. Errors in the data can affect statistical conclusion validity in two ways:

1. Errors in the data can affect the data **randomly** by contributing to error variance. Imagine a ruler printed on a stretchy piece of elastic. Despite your best efforts, you would probably get a slightly different measurement every time. This inconsistency is called unreliability. Unreliable measures are one possible source of random error. Another source of random error could occur if a researcher lost data accidentally and at random, which can happen with paper surveys. If a survey was lost, and the lost survey was effectively lost at random, then the missing data would contribute to random error. The effects of random error in our data are unpredictable. In most cases, higher error variance makes it harder to reject the null and thus increases the Type II error rate. Since we care about statistical power, we want to minimize the Type II error rate. However, when the true effect is near zero, unreliability can actually increase the observed effect size, which would increase the Type I error rate (Rogosa, 1980 as cited in Cook et al. (2002), 2012). Because the alpha level is our threshold of evidence, we want to minimize the Type I error rate. All this to say, we want to minimize random error in our data.
2. Errors in the data can affect the data **systematically**, resulting in **bias**. Inferential statistics helps us to make statistical conclusions despite the presence of **sampling error**, which is evident in variation across samples and is the difference between an estimate and population parameter. While inferential statistics can help us account for sampling error, our statistics still rely on random sampling. Random sampling is important for two reasons. First, random sampling is important because it allows

us to use our sample statistic for estimation, justified under the central limit theorem. Second, random sampling is important because it is **unbiased**. That is, in a random sample, the direction of sampling error (either over- or under-estimating) is unpredictable and averages to zero. Unfortunately for the researcher, it is not just the selection of participants that needs to be unbiased, it is also the representation of those participants in the data, as well. Violating this assumption can lead to **selection bias**, which is a research design problem that cannot always be solved by statistics. Imagine you were trying to estimate the prevalence of dementia in adults in a community. To avoid selection bias, participants would need to be sampled in an unbiased way. Conducting this research with only participants living in an assisted living facility would create a bias toward a greater prevalence of dementia. Conducting this research with only participants living in single-family homes would create a bias toward a lower prevalence of dementia.

So far, this is a research design issue, specifically in the selection of a sampling method. Because this issue also applies to the representation of participants in the data, an unbiased sample can also end up biased through **attrition** (leaving the study). I would imagine a psychological study about dementia and its impacts would be sensitive. If researchers began their interview with an insensitive question about dementia, participants with a history of dementia (or a family history of dementia) might be put off and decline to participate. If having a history of dementia or a family history of dementia was correlated with having dementia, then the researchers have unknowingly biased their sample by including fewer participants at risk of dementia. This problem can easily extend past the data collection when participants provide incomplete data or researchers remove participants from a data file in a biased way. Biased data can reduce the size of true effects, create effects that do not exist, or even reverse the direction of a true effect. Because of this, bias has the potential to inflate both Type I and Type II error.

Although we cannot solve major problems with the research design using statistics, we can and should examine our data to try and detect problems. In some cases, problems can be solvable. In other cases, our solution will be to qualify the conclusions we make with our data. Data cleaning thus involves understanding issues in the data and correcting them while preserving meaning and statistical conclusion validity.

7.4 A General Plan for Data Cleaning

Refer to this checklist as a general plan for dealing with new data to be used for research. We'll go into more detail about how to do these steps in the sections that follow.

0. Design the study to prevent problems
1. Examine your data
2. Outlier identification and analysis
3. Missing values analysis
4. Test-specific assumption checking
5. Describe your data cleaning in APA style

Next, lets look at each step in more detail.

7.5 Step 0. Design your Research to Minimize Data Problems

I am including this as a reminder, as it is a research design topic and the strategies will differ depending on your research question. As you will see in the sections that follow, we do not have robust ways to correct missing and erroneous data after the fact, so wherever it is possible to increase the reliability and accuracy of data collection and avoid bias, it is worthwhile to do so. If a researcher is manually recording data (i.e., taking notes on a clipboard), having a backup researcher recording data would increase the accuracy of the data. In an online study, adding **input response validation** can help prevent errors, as well. If you ask participants for their age in years, then you may wish to have your survey software prevent any characters except for numbers from being entered in the box for “age.”

Response validation does require some care, as it borders on the ethical principle of **respect for persons**, which includes informed consent and the avoidance of coercion. The APA Ethics Code requires us to tell participants what is involved in the study, get their permission to participate, and not use status, power, authority, or pressure to get individuals to participate (Assocation (2020)). This principle applies equally to the study and to individual responses during the study. Just as you cannot tell a participant to stay and complete a study, you also cannot tell a participant that they must answer an item on a survey. Although it may be tempting to mark all your electronic survey items as “required” (which would prevent participants from continuing unless they provide a response), you want to avoid any implication that participants are not free to choose to participate.

In all, designing backups and cross-checks in your data collection protocol can reduce the amount of missing and erroneous data. Because we follow the ethical principle of respect for persons, we cannot compell our participants to provide data to us, and we will usually have to deal with some missing data regardless.

7.6 Step 1. Examine Your Data

Descriptive statistics can help you make sense of a new dataset. After importing data into R, check your dataframe to see what variables are included using `summary()`. Check the minimum and maximum values. Are they reasonable? You can use the function `head()` to see the first few rows of each variable. As yourself these questions:

- What are the variables?
- What are the levels of measurement of the variables? Are they continuous or discrete?
- Are the data complete? Which variables have missing values?
- Does further coding need to be done?
- Run appropriate descriptive statistics
- Generate appropriate visualizations (histograms or bar charts)
- What are the shapes of the distributions?
- Are the distributions normal? What patterns are evident? Generate Q-Q plots
- Are the values reasonable?
- What is the structure of the data? If there are repeated measurements, are they represented as additional columns (called wide format) or additional rows (called long format)

```

percentages1 <- c(32, 3, 2, 32, 32, 45, 101) # participant percentage scores at time1
percentages2 <- c(12, 12, 12, 11, 18, 92, 14) # participant percentage scores at time2
df <- cbind(percentages1, percentages2) # create a dataframe
summary(df) # list min and max values

##   percentages1      percentages2
##   Min. : 2.00      Min. :11.00
##   1st Qu.:17.50    1st Qu.:12.00
##   Median :32.00    Median :12.00
##   Mean   :35.29    Mean   :24.43
##   3rd Qu.:38.50    3rd Qu.:16.00
##   Max.   :101.00   Max.  :92.00

head(df) # list first few values

```

```
##      percentages1 percentages2
## [1,]         32            12
## [2,]          3            12
## [3,]          2            12
## [4,]         32            11
## [5,]         32            18
## [6,]         45            92
```

In the example above, I have recorded scores for participants at two points in time. The scores are percentages, so acceptable values are between 0 and 100. Percentages are continuous, ratio measures. When I run `summary()`, I see that `percentages1` has a maximum score of 101, so that value is probably an error. I would want to go back to the original source of the data and see if I can find the intended value. I would finish this first pass by running descriptive statistics (note that some are provided by the `summary()` function) and generating histograms.

Finally, you should be aware that there are two ways to include repeated measurements from participants. Any study with a within-subjects design will result in repeated observations of the same participants. There are two ways to collect repeated measures for a participant. One option is to add additional columns for each measurement (as I did in my example, a `time1` variable and a `time2` variable), which is called wide format. The other option is to add additional rows for each measurement, which is called long format. As a side note, SPSS only works with data in wide format. Some procedures in R require wide format while others require long format.

Next, we will add some useful R functions for managing data and converting between wide and long formats.

Note that you may not need to use the following R functions immediately. I am including them because they can be useful during data cleaning.

7.6.1 Sorting, flipping and merging data

Text by Navarro (2018)

In this section I discuss a few useful operations that I feel are loosely related to one another: sorting a vector, sorting a data frame, binding two or more vectors together into a data frame (or matrix), and flipping a data frame (or matrix) on its side. They're all fairly straightforward tasks, at least in comparison to some of the more obnoxious data handling problems that turn up in real life.

7.6.1.1 Creating tables from vectors

Let's start with a simple example. As the father of a small child, I naturally spend a lot of time watching TV shows like *In the Night Garden*. In the `nightgarden.Rdata` file, I've transcribed a short section of the dialogue. The file contains two variables, `speaker` and `utterance`, and when we take a look at the data, it becomes very clear what happened to my sanity.

```
library(lsr)
load(file.path(projecthome, "data", "nightgarden.Rdata"))
who()

##   -- Name --      -- Class --      -- Size --
##   a               numeric          1
##   addArrow        function
##   addDistPlot     function
##   afl              data.frame    4296 x 12
##   afl.finalists   factor         400
##   afl.margins     numeric        176
##   afl2             data.frame    4296 x 2
##   b               numeric          1
##   clin.trial      data.frame    18 x 3
##   colour           logical         1
##   crit.hi          numeric          1
##   crit.lo          numeric          1
##   def.par           list            66
##   df               matrix          7 x 2
##   emphCol          character        1
##   emphColLight    character        1
##   emphGrey         character        1
##   eps              logical          1
##   estImg            list            0
##   Fibonacci        numeric          7
##   freq              integer         17
##   h                numeric          1
##   height            numeric          1
##   i                integer          1
##   IQ                numeric        10000
##   m                numeric          1
##   N                integer         100
##   nhstImg           list            0
##   onesample         numeric          5
##   percentages1     numeric          7
##   percentages2     numeric          7
##   plotOne          function
```

```

##   plotSamples      function
##   pop_sd          numeric     1
##   population      numeric    1000
##   pow              numeric    100
##   projecthome     character   1
##   s                numeric    1
##   sampling         numeric   2000
##   sampling_sd      numeric    1
##   setUpPlot        function
##   speaker          character  10
##   standard_error   numeric    1
##   suspicious.cases logical   176
##   teams             character  17
##   theta             numeric    1
##   utterance         character 10
##   width             numeric    1
##   x                 integer   101
##   X                 numeric   1000
##   y                 numeric   101
##   z                 logical   101

print( speaker )

##  [1] "upsy-daisy"  "upsy-daisy"  "upsy-daisy"  "upsy-daisy"  "tomboliboo"
##  [6] "tomboliboo"  "makka-pakka" "makka-pakka" "makka-pakka" "makka-pakka"

print( utterance )

##  [1] "pip" "pip" "onk" "onk" "ee"   "oo"   "pip" "pip" "onk" "onk"

```

With these as my data, one task I might find myself needing to do is construct a frequency count of the number of words each character speaks during the show. The `table()` function provides a simple way to do this. The basic usage of the `table()` function is as follows:

```

table(speaker)

## speaker
## makka-pakka  tomboliboo upsy-daisy
##           4            2            4

```

The output here tells us on the first line that what we're looking at is a tabulation of the `speaker` variable. On the second line it lists all the different speakers

that exist in the data, and on the third line it tells you how many times that speaker appears in the data. In other words, it's a frequency table¹ Notice that in the command above I didn't name the argument, since `table()` is another function that makes use of unnamed arguments. You just type in a list of the variables that you want R to tabulate, and it tabulates them. For instance, if I type in the name of two variables, what I get as the output is a cross-tabulation:

```
table(speaker, utterance)

##          utterance
## speaker      ee onk oo pip
##   makka-pakka  0   2   0   2
##   tombliboo    1   0   1   0
##   upsy-daisy   0   2   0   2
```

When interpreting this table, remember that these are counts: so the fact that the first row and second column corresponds to a value of 2 indicates that Makka-Pakka (row 1) says “onk” (column 2) twice in this data set. As you'd expect, you can produce three way or higher order cross tabulations just by adding more objects to the list of inputs. However, I won't discuss that in this section.

7.6.1.2 Creating tables from data frames

Most of the time your data are stored in a data frame, not kept as separate variables in the workspace. Let's create one:

```
itng <- data.frame( speaker, utterance )
itng

##          speaker utterance
## 1   upsy-daisy      pip
## 2   upsy-daisy      pip
## 3   upsy-daisy     onk
## 4   upsy-daisy     onk
## 5   tombliboo       ee
## 6   tombliboo       oo
## 7   makka-pakka     pip
## 8   makka-pakka     pip
## 9   makka-pakka     onk
## 10  makka-pakka     onk
```

¹As usual, you can assign this output to a variable. If you type `speaker.freq <- table(speaker)` at the command prompt R will store the table as a variable. If you then type `class(speaker.freq)` you'll see that the output is actually of class `table`. The key thing to note about a table object is that it's basically a matrix (see Section ??).

There's a couple of options under these circumstances. Firstly, if you just want to cross-tabulate all of the variables in the data frame, then it's really easy:

```
table(itng)
```

```
##             utterance
## speaker      ee onk oo pip
## makka-pakka  0   2   0   2
## tombliboo    1   0   1   0
## upsy-daisy   0   2   0   2
```

However, it's often the case that you want to select particular variables from the data frame to tabulate. This is where the `xtabs()` function is useful. In this function, you input a one sided `formula` in order to list all the variables you want to cross-tabulate, and the name of the `data` frame that stores the data:

```
xtabs( formula = ~ speaker + utterance, data = itng )
```

```
##             utterance
## speaker      ee onk oo pip
## makka-pakka  0   2   0   2
## tombliboo    1   0   1   0
## upsy-daisy   0   2   0   2
```

Clearly, this is a totally unnecessary command in the context of the `itng` data frame, but in most situations when you're analysing real data this is actually extremely useful, since your data set will almost certainly contain lots of variables and you'll only want to tabulate a few of them at a time.

7.6.1.3 Converting a table of counts to a table of proportions

The tabulation commands discussed so far all construct a table of raw frequencies: that is, a count of the total number of cases that satisfy certain conditions. However, often you want your data to be organised in terms of proportions rather than counts. This is where the `prop.table()` function comes in handy. It has two arguments:

- `x`. The frequency table that you want to convert.
- `margin`. Which “dimension” do you want to calculate proportions for. By default, R assumes you want the proportion to be expressed as a fraction of all possible events. See examples for details.

To see how this works:

```

itng.table <- table(itng) # create the table, and assign it to a variable
itng.table           # display the table again, as a reminder

##              utterance
## speaker      ee onk oo pip
## makka-pakka 0   2   0   2
## tombliboo   1   0   1   0
## upsy-daisy  0   2   0   2

prop.table( x = itng.table ) # express as proportion:

##              utterance
## speaker      ee onk oo pip
## makka-pakka 0.0 0.2 0.0 0.2
## tombliboo   0.1 0.0 0.1 0.0
## upsy-daisy  0.0 0.2 0.0 0.2

```

Notice that there were 10 observations in our original data set, so all that R has done here is divide all our raw frequencies by 10. That's a sensible default, but more often you actually want to calculate the proportions separately by row (`margin = 1`) or by column (`margin = 2`). Again, this is most clearly seen by looking at examples:

```

prop.table( x = itng.table, margin = 1)

##              utterance
## speaker      ee onk oo pip
## makka-pakka 0.0 0.5 0.0 0.5
## tombliboo   0.5 0.0 0.5 0.0
## upsy-daisy  0.0 0.5 0.0 0.5

```

Notice that each row now sums to 1, but that's not true for each column. What we're looking at here is the proportions of utterances made by each character. In other words, 50% of Makka-Pakka's utterances are "pip", and the other 50% are "onk". Let's contrast this with the following command:

```

prop.table( x = itng.table, margin = 2)

##              utterance
## speaker      ee onk oo pip
## makka-pakka 0.0 0.5 0.0 0.5
## tombliboo   1.0 0.0 1.0 0.0
## upsy-daisy  0.0 0.5 0.0 0.5

```

Now the columns all sum to 1 but the rows don't. In this version, what we're seeing is the proportion of characters associated with each utterance. For instance, whenever the utterance "ee" is made (in this data set), 100% of the time it's a Tombliboo saying it.

7.6.1.4 Sorting a numeric or character vector

One thing that you often want to do is sort a variable. If it's a numeric variable you might want to sort in increasing or decreasing order. If it's a character vector you might want to sort alphabetically, etc. The `sort()` function provides this capability.

```
numbers <- c(2,4,3)
sort( x = numbers )
```

```
## [1] 2 3 4
```

You can ask for R to sort in decreasing order rather than increasing:

```
sort( x = numbers, decreasing = TRUE )
```

```
## [1] 4 3 2
```

And you can ask it to sort text data in alphabetical order:

```
text <- c("aardvark", "zebra", "swing")
sort( text )
```

```
## [1] "aardvark" "swing"     "zebra"
```

That's pretty straightforward. That being said, it's important to note that I'm glossing over something here. When you apply `sort()` to a character vector it doesn't strictly sort into alphabetical order. R actually has a slightly different notion of how characters are ordered (see Section ?? and Table ??), which is more closely related to how computers store text data than to how letters are ordered in the alphabet. However, that's a topic we'll discuss later. For now, the only thing I should note is that the `sort()` function doesn't alter the original variable. Rather, it creates a new, sorted variable as the output. So if I inspect my original `text` variable:

```
text
## [1] "aardvark" "zebra"     "swing"
```

I can see that it has remained unchanged.

7.6.1.5 Sorting a factor

You can also sort factors, but the story here is slightly more subtle because there's two different ways you can sort a factor: alphabetically (by label) or by factor level. The `sort()` function uses the latter. To illustrate, let's look at the two different examples. First, let's create a factor in the usual way:

```
fac <- factor( text )
fac

## [1] aardvark zebra    swing
## Levels: aardvark swing zebra
```

Now let's sort it:

```
sort(fac)

## [1] aardvark swing    zebra
## Levels: aardvark swing zebra
```

This *looks* like it's sorted things into alphabetical order, but that's only because the factor levels themselves happen to be alphabetically ordered. Suppose I deliberately define the factor levels in a non-alphabetical order:

```
fac <- factor( text, levels = c("zebra", "swing", "aardvark") )
fac

## [1] aardvark zebra    swing
## Levels: zebra swing aardvark
```

Now what happens when we try to sort `fac` this time? The answer:

```
sort(fac)

## [1] zebra    swing    aardvark
## Levels: zebra swing aardvark
```

It sorts the data into the numerical order implied by the factor levels, not the alphabetical order implied by the labels attached to those levels. Normally you never notice the distinction, because by default the factor levels are assigned in alphabetical order, but it's important to know the difference:

7.6.1.6 Sorting a data frame

The `sort()` function doesn't work properly with data frames. If you want to sort a data frame the standard advice that you'll find online is to use the `order()` function (not described in this book) to determine what order the rows should be sorted, and then use square brackets to do the shuffling. There's nothing inherently wrong with this advice, I just find it tedious. To that end, the `lsr` package includes a function called `sortFrame()` that you can use to do the sorting. The first argument to the function is named `(x)`, and should correspond to the data frame that you want sorted. After that, all you do is type a list of the names of the variables that you want to use to do the sorting. For instance, if I type this:

```
load(file.path(projecthome, "data", "nightgarden2.Rdata"))
garden
```

```
##           speaker utterance line
## case.1    upsy-daisy      pip   1
## case.2    upsy-daisy      pip   2
## case.3    tombliboo      ee    5
## case.4    makka-pakka    pip   7
## case.5    makka-pakka    onk   9

sortFrame( garden, speaker, line)
```

```
##           speaker utterance line
## case.4    makka-pakka    pip   7
## case.5    makka-pakka    onk   9
## case.3    tombliboo      ee    5
## case.1    upsy-daisy      pip   1
## case.2    upsy-daisy      pip   2
```

what R does is first sort by `speaker` (factor level order). Any ties (i.e., data from the same speaker) are then sorted in order of `line` (increasing numerical order). You can use the minus sign to indicate that numerical variables should be sorted in reverse order:

```
sortFrame( garden, speaker, -line)
```

```
##           speaker utterance line
## case.5    makka-pakka    onk   9
## case.4    makka-pakka    pip   7
## case.3    tombliboo      ee    5
## case.2    upsy-daisy      pip   2
## case.1    upsy-daisy      pip   1
```

As of the current writing, the `sortFrame()` function is under development. I've started introducing functionality to allow you to use the `-` sign to non-numeric variables or to make a distinction between sorting factors alphabetically or by factor level. The idea is that you should be able to type in something like this:

```
sortFrame( garden, -speaker)
```

and have the output correspond to a sort of the `garden` data frame in *reverse* alphabetical order (or reverse factor level order) of `speaker`. As things stand right now, this will actually work, and it will produce sensible output:

```
sortFrame( garden, -speaker)
```

```
##           speaker utterance line
## case.1    upsy-daisy      pip    1
## case.2    upsy-daisy      pip    2
## case.3    tombliboo       ee     5
## case.4    makka-pakka     pip    7
## case.5    makka-pakka     onk    9
```

However, I'm not completely convinced that I've set this up in the ideal fashion, so this may change a little bit in the future.

7.6.2 Binding vectors together

A not-uncommon task that you might find yourself needing to undertake is to combine several vectors. For instance, let's suppose we have the following two numeric vectors:

```
cake.1 <- c(100, 80, 0, 0, 0)
cake.2 <- c(100, 100, 90, 30, 10)
```

The numbers here might represent the amount of each of the two cakes that are left at five different time points. Apparently the first cake is tastier, since that one gets devoured faster. We've already seen one method for combining these vectors: we could use the `data.frame()` function to convert them into a data frame with two variables, like so:

```
cake.df <- data.frame( cake.1, cake.2 )
cake.df
```

```
##   cake.1 cake.2
```

```
## 1     100    100
## 2      80    100
## 3      0     90
## 4      0     30
## 5      0     10
```

Two other methods that I want to briefly refer to are the `rbind()` and `cbind()` functions, which will convert the vectors into a matrix. I'll discuss matrices properly in Section ?? but the details don't matter too much for our current purposes. The `cbind()` function ("column bind") produces a very similar looking output to the data frame example:

```
cake.mat1 <- cbind( cake.1, cake.2 )
cake.mat1
```

```
##      cake.1 cake.2
## [1,]    100    100
## [2,]     80    100
## [3,]      0     90
## [4,]      0     30
## [5,]      0     10
```

but nevertheless it's important to keep in mind that `cake.mat1` is a matrix rather than a data frame, and so has a few differences from the `cake.df` variable. The `rbind()` function ("row bind") produces a somewhat different output: it binds the vectors together row-wise rather than column-wise, so the output now looks like this:

```
cake.mat2 <- rbind( cake.1, cake.2 )
cake.mat2
```

```
##      [,1] [,2] [,3] [,4] [,5]
## cake.1 100   80    0    0    0
## cake.2 100   100   90   30   10
```

You can add names to a matrix by using the `rownames()` and `colnames()` functions, and I should also point out that there's a fancier function in R called `merge()` that supports more complicated "database like" merging of vectors and data frames, but I won't go into details here.

7.6.2.1 Binding multiple copies of the same vector together

It is sometimes very useful to bind together multiple copies of the same vector. You could do this using the `rbind` and `cbind` functions, using commands like this one

```

fibonacci <- c( 1,1,2,3,5,8 )
rbind( fibonacci, fibonacci, fibonacci )

##      [,1] [,2] [,3] [,4] [,5] [,6]
## fibonacci  1    1    2    3    5    8
## fibonacci  1    1    2    3    5    8
## fibonacci  1    1    2    3    5    8

```

but that can be pretty annoying, especially if you need lots of copies. To make this a little easier, the `lsr` package has two additional functions `rowCopy` and `colCopy` that do the same job, but all you have to do is specify the number of copies that you want, instead of typing the name in over and over again. The two arguments you need to specify are `x`, the vector to be copied, and `times`, indicating how many copies should be created:²

```

rowCopy( x = fibonacci, times = 3 )

##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    1    2    3    5    8
## [2,]    1    1    2    3    5    8
## [3,]    1    1    2    3    5    8

```

Of course, in practice you don't need to name the arguments all the time. For instance, here's an example using the `colCopy()` function with the argument names omitted:

```

colCopy( fibonacci, 3 )

##      [,1] [,2] [,3]
## [1,]    1    1    1
## [2,]    1    1    1
## [3,]    2    2    2
## [4,]    3    3    3
## [5,]    5    5    5
## [6,]    8    8    8

```

²Note for advanced users: both of these functions are just wrappers to the `matrix()` function, which is pretty flexible in terms of the ability to convert vectors into matrices. Also, while I'm on this topic, I'll briefly mention the fact that if you're a Matlab user and looking for an equivalent of Matlab's `repmat()` function, I'd suggest checking out the `matlab` package which contains R versions of a lot of handy Matlab functions.

7.6.2.2 Transposing a matrix or data frame

One of the main reasons that I wanted to discuss the `rbind()` and `cbind()` functions in the same section as the `data.frame()` function is that it immediately raises the question of how to “flip” or *transpose* a matrix or data frame. Notice that in the last section I was able to produce two different matrices, `cake.mat1` and `cake.mat2` that were basically mirror images of one another. A natural question to ask is whether you can directly transform one into another. The transpose function `t()` allows us to do this in a straightforward fashion. To start with, I’ll show you how to transpose a matrix, and then I’ll move onto talk about data frames. Firstly, let’s load a matrix I prepared earlier, from the `cakes.Rdata` file:

```
load(file.path(projecthome, "data", "cakes.Rdata"))
cakes

##          time.1 time.2 time.3 time.4 time.5
## cake.1     100     80      0      0      0
## cake.2     100    100     90     30     10
## cake.3     100     20     20     20     20
## cake.4     100    100    100    100    100
```

And just to make sure you believe me that this is actually a matrix:

```
class( cakes )

## [1] "matrix" "array"
```

Okay, now let’s transpose the matrix:

```
cakes.flipped <- t( cakes )
cakes.flipped
```

```
##          cake.1 cake.2 cake.3 cake.4
## time.1     100     100     100     100
## time.2      80     100      20     100
## time.3       0      90      20     100
## time.4       0      30      20     100
## time.5       0      10      20     100
```

The output here is still a matrix:

```
class( cakes.flipped )

## [1] "matrix" "array"
```

At this point you should have two questions: (1) how do we do the same thing for data frames? and (2) why should we care about this? Let's start with the how question. First, I should note that you can transpose a data frame just fine using the `t()` function, but that has the slightly awkward consequence of converting the output from a data frame to a matrix, which isn't usually what you want. It's quite easy to convert the output back again, of course,³ but I hate typing two commands when I can do it with one. To that end, the `lsr` package has a simple “convenience” function called `tFrame()` which does exactly the same thing as `t()` but converts the output to a data frame for you. To illustrate this, let's transpose the `itng` data frame that we used earlier. Here's the original data frame:

```
itng
```

```
##           speaker utterance
## 1    upsy-daisy      pip
## 2    upsy-daisy      pip
## 3    upsy-daisy     onk
## 4    upsy-daisy     onk
## 5  tombliboo       ee
## 6  tombliboo       oo
## 7 makka-pakka      pip
## 8 makka-pakka      pip
## 9 makka-pakka     onk
## 10 makka-pakka    onk
```

and here's what happens when you transpose it using `tFrame()`:

```
tFrame( itng )
```

```
##          V1        V2        V3        V4        V5        V6
## speaker  upsy-daisy upsy-daisy upsy-daisy upsy-daisy tombliboo tombliboo
## utterance      pip      pip     onk     onk       ee       oo
##          V7        V8        V9        V10
## speaker  makka-pakka makka-pakka makka-pakka makka-pakka
## utterance      pip      pip     onk     onk
```

³The function you need for that is called `as.data.frame()`.

An important point to recognise is that transposing a data frame is not always a sensible thing to do: in fact, I'd go so far as to argue that it's usually *not* sensible. It depends a lot on whether the “cases” from your original data frame would make sense as variables, and to think of each of your original “variables” as cases. I think that's emphatically *not* true for our `itng` data frame, so I wouldn't advise doing it in this situation.

That being said, sometimes it really is true. For instance, had we originally stored our `cakes` variable as a data frame instead of a matrix, then it would absolutely be sensible to flip the data frame!⁴ There are some situations where it is useful to flip your data frame, so it's nice to know that you can do it. Indeed, that's the main reason why I have spent so much time talking about this topic. A lot of statistical tools make the assumption that the rows of your data frame (or matrix) correspond to observations, and the columns correspond to the variables. That's not unreasonable, of course, since that is a pretty standard convention. However, think about our `cakes` example here. This is a situation where you might want to do an analysis of the different cakes (i.e. cakes as variables, time points as cases), but equally you might want to do an analysis where you think of the times as being the things of interest (i.e., times as variables, cakes as cases). If so, then it's useful to know how to flip a matrix or data frame around.

7.6.3 Reshaping a data frame

Text by Navarro (2018)

One of the most annoying tasks that you need to undertake on a regular basis is that of reshaping a data frame. Framed in the most general way, reshaping the data means taking the data in whatever format it's given to you, and converting it to the format you need it. Of course, if we're going to characterise the problem that broadly, then about half of this chapter can probably be thought of as a kind of reshaping. So we're going to have to narrow things down a little bit. To that end, I'll talk about a few different tools that you can use for a few different tasks. In particular, I'll discuss a couple of easy to use (but limited) functions that I've included in the `lsr` package. In future versions of the book I plan to expand this discussion to include some of the more powerful tools that are available in R, but I haven't had the time to do so yet.

⁴In truth, I suspect that most of the cases when you can sensibly flip a data frame occur when all of the original variables are measurements of the same type (e.g., all variables are response times), and if so you could easily have chosen to encode your data as a matrix instead of as a data frame. But since people do sometimes prefer to work with data frames, I've written the `tFrame()` function for the sake of convenience. I don't really think it's something that is needed very often.

7.6.3.1 Long form and wide form data

The most common format in which you might obtain data is as a “case by variable” layout, commonly known as the *wide form* of the data.

```
load(file.path(projecthome, "data", "repeated.Rdata"))
who()
```

```
##   -- Name --      -- Class --      -- Size --
##   a                  numeric       1
##   addArrow          function
##   addDistPlot        function
##   afl                 data.frame  4296 x 12
##   afl.finalists      factor        400
##   afl.margins         numeric      176
##   afl2                data.frame  4296 x  2
##   b                  numeric       1
##   cake.1              numeric      5
##   cake.2              numeric      5
##   cake.df             data.frame  5 x  2
##   cake.mat1           matrix      5 x  2
##   cake.mat2           matrix      2 x  5
##   cakes               matrix      4 x  5
##   cakes.flipped       matrix      5 x  4
##   choice              data.frame  4 x 10
##   clin.trial          data.frame 18 x  3
##   colour              logical      1
##   crit.hi              numeric      1
##   crit.lo              numeric      1
##   def.par              list        66
##   df                  matrix      7 x  2
##   drugs               data.frame 10 x  8
##   emphCol              character    1
##   emphColLight         character    1
##   emphGrey             character    1
##   eps                 logical      1
##   estImg              list        0
##   fac                 factor       3
##   fibonacci            numeric      6
##   Fibonacci            numeric      7
##   freq                integer      17
##   garden              data.frame  5 x  3
##   h                   numeric      1
##   height              numeric      1
##   i                   integer      1
```

```

##   IQ      numeric    10000
##   itng    data.frame 10 x 2
##   itng.table  table    3 x 4
##   m      numeric    1
##   N      integer    100
##   nhstImg list      0
##   numbers numeric    3
##   onesample numeric    5
##   percentages1 numeric    7
##   percentages2 numeric    7
##   plotOne function
##   plotSamples function
##   pop_sd   numeric    1
##   population numeric    1000
##   pow      numeric    100
##   projecthome character 1
##   s       numeric    1
##   sampling numeric    2000
##   sampling_sd numeric    1
##   setUpPlot function
##   speaker  character 10
##   standard_error numeric    1
##   suspicious.cases logical   176
##   teams    character 17
##   text     character 3
##   theta    numeric    1
##   utterance character 10
##   width    numeric    1
##   x        integer    101
##   X        numeric    1000
##   y        numeric    101
##   z        logical    101

```

To get a sense of what I'm talking about, consider an experiment in which we are interested in the different effects that alcohol and caffeine have on people's working memory capacity (WMC) and reaction times (RT). We recruit 10 participants, and measure their WMC and RT under three different conditions: a "no drug" condition, in which they are not under the influence of either caffeine or alcohol, a "caffeine" condition, in which they are under the influence of caffeine, and an "alcohol" condition, in which... well, you can probably guess. Ideally, I suppose, there would be a fourth condition in which both drugs are administered, but for the sake of simplicity let's ignore that. The `drugs` data frame gives you a sense of what kind of data you might observe in an experiment like this:

```
drugs
```

```
##   id gender WMC_alcohol WMC_caffeine WMC_no.drug RT_alcohol RT_caffeine
## 1  1 female      3.7       3.7      3.9      488      236
## 2  2 female      6.4       7.3      7.9      607      376
## 3  3 female      4.6       7.4      7.3      643      226
## 4  4 male        6.4       7.8      8.2      684      206
## 5  5 female      4.9       5.2      7.0      593      262
## 6  6 male        5.4       6.6      7.2      492      230
## 7  7 male        7.9       7.9      8.9      690      259
## 8  8 male        4.1       5.9      4.5      486      230
## 9  9 female      5.2       6.2      7.2      686      273
## 10 10 female     6.2       7.4      7.8      645      240
##   RT_no.drug
## 1      371
## 2      349
## 3      412
## 4      252
## 5      439
## 6      464
## 7      327
## 8      305
## 9      327
## 10     498
```

This is a data set in “wide form”, in which each participant corresponds to a single row. We have two variables that are characteristics of the subject (i.e., their `id` number and their `gender`) and six variables that refer to one of the two measured variables (WMC or RT) in one of the three testing conditions (alcohol, caffeine or no drug). Because all of the testing conditions (i.e., the three drug types) are applied to all participants, drug type is an example of a *within-subject factor*.

7.6.3.2 Reshaping data using `wideToLong()`

The “wide form” of this data set is useful for some situations: it is often very useful to have each row correspond to a single subject. However, it is not the only way in which you might want to organise this data. For instance, you might want to have a separate row for each “testing occasion”. That is, “participant 1 under the influence of alcohol” would be one row, and “participant 1 under the influence of caffeine” would be another row. This way of organising the data is generally referred to as the *long form* of the data. It’s not too difficult to switch between wide and long form, and I’ll explain how it works in a moment; for now, let’s just have a look at what the long form of this data set looks like:

```
drugs.2 <- wideToLong( data = drugs, within = "drug" )
head(drugs.2)
```

```
##   id gender   drug WMC  RT
## 1  1 female alcohol 3.7 488
## 2  2 female alcohol 6.4 607
## 3  3 female alcohol 4.6 643
## 4  4   male alcohol 6.4 684
## 5  5 female alcohol 4.9 593
## 6  6   male alcohol 5.4 492
```

The `drugs.2` data frame that we just created has 30 rows: each of the 10 participants appears in three separate rows, one corresponding to each of the three testing conditions. And instead of having a variable like `WMC_caffeine` that indicates that we were measuring “WMC” in the “caffeine” condition, this information is now recorded in two separate variables, one called `drug` and another called `WMC`. Obviously, the long and wide forms of the data contain the same information, but they represent quite different ways of organising that information. Sometimes you find yourself needing to analyse data in wide form, and sometimes you find that you need long form. So it’s really useful to know how to switch between the two.

In the example I gave above, I used a function called `wideToLong()` to do the transformation. The `wideToLong()` function is part of the `lsr` package. The key to understanding this function is that it relies on the *variable names* to do all the work. Notice that the variable names in the `drugs` data frame follow a very clear scheme. Whenever you have a variable with a name like `WMC_caffeine` you know that the variable being measured is “WMC”, and that the specific condition in which it is being measured is the “caffeine” condition. Similarly, you know that `RT_no.drug` refers to the “RT” variable measured in the “no drug” condition. The measured variable comes first (e.g., `WMC`), followed by a separator character (in this case the separator is an underscore, `_`), and then the name of the condition in which it is being measured (e.g., `caffeine`). There are two different prefixes (i.e, the strings before the separator, `WMC`, `RT`) which means that there are two separate variables being measured. There are three different suffixes (i.e., the strings after the separator, `caffeine`, `alcohol`, `no.drug`) meaning that there are three different levels of the within-subject factor. Finally, notice that the separator string (i.e., `_`) does not appear anywhere in two of the variables (`id`, `gender`), indicating that these are *between-subject* variables, namely variables that do not vary within participant (e.g., a person’s `gender` is the same regardless of whether they’re under the influence of alcohol, caffeine etc.).

Because of the fact that the variable naming scheme here is so informative, it’s quite possible to reshape the data frame without any additional input from the user. For example, in this particular case, you could just type the following:

```
wideToLong( drugs )
```

```
##   id gender   within WMC  RT
## 1  1 female alcohol 3.7 488
## 2  2 female alcohol 6.4 607
## 3  3 female alcohol 4.6 643
## 4  4   male alcohol 6.4 684
## 5  5 female alcohol 4.9 593
## 6  6   male alcohol 5.4 492
## 7  7   male alcohol 7.9 690
## 8  8   male alcohol 4.1 486
## 9  9 female alcohol 5.2 686
## 10 10 female alcohol 6.2 645
## 11 11 female caffeine 3.7 236
## 12 12 female caffeine 7.3 376
## 13 13 female caffeine 7.4 226
## 14 14   male caffeine 7.8 206
## 15 15 female caffeine 5.2 262
## 16 16   male caffeine 6.6 230
## 17 17   male caffeine 7.9 259
## 18 18   male caffeine 5.9 230
## 19 19 female caffeine 6.2 273
## 20 20 female caffeine 7.4 240
## 21 21  1 female no.drug 3.9 371
## 22 22  2 female no.drug 7.9 349
## 23 23  3 female no.drug 7.3 412
## 24 24  4   male no.drug 8.2 252
## 25 25  5 female no.drug 7.0 439
## 26 26  6   male no.drug 7.2 464
## 27 27  7   male no.drug 8.9 327
## 28 28  8   male no.drug 4.5 305
## 29 29  9 female no.drug 7.2 327
## 30 30 10 female no.drug 7.8 498
```

This is pretty good, actually. The only think it has gotten wrong here is that it doesn't know what name to assign to the within-subject factor, so instead of calling it something sensible like `drug`, it has used the unimaginative name `within`. If you want to ensure that the `wideToLong()` function applies a sensible name, you have to specify the `within` argument, which is just a character string that specifies the name of the within-subject factor. So when I used this command earlier,

```
drugs.2 <- wideToLong( data = drugs, within = "drug" )
```

all I was doing was telling R to use `drug` as the name of the within subject factor.

Now, as I was hinting earlier, the `wideToLong()` function is very inflexible. It *requires* that the variable names all follow this naming scheme that I outlined earlier. If you don't follow this naming scheme it won't work.⁵ The only flexibility that I've included here is that you can change the separator character by specifying the `sep` argument. For instance, if you were using variable names of the form `WMC/caffeine`, for instance, you could specify that `sep="/"`, using a command like this

```
drugs.2 <- wideToLong( data = drugs, within = "drug", sep = "/" )
```

and it would still work.

7.6.3.3 Reshaping data using `longToWide()`

To convert data from long form to wide form, the `lsr` package also includes a function called `longToWide()`. Recall from earlier that the long form of the data (i.e., the `drugs.2` data frame) contains variables named `id`, `gender`, `drug`, `WMC` and `RT`. In order to convert from long form to wide form, all you need to do is indicate which of these variables are measured separately for each condition (i.e., `WMC` and `RT`), and which variable is the within-subject factor that specifies the condition (i.e., `drug`). You do this via a two-sided formula, in which the measured variables are on the left hand side, and the within-subject factor is on the right hand side. In this case, the formula would be `WMC + RT ~ drug`. So the command that we would use might look like this:

```
longToWide( data=drugs.2, formula= WMC+RT ~ drug )
```

	<code>id</code>	<code>gender</code>	<code>WMC_alcohol</code>	<code>RT_alcohol</code>	<code>WMC_caffeine</code>	<code>RT_caffeine</code>	<code>WMC_no.drug</code>
## 1	1	female	3.7	488	3.7	236	3.9
## 2	2	female	6.4	607	7.3	376	7.9
## 3	3	female	4.6	643	7.4	226	7.3
## 4	4	male	6.4	684	7.8	206	8.2
## 5	5	female	4.9	593	5.2	262	7.0
## 6	6	male	5.4	492	6.6	230	7.2
## 7	7	male	7.9	690	7.9	259	8.9
## 8	8	male	4.1	486	5.9	230	4.5
## 9	9	female	5.2	686	6.2	273	7.2
## 10	10	female	6.2	645	7.4	240	7.8

⁵This limitation is deliberate, by the way: if you're getting to the point where you want to do something more complicated, you should probably start learning how to use `reshape()`, `cast()` and `melt()` or some of other the more advanced tools. The `wideToLong()` and `longToWide()` functions are included only to help you out when you're first starting to use R.

```
##      RT_no.drug
## 1        371
## 2        349
## 3        412
## 4        252
## 5        439
## 6        464
## 7        327
## 8        305
## 9        327
## 10       498
```

or, if we chose to omit argument names, we could simplify it to this:

```
longToWide( drugs.2, WMC+RT ~ drug )
```

```
##      id gender WMC_alcohol RT_alcohol WMC_caffeine RT_caffeine WMC_no.drug
## 1    1 female     3.7      488      3.7      236      3.9
## 2    2 female     6.4      607      7.3      376      7.9
## 3    3 female     4.6      643      7.4      226      7.3
## 4    4 male       6.4      684      7.8      206      8.2
## 5    5 female     4.9      593      5.2      262      7.0
## 6    6 male       5.4      492      6.6      230      7.2
## 7    7 male       7.9      690      7.9      259      8.9
## 8    8 male       4.1      486      5.9      230      4.5
## 9    9 female     5.2      686      6.2      273      7.2
## 10   10 female    6.2      645      7.4      240      7.8
##      RT_no.drug
## 1        371
## 2        349
## 3        412
## 4        252
## 5        439
## 6        464
## 7        327
## 8        305
## 9        327
## 10       498
```

Note that, just like the `wideToLong()` function, the `longToWide()` function allows you to override the default separator character. For instance, if the command I used had been

```
longToWide( drugs.2, WMC+RT ~ drug, sep="/" )

##      id gender WMC/alcohol RT/alcohol WMC/caffeine RT/caffeine WMC/no.drug
## 1    1 female     3.7       488      3.7       236      3.9
## 2    2 female     6.4       607      7.3       376      7.9
## 3    3 female     4.6       643      7.4       226      7.3
## 4    4 male       6.4       684      7.8       206      8.2
## 5    5 female     4.9       593      5.2       262      7.0
## 6    6 male       5.4       492      6.6       230      7.2
## 7    7 male       7.9       690      7.9       259      8.9
## 8    8 male       4.1       486      5.9       230      4.5
## 9    9 female     5.2       686      6.2       273      7.2
## 10  10 female    6.2       645      7.4       240      7.8
##      RT/no.drug
## 1    371
## 2    349
## 3    412
## 4    252
## 5    439
## 6    464
## 7    327
## 8    305
## 9    327
## 10   498
```

the output would contain variables with names like `RT/alcohol` instead of `RT_alcohol`.

7.6.3.4 Reshaping with multiple within-subject factors

As I mentioned above, the `wideToLong()` and `longToWide()` functions are quite limited in terms of what they can do. However, they do handle a broader range of situations than the one outlined above. Consider the following, fairly simple psychological experiment. I'm interested in the effects of practice on some simple decision making problem. It doesn't really matter what the problem is, other than to note that I'm interested in two distinct outcome variables. Firstly, I care about people's accuracy, measured by the proportion of decisions that people make correctly, denoted `PC`. Secondly, I care about people's speed, measured by the mean response time taken to make those decisions, denoted `MRT`. That's standard in psychological experiments: the speed-accuracy trade-off is pretty ubiquitous, so we generally need to care about both variables.

To look at the effects of practice over the long term, I test each participant on two days, `day1` and `day2`, where for the sake of argument I'll assume that `day1`

and `day2` are about a week apart. To look at the effects of practice over the short term, the testing during each day is broken into two “blocks”, `block1` and `block2`, which are about 20 minutes apart. This isn’t the world’s most complicated experiment, but it’s still a fair bit more complicated than the last one. This time around we have two within-subject factors (i.e., `day` and `block`) and we have two measured variables for each condition (i.e., `PC` and `MRT`). The `choice` data frame shows what the wide form of this kind of data might look like:

```
choice

##   id gender MRT/block1/day1 MRT/block1/day2 MRT/block2/day1
## 1  1   male        415        400        455
## 2  2   male        500        490        532
## 3  3 female       478        468        499
## 4  4 female       550        502        602
##   MRT/block2/day2 PC/block1/day1 PC/block1/day2 PC/block2/day1
## 1            450         79         88         82
## 2            518         83         92         86
## 3            474         91         98         90
## 4            588         75         89         78
##   PC/block2/day2
## 1            93
## 2            97
## 3           100
## 4            95
```

Notice that this time around we have variable names of the form `MRT/block1/day2`. As before, the first part of the name refers to the measured variable (response time), but there are now two suffixes, one indicating that the testing took place in block 1, and the other indicating that it took place on day 2. And just to complicate matters, it uses `/` as the separator character rather than `_`. Even so, reshaping this data set is pretty easy. The command to do it is,

```
choice.2 <- wideToLong( choice, within=c("block","day"), sep="/" )
```

which is pretty much the exact same command we used last time. The only difference here is that, because there are two within-subject factors, the `within` argument is a vector that contains two names. When we look at the long form data frame that this creates, we get this:

```
choice.2
```

```

##   id gender MRT  PC  block  day
## 1   1 male  415 79 block1 day1
## 2   2 male  500 83 block1 day1
## 3   3 female 478 91 block1 day1
## 4   4 female 550 75 block1 day1
## 5   1 male  400 88 block1 day2
## 6   2 male  490 92 block1 day2
## 7   3 female 468 98 block1 day2
## 8   4 female 502 89 block1 day2
## 9   1 male  455 82 block2 day1
## 10  2 male  532 86 block2 day1
## 11  3 female 499 90 block2 day1
## 12  4 female 602 78 block2 day1
## 13  1 male  450 93 block2 day2
## 14  2 male  518 97 block2 day2
## 15  3 female 474 100 block2 day2
## 16  4 female 588 95 block2 day2

```

In this long form data frame we have two between-subject variables (`id` and `gender`), two variables that define our within-subject manipulations (`block` and `day`), and two more contain the measurements we took (`MRT` and `PC`).

To convert this back to wide form is equally straightforward. We use the `longToWide()` function, but this time around we need to alter the formula in order to tell it that we have two within-subject factors. The command is now

```

longToWide( choice.2, MRT+PC ~ block+day, sep="/" )

##   id gender MRT/block1/day1 PC/block1/day1 MRT/block1/day2 PC/block1/day2
## 1   1 male      415          79      400          88
## 2   2 male      500          83      490          92
## 3   3 female    478          91      468          98
## 4   4 female    550          75      502          89
##   MRT/block2/day1 PC/block2/day1 MRT/block2/day2 PC/block2/day2
## 1           455          82      450          93
## 2           532          86      518          97
## 3           499          90      474         100
## 4           602          78      588          95

```

and this produces a wide form data set containing the same variables as the original `choice` data frame.

7.6.3.5 What other options are there?

The advantage to the approach described in the previous section is that it solves a quite specific problem (but a commonly encountered one) with a minimum of

fuss. The disadvantage is that the tools are quite limited in scope. They allow you to switch your data back and forth between two different formats that are very common in everyday data analysis. However, there are a number of other tools that you can use if need be. Just within the core packages distributed with R there is the `reshape()` function, as well as the `stack()` and `unstack()` functions, all of which can be useful under certain circumstances. And there are of course thousands of packages on CRAN that you can use to help you with different tasks. One popular package for this purpose is the `reshape` package, written by Hadley Wickham ?, for details see Wickham2007. There are two key functions in this package, called `melt()` and `cast()` that are pretty useful for solving a lot of reshaping problems. In a future version of this book I intend to discuss `melt()` and `cast()` in a fair amount of detail.

7.7 Step 2. Outlier Analysis

An **Outlier** is a low-frequency extreme score. Remember that the mean is the balance point of a distribution, so a single extreme score can cause the mean to shift dramatically.

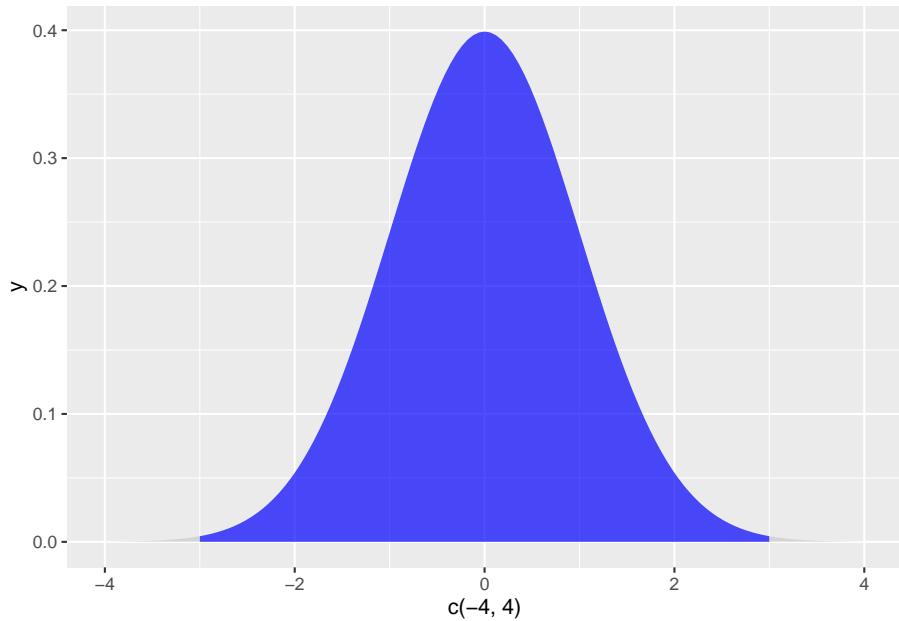
Outlier analysis is detecting which variables have outliers, if any, and reasoning about why outliers exist. Often, an extreme value is a typo. Someone meant to enter 10 and they entered 100. Where you can identify such an error and are confident of its true value, then you can correct it. This is an error, not an outlier.

Here are some helpful questions to ask yourself as you look for outliers:

- How do you define an outlier? A commonly accepted definition for outliers is beyond three standard deviations from the mean in either direction.
- Which variables have outliers? How many outliers are present?
- How many outliers would you expect to find in sample of this size? To solve this, find the total area under the curve beyond three standard deviations:

```
area_above_z3 <- pnorm(3, mean = 0, sd = 1, lower.tail = FALSE) # find area under curve above z = 3
area_below_z3 <- area_above_z3 # there will be the same sized area below z = -3
p_z3_outlier <- area_above_z3 + area_below_z3 # add the two tails together to get probability of being an outlier

library(ggplot2)
ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function", fun = dnorm, xlim = c(-4, -3), alpha = 0.5, fill=alpha("grey",0)) +
  geom_area(stat = "function", fun = dnorm, fill="blue", xlim = c(-3, 3), alpha = 0.7) +
  geom_area(stat = "function", fun = dnorm, xlim = c(3, 4), alpha = 0.5, fill=alpha("grey",0))
```



This gives a probability around .001 of obtaining an outlier beyond $z = 3$. If you multiply this by your sample size, you have a rough estimate of the number of extreme scores you would expect if you were using random sampling. In a sample of $N = 10000$, you should expect to have some extreme scores and nothing is “wrong,” as this is how random sampling is supposed to work. In a sample of $N = 100$, the presence of an outlier is more concerning, and additional effort is needed to explain it.

- Why do you think the the outlier scores are so extreme? If one individual had exceptionally high or low performance, you may be tempted to remove them. However, be careful, as these exclusions can lead to bias. Ask yourself if the outlier is a member of your population, and thus should be included in your sample. This question is key for deciding what, if any, action to take. For example, if you are studying the accuracy of college students’ answers to Jeopardy! questions, one participant getting them all correct may be an outlier. If that participant was an exceptional college student, they should remain in the data to avoid bias to the sample. If, however, that participant was a former Jeopardy! champion, the researchers may decide that this is a study of the amateur public, not actual contests, and exclude that participant on the basis of them not being a true member of the population of interest. Note that regardless of the path, researchers must explain all data exclusions in their results section. Also notice that dealing with outlier scores requires more justification than simply excluding scores because they are unusually high or low.
- Do you have any reason to suspect bivariate outliers? Finally, participants

may be outliers because of their scores on two measures (Cohen (2013)). Cohen (2013) gives the example of a man 74 inches tall who weighs 140 pounds. Although neither score is extreme, the combination of these scores occurs with a low frequency.

7.8 Step 3. Missing values analysis

Text by David Schuster

Do not be daunted by missing values analysis. It boils down to three issues:

- R will complain if we try to do data analysis with variables with missing data. We need to know how R handles missing data.
- What proportion of the sample is missing (computed as the number missing divided by the sample size)? The lower the proportion that are missing, the lower the impact.
- Why are the data missing? We will attach one of three labels (MCAR, MAR, MNAR). The reasons matter because MAR and MNAR will lead to biased estimates.

7.8.1 Special values in R

Text by Navarro (2018)

The first thing I want to mention are some of the “special” values that you might see R produce. Most likely you’ll see them in situations where you were expecting a number, but there are quite a few other ways you can encounter them. These values are `Inf`, `NaN`, `NA` and `NULL`. These values can crop up in various different places, and so it’s important to understand what they mean.

- *Infinity (Inf)*. The easiest of the special values to explain is `Inf`, since it corresponds to a value that is infinitely large. You can also have `-Inf`. The easiest way to get `Inf` is to divide a positive number by 0:

```
1 / 0
```

```
## [1] Inf
```

In most real world data analysis situations, if you’re ending up with infinite numbers in your data, then something has gone awry. Hopefully you’ll never have to see them.

- *Not a Number (NaN)*. The special value of `NaN` is short for “not a number”, and it’s basically a reserved keyword that means “there isn’t a mathematically defined number for this”. If you can remember your high school maths, remember that it is conventional to say that $0/0$ doesn’t have a proper answer: mathematicians would say that $0/0$ is *undefined*. R says that it’s not a number:

```
0 / 0
```

```
## [1] NaN
```

Nevertheless, it’s still treated as a “numeric” value. To oversimplify, `NaN` corresponds to cases where you asked a proper numerical question that genuinely has *no meaningful answer*.

- *Not available (NA)* - used for missing data. `NA` indicates that the value that is “supposed” to be stored here is missing. To understand what this means, it helps to recognise that the `NA` value is something that you’re most likely to see when analysing data from real world experiments. Sometimes you get equipment failures, or you lose some of the data, or whatever. The point is that some of the information that you were “expecting” to get from your study is just plain missing. Note the difference between `NA` and `NaN`. For `NaN`, we really do know what’s supposed to be stored; it’s just that it happens to correspond to something like $0/0$ that doesn’t make any sense at all. In contrast, `NA` indicates that we actually don’t know what was supposed to be there. The information is *missing*.
- *No value (NULL)*. The `NULL` value takes this “absence” concept even further. It basically asserts that the variable genuinely has no value whatsoever. This is quite different to both `NaN` and `NA`. For `NaN` we actually know what the value is, because it’s something insane like $0/0$. For `NA`, we believe that there is supposed to be a value “out there”, but a dog ate our homework and so we don’t quite know what it is. But for `NULL` we strongly believe that there is *no value at all*.

7.8.2 Handling missing values in R

Text by Navarro (2018)

Real data sets very frequently turn out to have missing values: perhaps someone forgot to fill in a particular survey question, for instance. Missing data can be the source of a lot of tricky issues, most of which I’m going to gloss over. However, at a minimum, you need to understand the basics of handling missing data in R.

7.8.2.1 The single variable case

Let's start with the simplest case, in which you're trying to calculate descriptive statistics for a single variable which has missing data. In R, this means that there will be `NA` values in your data vector. Let's create a variable like that:

```
> partial <- c(10, 20, NA, 30)
```

Let's assume that you want to calculate the mean of this variable. By default, R assumes that you want to calculate the mean using all four elements of this vector, which is probably the safest thing for a dumb automaton to do, but it's rarely what you actually want. Why not? Well, remember that the basic interpretation of `NA` is "I don't know what this number is". This means that $1 + \text{NA} = \text{NA}$: if I add 1 to some number that I don't know (i.e., the `NA`) then the answer is *also* a number that I don't know. As a consequence, if you don't explicitly tell R to ignore the `NA` values, and the data set does have missing values, then the output will itself be a missing value. If I try to calculate the mean of the `partial` vector, without doing anything about the missing value, here's what happens:

```
> mean( x = partial )
[1] NA
```

Technically correct, but deeply unhelpful.

To fix this, all of the descriptive statistics functions that I've discussed in this chapter (with the exception of `cor()` which is a special case I'll discuss below) have an optional argument called `na.rm`, which is shorthand for "remove NA values". By default, `na.rm = FALSE`, so R does nothing about the missing data problem. Let's try setting `na.rm = TRUE` and see what happens:

When calculating sums and means when missing data are present (i.e., when there are `NA` values) there's actually an additional argument to the function that you should be aware of. This argument is called `na.rm`, and is a logical value indicating whether R should ignore (or "remove") the missing data for the purposes of doing the calculations. By default, R assumes that you want to keep the missing values, so unless you say otherwise it will set `na.rm = FALSE`. However, R assumes that $1 + \text{NA} = \text{NA}$: if I add 1 to some number that I don't know (i.e., the `NA`) then the answer is *also* a number that I don't know. As a consequence, if you don't explicitly tell R to ignore the `NA` values, and the data set does have missing values, then the output will itself be a missing value. This is illustrated in the following extract:

```
> mean( x = partial, na.rm = TRUE )
[1] 20
```

Notice that the mean is 20 (i.e., $60 / 3$) and *not* 15. When R ignores a NA value, it genuinely ignores it. In effect, the calculation above is identical to what you'd get if you asked for the mean of the three-element vector `c(10, 20, 30)`.

As indicated above, this isn't unique to the `mean()` function. Pretty much all of the other functions that I've talked about in this chapter have an `na.rm` argument that indicates whether it should ignore missing values. However, its behaviour is the same for all these functions, so I won't waste everyone's time by demonstrating it separately for each one.

7.8.2.2 Missing values in pairwise calculations

I mentioned earlier that the `cor()` function is a special case. It doesn't have an `na.rm` argument, because the story becomes a lot more complicated when more than one variable is involved. What it does have is an argument called `use` which does roughly the same thing, but you need to think little more carefully about what you want this time. To illustrate the issues, let's open up a data set that has missing values, `parenthood2.Rdata`. This file contains the same data as the original `parenthood` data, but with some values deleted. It contains a single data frame, `parenthood2`:

```
> load("parenthood2.Rdata")
> print(parenthood2)
  dan.sleep baby.sleep dan.grump day
1      7.59        NA       56   1
2      7.91     11.66      60   2
3      5.14      7.92      82   3
4      7.71      9.61      55   4
5      6.68      9.75      NA   5
6      5.99      5.04      72   6
BLAH BLAH BLAH
```

If I calculate my descriptive statistics using the `describe()` function

```
> describe(parenthood2)
      var   n   mean      sd median trimmed    mad    min     max   BLAH
dan.sleep   1  91  6.98  1.02    7.03    7.02  1.13  4.84  9.00  BLAH
baby.sleep  2  89  8.11  2.05    8.20    8.13  2.28  3.25 12.07  BLAH
dan.grump   3  92 63.15  9.85   61.00   62.66 10.38 41.00 89.00  BLAH
day         4 100 50.50 29.01   50.50   50.50 37.06  1.00 100.00  BLAH
```

we can see from the `n` column that there are 9 missing values for `dan.sleep`, 11

missing values for `baby.sleep` and 8 missing values for `dan.grump`.⁶ Suppose what I would like is a correlation matrix. And let's also suppose that I don't bother to tell R how to handle those missing values. Here's what happens:

```
> cor( parenthood2 )
    dan.sleep baby.sleep dan.grump day
dan.sleep      1          NA        NA  NA
baby.sleep     NA         1          NA  NA
dan.grump      NA         NA         1  NA
day            NA         NA        NA   1
```

Annoying, but it kind of makes sense. If I don't *know* what some of the values of `dan.sleep` and `baby.sleep` actually are, then I can't possibly *know* what the correlation between these two variables is either, since the formula for the correlation coefficient makes use of every single observation in the data set. Once again, it makes sense: it's just not particularly *helpful*.

To make R behave more sensibly in this situation, you need to specify the `use` argument to the `cor()` function. There are several different values that you can specify for this, but the two that we care most about in practice tend to be "`complete.obs`" and "`pairwise.complete.obs`". If we specify `use = "complete.obs"`, R will completely ignore all cases (i.e., all rows in our `parenthood2` data frame) that have any missing values at all. So, for instance, if you look back at the extract earlier when I used the `head()` function, notice that observation 1 (i.e., day 1) of the `parenthood2` data set is missing the value for `baby.sleep`, but is otherwise complete? Well, if you choose `use = "complete.obs"` R will ignore that row completely: that is, even when it's trying to calculate the correlation between `dan.sleep` and `dan.grump`, observation 1 will be ignored, because the value of `baby.sleep` is missing for that observation. Here's what we get:

```
> cor(parenthood2, use = "complete.obs")
    dan.sleep baby.sleep  dan.grump       day
dan.sleep  1.00000000  0.6394985 -0.89951468  0.06132891
baby.sleep  0.63949845  1.0000000 -0.58656066  0.14555814
dan.grump -0.89951468 -0.5865607  1.00000000 -0.06816586
day        0.06132891  0.1455581 -0.06816586  1.00000000
```

The other possibility that we care about, and the one that tends to get used more often in practice, is to set `use = "pairwise.complete.obs"`. When we do that, R only looks at the variables that it's trying to correlate when determining what to drop. So, for instance, since the only missing value for observation 1 of `parenthood2` is for `baby.sleep` R will only drop observation 1 when

⁶It's worth noting that, even though we have missing data for each of these variables, the output doesn't contain any `NA` values. This is because, while `describe()` also has an `na.rm` argument, the default value for this function is `na.rm = TRUE`.

`baby.sleep` is one of the variables involved: and so R keeps observation 1 when trying to correlate `dan.sleep` and `dan.grump`. When we do it this way, here's what we get:

```
> cor(parenthood2, use = "pairwise.complete.obs")
      dan.sleep  baby.sleep  dan.grump       day
dan.sleep    1.00000000  0.61472303 -0.903442442 -0.076796665
baby.sleep   0.61472303  1.00000000 -0.567802669  0.058309485
dan.grump   -0.90344244 -0.56780267  1.000000000  0.005833399
day        -0.07679667  0.05830949  0.005833399  1.000000000
```

Similar, but not quite the same. It's also worth noting that the `correlate()` function (in the `lsr` package) automatically uses the “pairwise complete” method:

```
> correlate(parenthood2)

CORRELATIONS
=====
- correlation type: pearson
- correlations shown only when both variables are numeric

      dan.sleep  baby.sleep  dan.grump       day
dan.sleep      .       0.615     -0.903 -0.077
baby.sleep    0.615      .     -0.568  0.058
dan.grump    -0.903    -0.568      .   0.006
day        -0.077     0.058     0.006     .
```

The two approaches have different strengths and weaknesses. The “pairwise complete” approach has the advantage that it keeps more observations, so you're making use of more of your data and (as we'll discuss in tedious detail in Chapter 4.2 and it improves the reliability of your estimated correlation. On the other hand, it means that every correlation in your correlation matrix is being computed from a slightly different set of observations, which can be awkward when you want to compare the different correlations that you've got.

So which method should you use? It depends a lot on *why* you think your values are missing, and probably depends a little on how paranoid you are. For instance, if you think that the missing values were “chosen” completely randomly⁷ then you'll probably want to use the pairwise method. If you think that missing data are a cue to thinking that the whole observation might be rubbish (e.g., someone just selecting arbitrary responses in your questionnaire),

⁷The technical term here is “missing completely at random” (often written MCAR for short). Makes sense, I suppose, but it does sound ungrammatical to me.

but that there's no pattern to which observations are "rubbish" then it's probably safer to keep only those observations that are complete. If you think there's something systematic going on, in that some observations are more likely to be missing than others, then you have a much trickier problem to solve, and one that is beyond the scope of this book.

7.8.3 Why values are missing: MCAR, MAR, and MNAR

Text by David Schuster

Next, we will expand the definitions of MCAR, MAR, and MNAR:

7.8.3.1 Missing completely at random: MCAR

In MCAR, the distribution of missing data is totally unpredictable from other variables in data set, including the value of missing values. For example: participants may overlook a survey item on the back page. It could be related to a third, unmeasured variable. The implication is that the missingness is effectively random because it cannot be predicted by any variable of interest.

How to handle: While there are no acceptable limits, < 5% is unlikely to make a difference. Deletion of missing cases can be appropriate when data are MCAR.

7.8.3.2 Missing at random: MAR

A better name for this would be missing not-quite-so-randomly. The distribution of missing data is predictable from other variables in the data set. For example: people in a treatment program were not encouraged to complete the survey completely but participants in the control condition were encouraged to complete the survey completely. This would create a pattern where treatment participants would be more likely to have missing data.

How to handle: While there are no easy solutions, multiple imputation can be used with MAR data to estimate missing values.

7.8.3.3 Missing not at random: MNAR

Missing data are related to the variable itself. For example: people with low income are less likely to report their income. This type is the most problematic. You cannot ignore this because it may bias results, even when using multiple imputation (Craig, 2010).

How to handle: If you have variables in the dataset that could explain randomness, you might have MAR instead, which would allow multiple imputation. Therefore, when you suspect MNAR, it is worthwhile to see if you can classify the data as MAR.

7.8.3.4 Planned missingness is not missing data

Data that you have control over not including is not missing data. For example, if you ask participants how well they get along with their siblings, it is not missing data if that field is blank for participants without siblings. Planned missingness is not missing data.

7.8.3.5 Dealing with missingness

I recommend viewing the Video: Dealing with missing data - 30 min for a more complete discussion of the solutions to missing data. Some of your options are:

- Little's test for MCAR tests the null hypothesis that data are MCAR. This may be appropriate as a diagnostic tool, but even a non-significant Little's test does not guarantee MCAR. Further, a significant Little's test does not tell you if the data are MAR or MNAR.
- You may be able to exclude the variable with missing data completely. This may be the best option when data are MNAR, and it can be used no matter the category. If the variable is not central to the research question, this is an easy strategy. I would recommend reporting that you suspect the data are MNAR to explain why you are not reporting them. Of course, if the variable is your IV or DV, proceeding without the variable may not be an option.
- Excluding the missing data. This is the default action in SPSS and available in R using the `na.rm = TRUE`$` argument. The two ways of excluding missing data are listwise (all must be complete to be included in an analysis) and pairwise (use cases wherever possible, even if some variables are missing). **Exclusion of missing data assumes MCAR** to avoid biasing results and also assumes a low proportion are missing.
- Do nothing. If you are reasonably confident you have MCAR and under 5% missing, you probably do not have an issue.
- Using missing data as new data: If know why data are missing, you might use it in your analysis.
- Imputation, which is estimating missing data. There are a variety of methods that have not been demonstrated to have good performance, including prior knowledge, mean substitution (overall or group mean), and using regression. Multiple imputation is the best of these methods but it is complicated. It improves on a regression approach by using sampling (bootstrapping). It uses regression to create a function that estimates the missing cases; random samples are taken with replacement to identify distribution of the variable with missing data. Random samples (about 5) are taken from the distribution of variable with missing data and filled

into the dataset. Statistical analysis is done on each new complete dataset and the statistics are averaged. This procedure assumes MAR or MCAR (Craig, 2010).

7.9 Step 4. Test-specific assumption checking

Text by David Schuster

I would like to note that every statistical test you will run carries certain assumptions. Assumptions may include a required level of measurement or a minimum sample size. Often your data need to be checked to see if they meet the assumptions for the statistical test.

The good news about the statistical techniques we use is that they are sometimes **robust** to violations of some assumptions. Robustness means that statistical conclusion validity is not threatened despite the assumption violation. The bad news about the statistical techniques we use is that some assumptions are more critical than others, and the violation of multiple assumptions can be more severe than just one. Because of this, we need to learn the assumptions of our statistical tests as well as strategies for dealing with assumption violations when they occur. We will look at the assumptions for each test in detail. For now, you should be aware that assumption checking is an important step in preparing for analysis.

7.10 Communicate results of data cleaning in APA style

The first part of your results section should include one or more paragraphs describing the results of your data cleaning. Generally, you can report less detail if you did not find major issues. The bigger the issues, the more detail you will need to explain and justify your methods.

Assume that reader has a professional knowledge of statistical methods. At minimum, the frequency and proportion of missing data should be reported. Empirical evidence and/or theoretical arguments for the causes of data that are missing are helpful. Describe your methods for addressing missing data, if any were used. Finding examples in journal articles can be helpful. Unless you have implemented a new method, do not include R code or output in your manuscript.

Chapter 8

Regression

Text by Navarro (2018)

8.1 Videos

While I had intended on producing my own videos for this chapter, the existing videos I found were just too good.

- Video: Correlation Doesn't Equal Causation - 13 min Quick note about this video: This one and the one on regression are excellent. However, I was surprised to see Karl Pearson's work on the heights of fathers and sons included without editorial comment. The reason *why* Pearson was studying this problem is worth discussing. Oddly, I found an excellent video describing this from another channel from the same organization. Pearson isn't discussed specifically, but both Pearson and Galton were British mathematicians who developed many of techniques we are using today (see their Wikipedia pages for a list). Therefore, their legacy lives on in our work, and you'll find connections between modern psychology and these two in this video: Video: Eugenics and Francis Galton - 13 min
- Video: Regression - 13 min. This is an excellent introduction to GLM and regression. Worth watching more than once.
- Video: How to do regression in R - 6 min

Because it's longer and lower quality than the videos above, you don't need to watch the following video. But I am including it here to provide as many resources as possible.

- Dave's Video: Correlation Basics - 20 min

8.2 Introduction

Text by Navarro (2018)

The goal in this chapter is to introduce *linear regression*, the standard tool that statisticians rely on when analysing the relationship between interval scale predictors and interval scale outcomes. Stripped to its bare essentials, linear regression models are basically a slightly fancier version of the Pearson correlation (Section 8.4) though as we'll see, regression models are much more powerful tools.

8.3 The General Linear Model (GLM)

Text by David Schuster

We have not even done *t*-tests in R yet and we are already doing multiple regression! This is intentional. One theme of this course is that statistical concepts that seem to be independent are sometimes mathematically related. Such is the case with regression. Regression is the mathematical basis of most of the inferential statistical procedures we use, including *t*-tests, *F*-tests (including all variations of analysis of variance, also called ANOVA), and, of course, correlation and multiple regression. Besides introducing linear regression, which is powerful and versatile, this chapter will introduce the mathematical concept that underlies all of these procedures, the **general linear model** (GLM). As an example of GLM, we will see that linear regression is perhaps the more powerful and versatile technique we will discuss in this class, because everything that follows can be implemented as a GLM model.

The key concept of the GLM is that it measures a linear relationship between one or more outcome variables (in our research they will usually be the dependent variable) and one or more predictor variables (usually independent variables). GLM models that have a single outcome variable are **univariate statistics**. GLM modeling with a single predictor is called **simple regression**. In this chapter, we will discover that we can add unlimited number of predictors in a linear model (although we will discover that there are situations where we can have too many variables), which is what defines **multiple regression** (it contains *multiple* independent variables). Our course will focus on univariate statistics. You should be aware, however, that a more complex version exists; we can combine multiple dependent variables on the left side of the equation. Including more than one outcome variable in a model is called **multivariate statistics**.

8.3.1 The Traditional approach: Two kinds of parametric statistical tests

Statistics teachers and textbook authors sometimes de-emphasize GLM in favor of presenting statistical tests as being one of two types:

- Statistical tests that measure linear relationships (e.g., correlation)
- Statistical tests that measure differences between groups (e.g., z-test)

This characterization can be useful, as researchers find themselves wanting to measure a linear relationship (i.e., Is there a correlation between A and B?) or measure differences between groups (i.e., Is there a difference between Condition A and Condition B?). Consequently, linear relationships are useful when researchers are doing observational research. You can measure two quantitative variables and see if they are related. Meanwhile, experimental researchers want to see if their discrete manipulation (participants are either in Condition A or Condition B; it is a discrete IV) causes a change in their dependent variable. This makes multiple regression more convenient for researchers because they have continuous variables. Meanwhile, experimental researchers often have a discrete IV, so software designed to run an ANOVA can make this analysis more convenient than adapting it to run as a multiple regression. It is important to note that this is merely convenience. Both procedures use GLM.

This distinction between linear relationships and group differences can be misleading when researchers think that these two techniques are separate. Any t-test or ANOVA could be run as a multiple regression. Or, worse, researchers may mix up the *research design* and the *statistical analysis*. Therefore, it is important to state:

- Regression can be used with continuous or discrete predictor variables. There is a rule about this; discrete predictors in a regression must be **dichotomous** (only having two possible values). A yes/no question could be added to a regression model. A question asking participants their favorite color could not be added to a regression model without an additional step. Non-dichotomous, discrete predictors are added to regression models using a technique called **dummy coding**, which turns a categorical variable into a series of dichotomous variables (we will learn how to do this later on). Dummy coding would lead researchers to the same answer as running an ANOVA.
- ANOVA is designed for discrete predictor variables (with as many categories or levels as you would like). ANOVA can be used with continuous predictor variables if we make them discrete, but that is rarely advisable. One way to do this is a **median split** (scores below the median are 0 and scores above the median are 1). If one participant scores just below the

median and another very far below the median, they will both be scored the same. Therefore, this throws away variance and can reduce statistical power. For this reason, multiple regression is the more adaptable technique. That said, we will see that we use some ANOVA calculations on our regression models when we do null hypothesis significance testing (NHST).

- All GLM procedures can be used with non-experimental, quasi-experimental, or experimental research designs. **Whether inferences can be made about causality is affected by the research design, not the choice of statistical technique.** Both regression and ANOVA can show us if two variables are related. Thus, which one is the right statistic depends on the type of measurement used in the study, not whether the study is an experiment, quasi-experiment, or non-experiment. If you have two continuous variables, you should use a correlation. If you have a discrete IV and a continuous DV, you could use either correlation or a t-test. You can use a correlation to analyze experiments, quasi-experiments, or non-experiments. You can use a t-test to analyze experiments, quasi-experiments, or non-experiments. A correlation analysis is not the same thing as a “correlational research design.” For this reason, “experiment, quasi-experiment, and non-experiment” are much clearer labels.
- The same model and data tested as a regression or as an ANOVA will yield the same mean differences, effect size, and p-value. **Regression and ANOVA, because they are both GLM, have the same statistical power.**

All GLM procedures are **parametric statistics**, which means that they make assumptions about the population distributions from which we sample. In practice, this means that parametric statistics typically have more assumptions than their **nonparametric** alternatives. We need to learn which assumptions are important and how we can check our data to see if assumptions are met. It is an oversimplification to say that nonparametric stats have no assumptions; they just have fewer assumptions.

8.3.2 The GLM approach

$$Y = BX + I + E$$

With GLM, we are using the equation for a line to assess and describe the relationship between variables. You will see this equation several times in this chapter with different letters labeling each element. The elements are:

- Y , an outcome variable, which can also be called the regressand, response variable, predicted variable, or output variable

- X , a predictor variable, which can also be called the regressor, independent variable, treatment variable, or factor
- B , a model parameter, also called the weight, slope or the coefficient, which defines the relationship between X and Y
- I , the intercept, which gives the estimated value of Y when $X = 0$. It is also the y-intercept of the regression line.
- E , error, also called the residual

The techniques in this chapter expand upon this approach of creating a model and using it to estimate an outcome variable. For example, each of these elements can actually be a list (a matrix) of observations. Further, we can expand this model such that XB is a combination of several predictor variables, not just one. Or, we could create an even simpler model with no predictors.

It might help to illustrate this with an example. In Kozma and Stones (1983) found that good health predicted happiness. Imagine that I wanted to predict your score on this measure right now. However, I know nothing about you. What could serve as my best guess of your happiness? I could use the population mean. That would result in a model $Y = I + E$, where Y is your predicted happiness, predicted on the basis of the population mean ($I = \mu$) and nothing else. As there are no variables, there would be no XB term. I could probably do a bit better if I also knew your health score. This new model would be $Y = XB + I + E$, where X is your health score, and B is a model parameter that explains how to adjust your predicted happiness score based on the health score (technically, it says to increase the predicted Y by B units for every 1 unit increase of X). We have just created a linear regression model, which is a use of the GLM. We have also shown how the addition of predictors (assuming they are good predictors that independently correlate with our outcome variable) can increase the predictive power of a model. This pattern could continue if we added an additional explanatory variable, leading to a multiple regression equation in the form $Y = X_1B_1 + X_2B_2 + I + E$.

8.4 Correlations

Text by Navarro (2018)

Up to this point we have focused entirely on how to construct descriptive statistics for a single variable. What we haven't done is talked about how to describe the relationships *between* variables in the data. To do that, we want to talk mostly about the **correlation** between variables. But first, we need some data.

8.4.1 The data

After spending so much time looking at the AFL data, I'm starting to get bored with sports. Instead, let's turn to a topic close to every parent's heart: sleep. The following data set is fictitious, but based on real events. Suppose I'm curious to find out how much my infant son's sleeping habits affect my mood. Let's say that I can rate my grumpiness very precisely, on a scale from 0 (not at all grumpy) to 100 (grumpy as a very, very grumpy old man). And, lets also assume that I've been measuring my grumpiness, my sleeping patterns and my son's sleeping patterns for quite some time now. Let's say, for 100 days. And, being a nerd, I've saved the data as a file called `parenthood.Rdata`. If we load the data...

```
load( "./data/parenthood.Rdata" )
who(TRUE)
```

```
##   -- Name --      -- Class --      -- Size --
##   parenthood     data.frame     100 x 4
##   $dan.sleep     numeric       100
##   $baby.sleep    numeric       100
##   $dan.grump    numeric       100
##   $day           integer       100
```

... we see that the file contains a single data frame called `parenthood`, which contains four variables `dan.sleep`, `baby.sleep`, `dan.grump` and `day`. If we peek at the data using `head()` out the data, here's what we get:

```
head(parenthood, 10)
```

```
##   dan.sleep baby.sleep dan.grump day
## 1      7.59     10.18      56    1
## 2      7.91     11.66      60    2
## 3      5.14      7.92      82    3
## 4      7.71      9.61      55    4
## 5      6.68      9.75      67    5
## 6      5.99      5.04      72    6
## 7      8.19     10.45      53    7
## 8      7.19      8.27      60    8
## 9      7.40      6.06      60    9
## 10     6.58      7.09      71   10
```

Next, I'll calculate some basic descriptive statistics:

```
describe( parenthood )
```

```
##          vars   n   mean     sd median trimmed    mad   min   max range
## dan.sleep     1 100  6.97  1.02    7.03     7.00  1.09  4.84  9.00  4.16
## baby.sleep    2 100  8.05  2.07    7.95     8.05  2.33  3.25 12.07  8.82
## dan.grump     3 100 63.71 10.05   62.00    63.16  9.64 41.00 91.00 50.00
## day           4 100 50.50 29.01   50.50    50.50 37.06  1.00 100.00 99.00
##                      skew kurtosis   se
## dan.sleep   -0.29    -0.72 0.10
## baby.sleep  -0.02    -0.69 0.21
## dan.grump   0.43    -0.16 1.00
## day         0.00    -1.24 2.90
```

Finally, to give a graphical depiction of what each of the three interesting variables looks like, Figure 8.1 plots histograms.

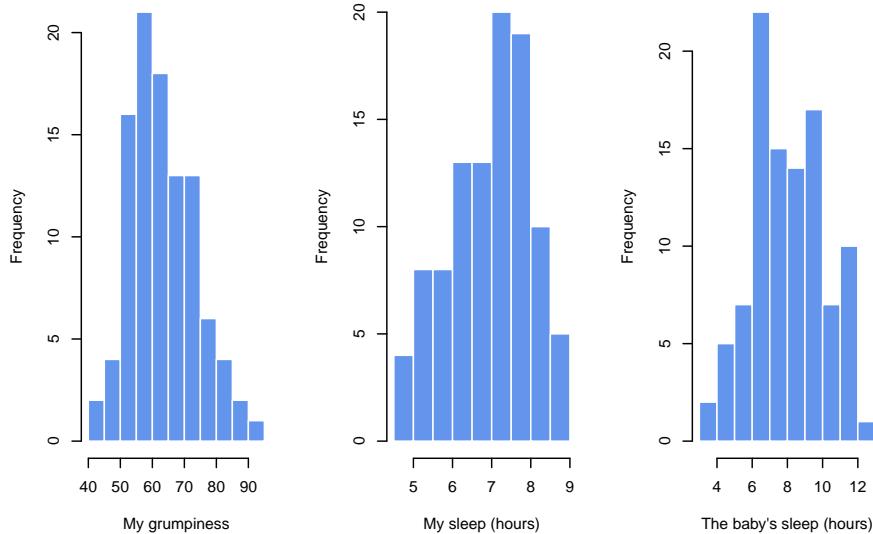


Figure 8.1: Histograms for the three interesting variables in the `parenthood` data set

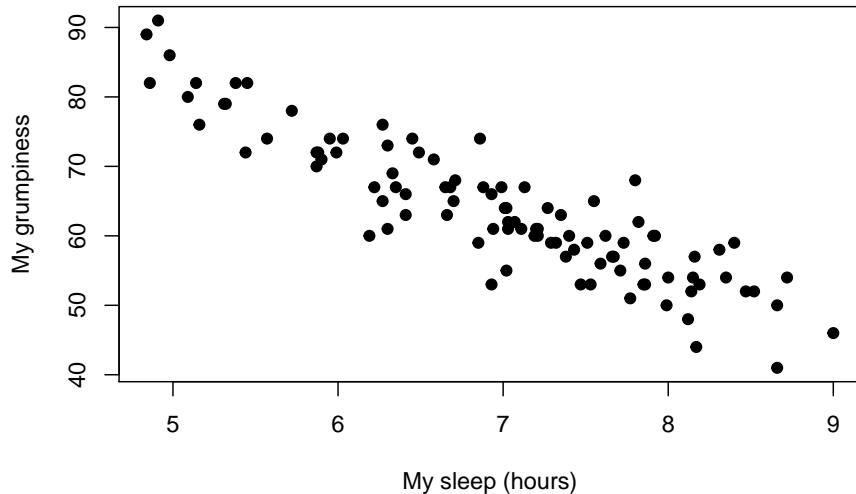
One thing to note: just because R can calculate dozens of different statistics doesn't mean you should report all of them. If I were writing this up for a report, I'd probably pick out those statistics that are of most interest to me (and to my readership), and then put them into a nice, simple table like the

Table 8.1: Descriptive statistics for the parenthood data.

variable	min	max	mean	median	std. dev	IQR
Dan's grumpiness	41	91	63.71	62	10.05	14
Dan's hours slept	4.84	9	6.97	7.03	1.02	1.45
Dan's son's hours slept	3.25	12.07	8.05	7.95	2.07	3.21

one in Table 8.1.¹ Notice that when I put it into a table, I gave everything “human readable” names. This is always good practice. Notice also that I’m not getting enough sleep. This isn’t good practice, but other parents tell me that it’s standard practice.

8.4.2 The strength and direction of a relationship

Figure 8.2: Scatterplot showing the relationship between `dan.sleep` and `dan.grump`

We can draw scatterplots to give us a general sense of how closely related two variables are. Ideally though, we might want to say a bit more about it

¹ Actually, even that table is more than I’d bother with. In practice most people pick *one* measure of central tendency, and *one* measure of variability only.

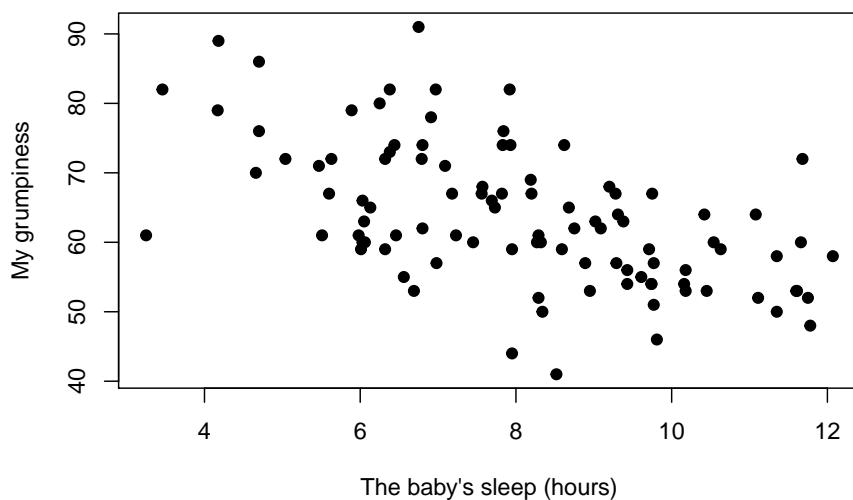


Figure 8.3: Scatterplot showing the relationship between `baby.sleep` and `dan.grump`

than that. For instance, let's compare the relationship between `dan.sleep` and `dan.grump` (Figure 8.2) with that between `baby.sleep` and `dan.grump` (Figure 8.3). When looking at these two plots side by side, it's clear that the relationship is *qualitatively* the same in both cases: more sleep equals less grump! However, it's also pretty obvious that the relationship between `dan.sleep` and `dan.grump` is *stronger* than the relationship between `baby.sleep` and `dan.grump`. The plot on the left is “neater” than the one on the right. What it feels like is that if you want to predict what my mood is, it'd help you a little bit to know how many hours my son slept, but it'd be *more* helpful to know how many hours I slept.

In contrast, let's consider Figure 8.3 vs. Figure 8.4. If we compare the scatterplot of “`baby.sleep v dan.grump`” to the scatterplot of “`baby.sleep v dan.sleep`”, the overall strength of the relationship is the same, but the direction is different. That is, if my son sleeps more, I get *more* sleep (positive relationship), but if he sleeps more then I get *less* grumpy (negative relationship).

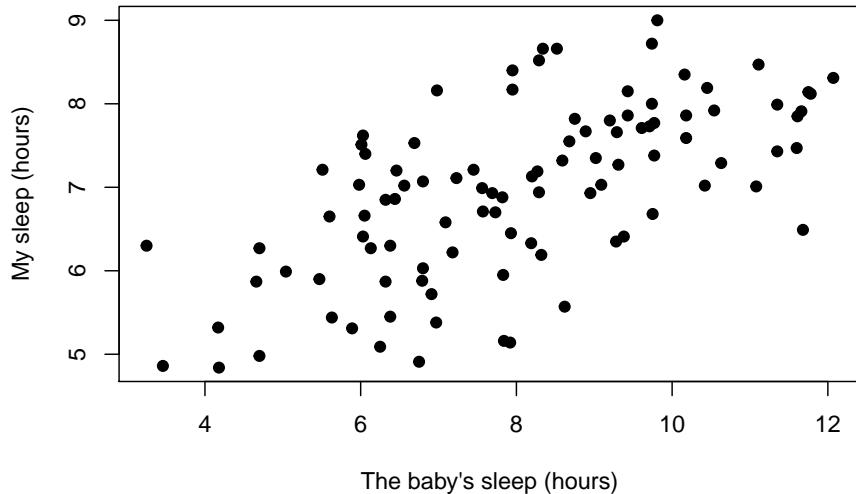


Figure 8.4: Scatterplot showing the relationship between `baby.sleep` and `dan.sleep`

8.4.3 The correlation coefficient

We can make these ideas a bit more explicit by introducing the idea of a **correlation coefficient** (or, more specifically, Pearson's correlation coefficient), which is traditionally denoted by r . The correlation coefficient between two

variables X and Y (sometimes denoted r_{XY}), which we'll define more precisely in the next section, is a measure that varies from -1 to 1 . When $r = -1$ it means that we have a perfect negative relationship, and when $r = 1$ it means we have a perfect positive relationship. When $r = 0$, there's no relationship at all. If you look at Figure 8.5, you can see several plots showing what different correlations look like.

The formula for the Pearson's correlation coefficient can be written in several different ways. I think the simplest way to write down the formula is to break it into two steps. Firstly, let's introduce the idea of a **covariance**. The covariance between two variables X and Y is a generalisation of the notion of the variance; it's a mathematically simple way of describing the relationship between two variables that isn't terribly informative to humans:

$$\text{Cov}(X, Y) = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

Because we're multiplying (i.e., taking the “product” of) a quantity that depends on X by a quantity that depends on Y and then averaging², you can think of the formula for the covariance as an “average cross product” between X and Y . The covariance has the nice property that, if X and Y are entirely unrelated, then the covariance is exactly zero. If the relationship between them is positive (in the sense shown in Figure@reffig:corr) then the covariance is also positive; and if the relationship is negative then the covariance is also negative. In other words, the covariance captures the basic qualitative idea of correlation. Unfortunately, the raw magnitude of the covariance isn't easy to interpret: it depends on the units in which X and Y are expressed, and worse yet, the actual units that the covariance itself is expressed in are really weird. For instance, if X refers to the `dan.sleep` variable (units: hours) and Y refers to the `dan.grump` variable (units: grumps), then the units for their covariance are “hours \times grumps”. And I have no freaking idea what that would even mean.

The Pearson correlation coefficient r fixes this interpretation problem by standardising the covariance, in pretty much the exact same way that the z -score standardises a raw score: by dividing by the standard deviation. However, because we have two variables that contribute to the covariance, the standardisation only works if we divide by both standard deviations.³ In other words, the correlation between X and Y can be written as follows:

$$r_{XY} = \frac{\text{Cov}(X, Y)}{\hat{\sigma}_X \hat{\sigma}_Y}$$

By doing this standardisation, not only do we keep all of the nice properties of the covariance discussed earlier, but the actual values of r are on a meaningful

²Just like we saw with the variance and the standard deviation, in practice we divide by $N-1$ rather than N .

³This is an oversimplification, but it'll do for our purposes.

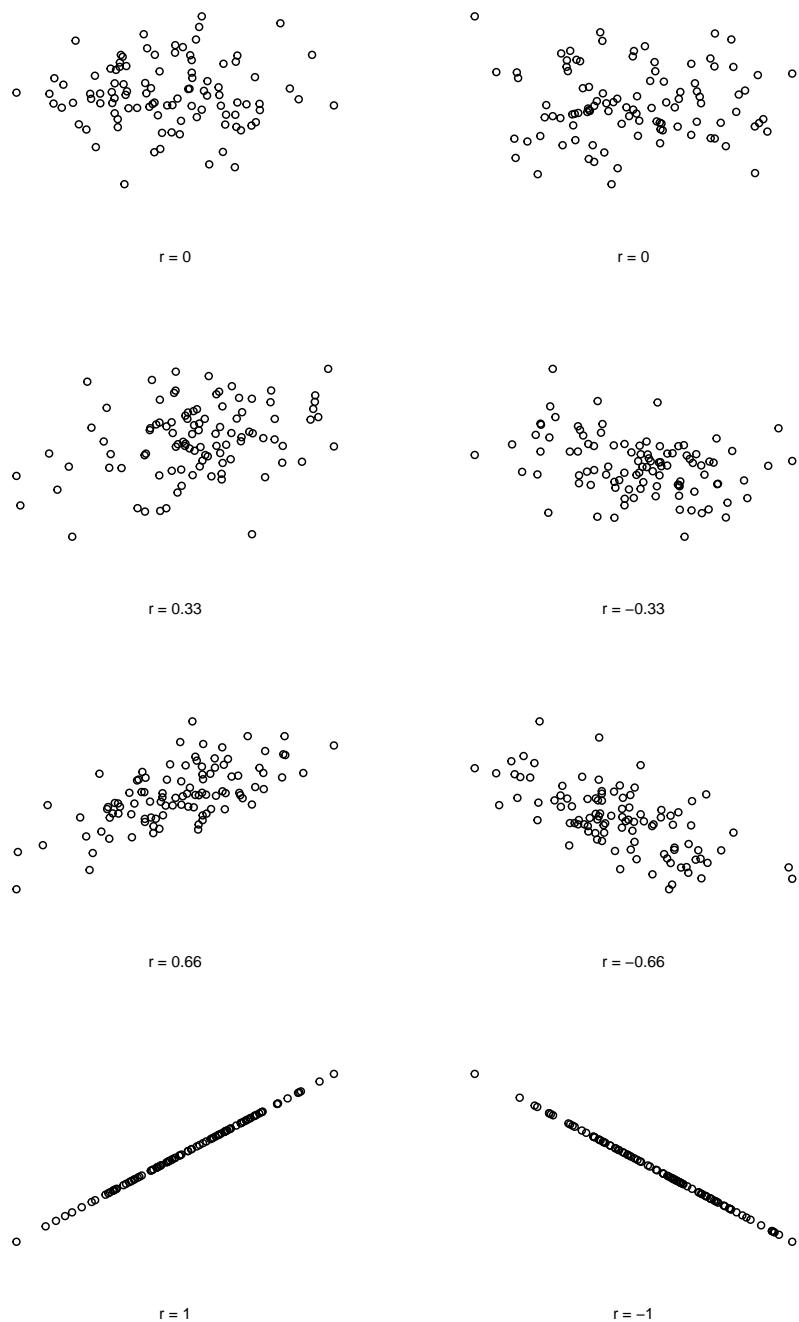


Figure 8.5: Illustration of the effect of varying the strength and direction of a correlation

scale: $r = 1$ implies a perfect positive relationship, and $r = -1$ implies a perfect negative relationship. I'll expand a little more on this point later, in Section@refsec:interpretingcorrelations. But before I do, let's look at how to calculate correlations in R.

8.4.4 Calculating correlations in R

Calculating correlations in R can be done using the `cor()` command. The simplest way to use the command is to specify two input arguments `x` and `y`, each one corresponding to one of the variables. The following extract illustrates the basic usage of the function:⁴

```
cor( x = parenthood$dan.sleep, y = parenthood$dan.grump )
## [1] -0.903384
```

However, the `cor()` function is a bit more powerful than this simple example suggests. For example, you can also calculate a complete “correlation matrix”, between all pairs of variables in the data frame:⁵

```
# correlate all pairs of variables in "parenthood":
cor( x = parenthood )

##          dan.sleep  baby.sleep  dan.grump       day
## dan.sleep  1.00000000  0.62794934 -0.90338404 -0.09840768
## baby.sleep  0.62794934  1.00000000 -0.56596373 -0.01043394
## dan.grump -0.90338404 -0.56596373  1.00000000  0.07647926
## day        -0.09840768 -0.01043394  0.07647926  1.00000000
```

8.4.5 Interpreting a correlation

Naturally, in real life you don't see many correlations of 1. So how should you interpret a correlation of, say $r = .4$? The honest answer is that it really depends on what you want to use the data for, and on how strong the correlations in your field tend to be. A friend of mine in engineering once argued that any

⁴If you are reading this after having already completed Chapter 5 you might be wondering about hypothesis tests for correlations. R has a function called `cor.test()` that runs a hypothesis test for a single correlation, and the `psych` package contains a version called `corr.test()` that can run tests for every correlation in a correlation matrix; hypothesis tests for correlations are discussed in more detail in Section 8.10.

⁵An alternative usage of `cor()` is to correlate one set of variables with another subset of variables. If `X` and `Y` are both data frames with the same number of rows, then `cor(x = X, y = Y)` will produce a correlation matrix that correlates all variables in `X` with all variables in `Y`.

Table 8.2: Rough guide to interpreting correlations

Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive

correlation less than .95 is completely useless (I think he was exaggerating, even for engineering). On the other hand there are real cases – even in psychology – where you should really expect correlations that strong. For instance, one of the benchmark data sets used to test theories of how people judge similarities is so clean that any theory that can't achieve a correlation of at least .9 really isn't deemed to be successful. However, when looking for (say) elementary correlates of intelligence (e.g., inspection time, response time), if you get a correlation above .3 you're doing very very well. In short, the interpretation of a correlation depends a lot on the context. That said, the rough guide in Table 8.2 is pretty typical.

However, something that can never be stressed enough is that you should *always* look at the scatterplot before attaching any interpretation to the data. A correlation might not mean what you think it means. The classic illustration of this is “Anscombe’s Quartet” (Anscombe, 1973), which is a collection of four data sets. Each data set has two variables, an X and a Y . For all four data sets the mean value for X is 9 and the mean for Y is 7.5. The standard deviations for all X variables are almost identical, as are those for the the Y variables. And in each case the correlation between X and Y is $r = 0.816$. You can verify this yourself, since the dataset comes distributed with R. The commands would be:

```
cor( anscombe$x1, anscombe$y1 )
```

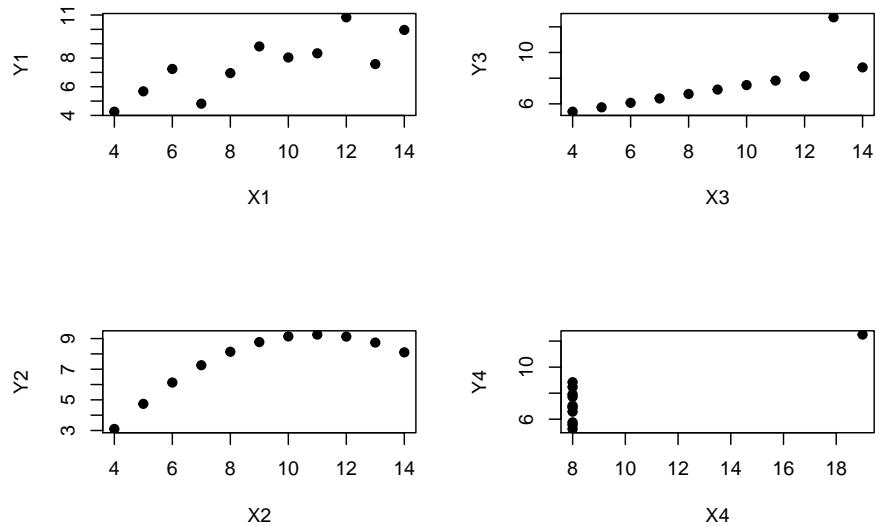
```
## [1] 0.8164205
```

```
cor( anscombe$x2, anscombe$y2 )
```

```
## [1] 0.8162365
```

and so on.

You'd think that these four data sets would look pretty similar to one another. They do not. If we draw scatterplots of X against Y for all four variables, as shown in Figure ?? we see that all four of these are *spectacularly* different to each



other.

The lesson here, which so very many people seem to forget in real life is “*always graph your raw data*”. This will be the focus of Chapter 3.9.

8.4.6 Spearman's rank correlations

The Pearson correlation coefficient is useful for a lot of things, but it does have shortcomings. One issue in particular stands out: what it actually measures is the strength of the *linear* relationship between two variables. In other words, what it gives you is a measure of the extent to which the data all tend to fall on a single, perfectly straight line. Often, this is a pretty good approximation to what we mean when we say “relationship”, and so the Pearson correlation is a good thing to calculate. Sometimes, it isn't.

One very common situation where the Pearson correlation isn't quite the right thing to use arises when an increase in one variable X really is reflected in an increase in another variable Y , but the nature of the relationship isn't necessarily linear. An example of this might be the relationship between effort and reward when studying for an exam. If you put in zero effort (X) into learning a subject, then you should expect a grade of 0% (Y). However, a little bit of effort will cause a *massive* improvement: just turning up to lectures means that you learn

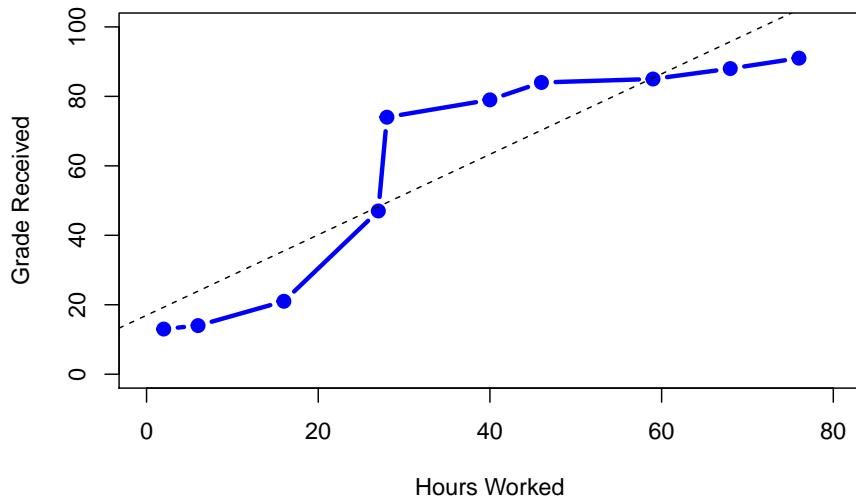


Figure 8.6: The relationship between hours worked and grade received, for a toy data set consisting of only 10 students (each circle corresponds to one student). The dashed line through the middle shows the linear relationship between the two variables. This produces a strong Pearson correlation of $r = .91$. However, the interesting thing to note here is that there's actually a perfect monotonic relationship between the two variables: in this toy example at least, increasing the hours worked always increases the grade received, as illustrated by the solid line. This is reflected in a Spearman correlation of $\rho = 1$. With such a small data set, however, it's an open question as to which version better describes the actual relationship involved.

a fair bit, and if you just turn up to classes, and scribble a few things down so your grade might rise to 35%, all without a lot of effort. However, you just don't get the same effect at the other end of the scale. As everyone knows, it takes *a lot* more effort to get a grade of 90% than it takes to get a grade of 55%. What this means is that, if I've got data looking at study effort and grades, there's a pretty good chance that Pearson correlations will be misleading.

To illustrate, consider the data plotted in Figure 8.6, showing the relationship between hours worked and grade received for 10 students taking some class. The curious thing about this – highly fictitious – data set is that increasing your effort *always* increases your grade. It might be by a lot or it might be by a little, but increasing effort will never decrease your grade. The data are stored in `effort.Rdata`:

```
> load( "effort.Rdata" )
> who(TRUE)
-- Name -- -- Class -- -- Size --
effort      data.frame   10 x 2
$hours       numeric      10
$grade       numeric      10
```

The raw data look like this:

```
> effort
  hours grade
1     2    13
2    76    91
3    40    79
4     6    14
5    16    21
6    28    74
7    27    47
8    59    85
9    46    84
10   68    88
```

If we run a standard Pearson correlation, it shows a strong relationship between hours worked and grade received,

```
> cor( effort$hours, effort$grade )
[1] 0.909402
```

but this doesn't actually capture the observation that increasing hours worked *always* increases the grade. There's a sense here in which we want to be able to say that the correlation is *perfect* but for a somewhat different notion of what

a “relationship” is. What we’re looking for is something that captures the fact that there is a perfect *ordinal relationship* here. That is, if student 1 works more hours than student 2, then we can guarantee that student 1 will get the better grade. That’s not what a correlation of $r = .91$ says at all.

How should we address this? Actually, it’s really easy: if we’re looking for ordinal relationships, all we have to do is treat the data as if it were ordinal scale! So, instead of measuring effort in terms of “hours worked”, let’s rank all 10 of our students in order of hours worked. That is, student 1 did the least work out of anyone (2 hours) so they get the lowest rank (rank = 1). Student 4 was the next laziest, putting in only 6 hours of work in over the whole semester, so they get the next lowest rank (rank = 2). Notice that I’m using “rank = 1” to mean “low rank”. Sometimes in everyday language we talk about “rank = 1” to mean “top rank” rather than “bottom rank”. So be careful: you can rank “from smallest value to largest value” (i.e., small equals rank 1) or you can rank “from largest value to smallest value” (i.e., large equals rank 1). In this case, I’m ranking from smallest to largest, because that’s the default way that R does it. But in real life, it’s really easy to forget which way you set things up, so you have to put a bit of effort into remembering!

Okay, so let’s have a look at our students when we rank them from worst to best in terms of effort and reward:

	rank (hours worked)	rank (grade received)
student 1	1	1
student 2		10
student 3		6
student 4		2
student 5		3
student 6		5
student 7		4
student 8		8
student 9		7
student 10		9

Hm. These are *identical*. The student who put in the most effort got the best grade, the student with the least effort got the worst grade, etc. We can get R to construct these rankings using the `rank()` function, like this:

```
> hours.rank <- rank( effort$hours ) # rank students by hours worked
> grade.rank <- rank( effort$grade ) # rank students by grade received
```

As the table above shows, these two rankings are identical, so if we now correlate them we get a perfect relationship:

```
> cor( hours.rank, grade.rank )
[1] 1
```

What we've just re-invented is *Spearman's rank order correlation*, usually denoted ρ to distinguish it from the Pearson correlation r . We can calculate Spearman's ρ using R in two different ways. Firstly we could do it the way I just showed, using the `rank()` function to construct the rankings, and then calculate the Pearson correlation on these ranks. However, that's way too much effort to do every time. It's much easier to just specify the `method` argument of the `cor()` function.

```
> cor( effort$hours, effort$grade, method = "spearman")
[1] 1
```

The default value of the `method` argument is "`pearson`", which is why we didn't have to specify it earlier on when we were doing Pearson correlations.

8.4.7 The `correlate()` function

As we've seen, the `cor()` function works pretty well, and handles many of the situations that you might be interested in. One thing that many beginners find frustrating, however, is the fact that it's not built to handle non-numeric variables. From a statistical perspective, this is perfectly sensible: Pearson and Spearman correlations are only designed to work for numeric variables, so the `cor()` function spits out an error.

Here's what I mean. Suppose you were keeping track of how many `hours` you worked in any given day, and counted how many `tasks` you completed. If you were doing the tasks for money, you might also want to keep track of how much `pay` you got for each job. It would also be sensible to keep track of the `weekday` on which you actually did the work: most of us don't work as much on Saturdays or Sundays. If you did this for 7 weeks, you might end up with a data set that looks like this one:

```
> load("work.Rdata")
> who(TRUE)
-- Name --  -- Class --  -- Size --
work       data.frame   49 x 7
$hours     numeric      49
$tasks     numeric      49
$pay       numeric      49
$day       integer      49
$weekday   factor       49
```

```
$week      numeric      49
$day.type  factor       49

> head(work)
  hours tasks pay day   weekday week day.type
1  7.2    14  41  1 Tuesday     1 weekday
2  7.4    11  39  2 Wednesday    1 weekday
3  6.6    14  13  3 Thursday    1 weekday
4  6.5    22  47  4 Friday     1 weekday
5  3.1     5   4  5 Saturday    1 weekend
6  3.0     7  12  6 Sunday     1 weekend
```

Obviously, I'd like to know something about how all these variables correlate with one another. I could correlate `hours` with `pay` quite using `cor()`, like so:

```
> cor(work$hours,work$pay)
[1] 0.7604283
```

But what if I wanted a quick and easy way to calculate all pairwise correlations between the numeric variables? I can't just input the `work` data frame, because it contains two factor variables, `weekday` and `day.type`. If I try this, I get an error:

```
> cor(work)
Error in cor(work) : 'x' must be numeric
```

In order to get the correlations that I want using the `cor()` function, is create a new data frame that doesn't contain the factor variables, and then feed that new data frame into the `cor()` function. It's not actually very hard to do that, and I'll talk about how to do it properly in Section ???. But it would be nice to have some function that is smart enough to just ignore the factor variables. That's where the `correlate()` function in the `lsr` package can be handy. If you feed it a data frame that contains factors, it knows to ignore them, and returns the pairwise correlations only between the numeric variables:

```
> correlate(work)

CORRELATIONS
=====
- correlation type: pearson
- correlations shown only when both variables are numeric

  hours tasks pay day weekday week day.type
hours          . 0.800 0.760 -0.049      . 0.018 .
```

```

tasks      0.800      . 0.720 -0.072      . -0.013      .
pay        0.760  0.720      . 0.137      . 0.196      .
day       -0.049 -0.072 0.137      .      . 0.990      .
weekday     .      .      .      .      .      .      .
week       0.018 -0.013 0.196  0.990      .      .      .
day.type    .      .      .      .      .      .      .

```

The output here shows a `.` whenever one of the variables is non-numeric. It also shows a `.` whenever a variable is correlated with itself (it's not a meaningful thing to do). The `correlate()` function can also do Spearman correlations, by specifying the `corr.method` to use:

```

> correlate( work, corr.method="spearman" )

CORRELATIONS
=====
- correlation type: spearman
- correlations shown only when both variables are numeric

      hours  tasks   pay   day weekday   week day.type
hours      . 0.805 0.745 -0.047      . 0.010      .
tasks      0.805      . 0.730 -0.068      . -0.008      .
pay        0.745  0.730      . 0.094      . 0.154      .
day       -0.047 -0.068 0.094      .      . 0.990      .
weekday     .      .      .      .      .      .      .
week       0.010 -0.008 0.154  0.990      .      .      .
day.type    .      .      .      .      .      .      .

```

Obviously, there's no new functionality in the `correlate()` function, and any advanced R user would be perfectly capable of using the `cor()` function to get these numbers out. But if you're not yet comfortable with extracting a subset of a data frame, the `correlate()` function is for you.

8.4.8 Missing values in pairwise calculations

I mentioned earlier that the `cor()` function is a special case. It doesn't have an `na.rm` argument, because the story becomes a lot more complicated when more than one variable is involved. What it does have is an argument called `use` which does roughly the same thing, but you need to think little more carefully about what you want this time. To illustrate the issues, let's open up a data set that has missing values, `parenthood2.Rdata`. This file contains the same data as the original `parenthood` data, but with some values deleted. It contains a single data frame, `parenthood2`:

```
> load( "parenthood2.Rdata" )
> print( parenthood2 )
  dan.sleep baby.sleep dan.grump day
1      7.59        NA       56   1
2      7.91     11.66      60   2
3      5.14      7.92      82   3
4      7.71      9.61      55   4
5      6.68      9.75      NA   5
6      5.99      5.04      72   6
BLAH BLAH BLAH
```

If I calculate my descriptive statistics using the `describe()` function

```
> describe( parenthood2 )
    var   n   mean     sd median trimmed   mad   min   max   BLAH
dan.sleep  1  91  6.98  1.02    7.03    7.02  1.13  4.84  9.00  BLAH
baby.sleep 2  89  8.11  2.05    8.20    8.13  2.28  3.25 12.07  BLAH
dan.grump  3  92 63.15  9.85   61.00   62.66 10.38 41.00 89.00  BLAH
day        4 100 50.50 29.01   50.50   50.50 37.06  1.00 100.00  BLAH
```

we can see from the `n` column that there are 9 missing values for `dan.sleep`, 11 missing values for `baby.sleep` and 8 missing values for `dan.grump`.⁶ Suppose what I would like is a correlation matrix. And let's also suppose that I don't bother to tell R how to handle those missing values. Here's what happens:

```
> cor( parenthood2 )
  dan.sleep baby.sleep dan.grump day
dan.sleep      1        NA       NA  NA
baby.sleep     NA        1       NA  NA
dan.grump      NA        NA        1  NA
day            NA        NA       NA  1
```

Annoying, but it kind of makes sense. If I don't *know* what some of the values of `dan.sleep` and `baby.sleep` actually are, then I can't possibly *know* what the correlation between these two variables is either, since the formula for the correlation coefficient makes use of every single observation in the data set. Once again, it makes sense: it's just not particularly *helpful*.

To make R behave more sensibly in this situation, you need to specify the `use` argument to the `cor()` function. There are several different values that you can specify for this, but the two that we care most about in practice tend to be "`complete.obs`" and "`pairwise.complete.obs`". If we specify

⁶It's worth noting that, even though we have missing data for each of these variables, the output doesn't contain any `NA` values. This is because, while `describe()` also has an `na.rm` argument, the default value for this function is `na.rm = TRUE`.

`use = "complete.obs"`, R will completely ignore all cases (i.e., all rows in our `parenthood2` data frame) that have any missing values at all. So, for instance, if you look back at the extract earlier when I used the `head()` function, notice that observation 1 (i.e., day 1) of the `parenthood2` data set is missing the value for `baby.sleep`, but is otherwise complete? Well, if you choose `use = "complete.obs"` R will ignore that row completely: that is, even when it's trying to calculate the correlation between `dan.sleep` and `dangrump`, observation 1 will be ignored, because the value of `baby.sleep` is missing for that observation. Here's what we get:

```
> cor(parenthood2, use = "complete.obs")
      dan.sleep baby.sleep   dan.grump       day
dan.sleep    1.00000000  0.6394985 -0.89951468  0.06132891
baby.sleep   0.63949845  1.0000000 -0.58656066  0.14555814
dan.grump   -0.89951468 -0.5865607  1.00000000 -0.06816586
day         0.06132891  0.1455581 -0.06816586  1.00000000
```

The other possibility that we care about, and the one that tends to get used more often in practice, is to set `use = "pairwise.complete.obs"`. When we do that, R only looks at the variables that it's trying to correlate when determining what to drop. So, for instance, since the only missing value for observation 1 of `parenthood2` is for `baby.sleep` R will only drop observation 1 when `baby.sleep` is one of the variables involved: and so R keeps observation 1 when trying to correlate `dan.sleep` and `dangrump`. When we do it this way, here's what we get:

```
> cor(parenthood2, use = "pairwise.complete.obs")
      dan.sleep baby.sleep   dan.grump       day
dan.sleep    1.00000000  0.61472303 -0.903442442 -0.076796665
baby.sleep   0.61472303  1.00000000 -0.567802669  0.058309485
dan.grump   -0.90344244 -0.56780267  1.000000000  0.005833399
day         -0.07679667  0.05830949  0.005833399  1.000000000
```

Similar, but not quite the same. It's also worth noting that the `correlate()` function (in the `lsr` package) automatically uses the “pairwise complete” method:

```
> correlate(parenthood2)

CORRELATIONS
=====
- correlation type: pearson
- correlations shown only when both variables are numeric
```

	dan.sleep	baby.sleep	dan.grump	day
dan.sleep	.	0.615	-0.903	-0.077
baby.sleep	0.615	.	-0.568	0.058
dan.grump	-0.903	-0.568	.	0.006
day	-0.077	0.058	0.006	.

The two approaches have different strengths and weaknesses. The “pairwise complete” approach has the advantage that it keeps more observations, so you’re making use of more of your data and (as we’ll discuss in tedious detail in Chapter 4.2) it improves the reliability of your estimated correlation. On the other hand, it means that every correlation in your correlation matrix is being computed from a slightly different set of observations, which can be awkward when you want to compare the different correlations that you’ve got.

So which method should you use? It depends a lot on *why* you think your values are missing, and probably depends a little on how paranoid you are. For instance, if you think that the missing values were “chosen” completely randomly⁷ then you’ll probably want to use the pairwise method. If you think that missing data are a cue to thinking that the whole observation might be rubbish (e.g., someone just selecting arbitrary responses in your questionnaire), but that there’s no pattern to which observations are “rubbish” then it’s probably safer to keep only those observations that are complete. If you think there’s something systematic going on, in that some observations are more likely to be missing than others, then you have a much trickier problem to solve, and one that is beyond the scope of this book.

8.5 Linear regression

Text by Navarro (2018)

Since the basic ideas in regression are closely tied to correlation, we’ll return to the `parenthood.Rdata` file that we were using to illustrate how correlations work. Recall that, in this data set, we were trying to find out why Dan is so very grumpy all the time, and our working hypothesis was that I’m not getting enough sleep. We drew some scatterplots to help us examine the relationship between the amount of sleep I get, and my grumpiness the following day. The actual scatterplot that we draw is the one shown in Figure 8.7, and as we saw previously this corresponds to a correlation of $r = -.90$, but what we find ourselves secretly imagining is something that looks closer to Figure 8.8. That is, we mentally draw a straight line through the middle of the data. In statistics, this line that we’re drawing is called a **regression line**. Notice that the regression line goes through the middle of the data. We don’t find ourselves imagining anything like the rather silly plot shown in Figure 8.9.

⁷The technical term here is “missing completely at random” (often written MCAR for short). Makes sense, I suppose, but it does sound ungrammatical to me.

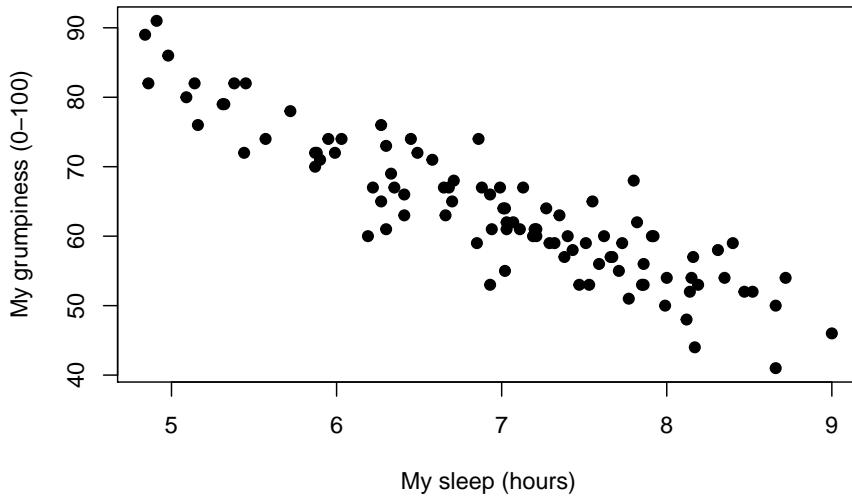


Figure 8.7: Scatterplot showing grumpiness as a function of hours slept.

This is not highly surprising: the line that I've drawn in Figure 8.9 doesn't "fit" the data very well, so it doesn't make a lot of sense to propose it as a way of summarising the data, right? This is a very simple observation to make, but it turns out to be very powerful when we start trying to wrap just a little bit of maths around it. To do so, let's start with a refresher of some high school maths. The formula for a straight line is usually written like this:

$$y = mx + c$$

Or, at least, that's what it was when I went to high school all those years ago. The two *variables* are x and y , and we have two *coefficients*, m and c . The coefficient m represents the *slope* of the line, and the coefficient c represents the *y-intercept* of the line. Digging further back into our decaying memories of high school (sorry, for some of us high school was a long time ago), we remember that the intercept is interpreted as "the value of y that you get when $x = 0$ ". Similarly, a slope of m means that if you increase the x -value by 1 unit, then the y -value goes up by m units; a negative slope means that the y -value would go down rather than up. Ah yes, it's all coming back to me now.

Now that we've remembered that, it should come as no surprise to discover that we use the exact same formula to describe a regression line. If Y is the outcome variable (the DV) and X is the predictor variable (the IV), then the formula

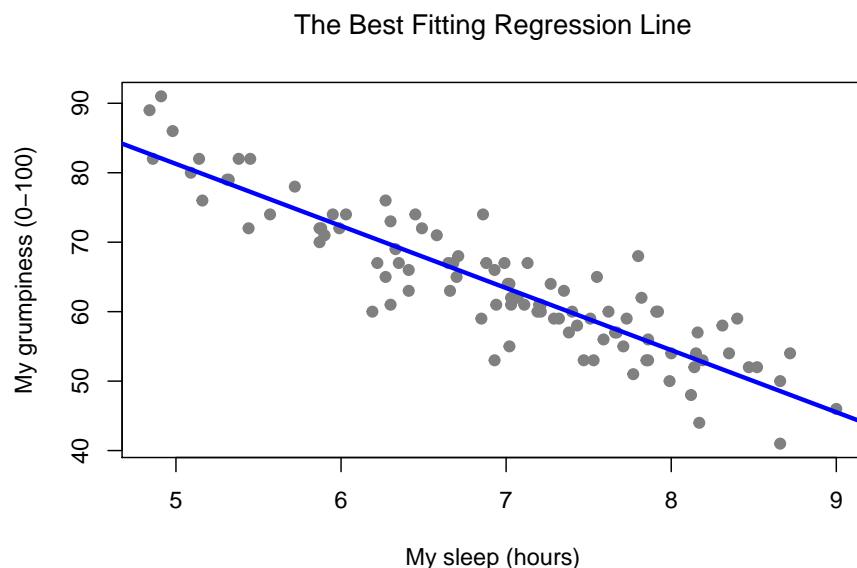


Figure 8.8: Panel a shows the sleep-grumpiness scatterplot from above with the best fitting regression line drawn over the top. Not surprisingly, the line goes through the middle of the data.

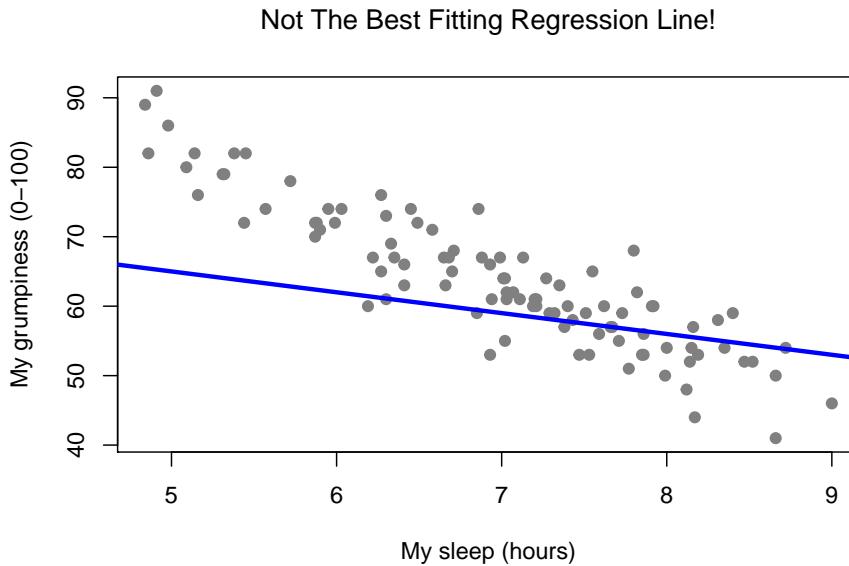


Figure 8.9: In contrast, this plot shows the same data, but with a very poor choice of regression line drawn over the top.

that describes our regression is written like this:

$$\hat{Y}_i = b_1 X_i + b_0$$

Hm. Looks like the same formula, but there's some extra frilly bits in this version. Let's make sure we understand them. Firstly, notice that I've written X_i and \hat{Y}_i rather than just plain old X and Y . This is because we want to remember that we're dealing with actual data. In this equation, X_i is the value of predictor variable for the i th observation (i.e., the number of hours of sleep that I got on day i of my little study), and \hat{Y}_i is the corresponding value of the outcome variable (i.e., my grumpiness on that day). And although I haven't said so explicitly in the equation, what we're assuming is that this formula works for all observations in the data set (i.e., for all i). Secondly, notice that I wrote \hat{Y}_i and not Y_i . This is because we want to make the distinction between the *actual data* Y_i , and the *estimate* \hat{Y}_i (i.e., the prediction that our regression line is making). Thirdly, I changed the letters used to describe the coefficients from m and c to b_1 and b_0 . That's just the way that statisticians like to refer to the coefficients in a regression model. I've no idea why they chose b , but that's what they did. In any case b_0 always refers to the intercept term, and b_1 refers to the slope.

Excellent, excellent. Next, I can't help but notice that – regardless of whether

we're talking about the good regression line or the bad one – the data don't fall perfectly on the line. Or, to say it another way, the data Y_i are not identical to the predictions of the regression model \hat{Y}_i . Since statisticians love to attach letters, names and numbers to everything, let's refer to the difference between the model prediction and that actual data point as a *residual*, and we'll refer to it as ϵ_i .⁸ Written using mathematics, the residuals are defined as:

$$\epsilon_i = Y_i - \hat{Y}_i$$

which in turn means that we can write down the complete linear regression model as:

$$Y_i = b_1 X_i + b_0 + \epsilon_i$$

8.6 Estimating a linear regression model

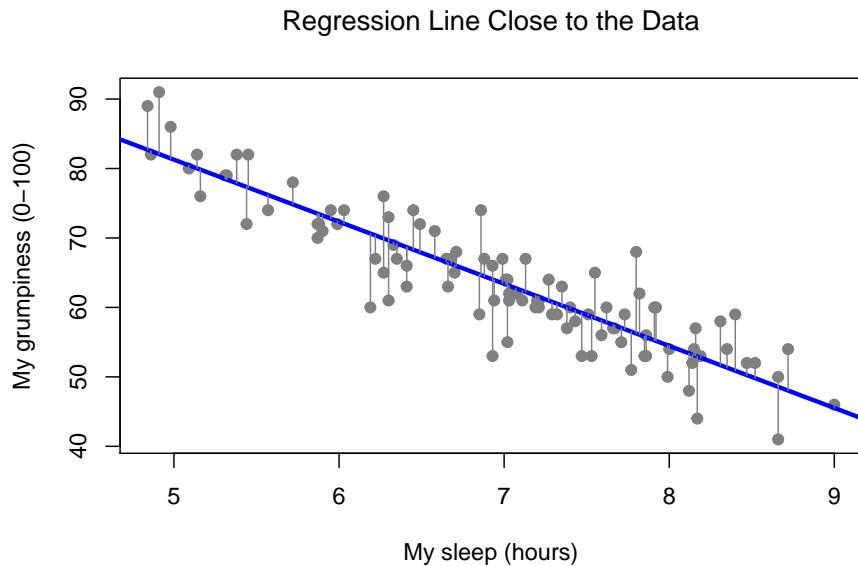


Figure 8.10: A depiction of the residuals associated with the best fitting regression line

Okay, now let's redraw our pictures, but this time I'll add some lines to show the size of the residual for all observations. When the regression line is good, our residuals (the lengths of the solid black lines) all look pretty small, as shown

⁸The ϵ symbol is the Greek letter epsilon. It's traditional to use ϵ_i or e_i to denote a residual.

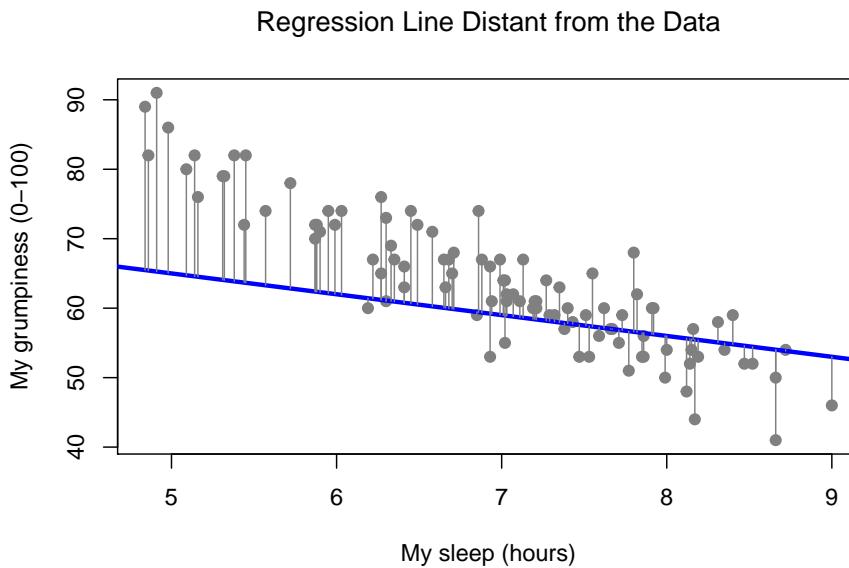


Figure 8.11: The residuals associated with a poor regression line

in Figure 8.10, but when the regression line is a bad one, the residuals are a lot larger, as you can see from looking at Figure 8.11. Hm. Maybe what we “want” in a regression model is *small* residuals. Yes, that does seem to make sense. In fact, I think I’ll go so far as to say that the “best fitting” regression line is the one that has the smallest residuals. Or, better yet, since statisticians seem to like to take squares of everything why not say that ...

The estimated regression coefficients, \hat{b}_0 and \hat{b}_1 are those that minimise the sum of the squared residuals, which we could either write as $\sum_i(Y_i - \hat{Y}_i)^2$ or as $\sum_i \epsilon_i^2$.

Yes, yes that sounds even better. And since I’ve indented it like that, it probably means that this is the right answer. And since this is the right answer, it’s probably worth making a note of the fact that our regression coefficients are *estimates* (we’re trying to guess the parameters that describe a population!), which is why I’ve added the little hats, so that we get \hat{b}_0 and \hat{b}_1 rather than b_0 and b_1 . Finally, I should also note that – since there’s actually more than one way to estimate a regression model – the more technical name for this estimation process is ***ordinary least squares (OLS) regression***.

At this point, we now have a concrete definition for what counts as our “best” choice of regression coefficients, \hat{b}_0 and \hat{b}_1 . The natural question to ask next is,

if our optimal regression coefficients are those that minimise the sum squared residuals, how do we *find* these wonderful numbers? The actual answer to this question is complicated, and it doesn't help you understand the logic of regression.⁹ As a result, this time I'm going to let you off the hook. Instead of showing you how to do it the long and tedious way first, and then "revealing" the wonderful shortcut that R provides you with, let's cut straight to the chase... and use the `lm()` function (short for "linear model") to do all the heavy lifting.

8.6.1 Using the `lm()` function

The `lm()` function is a fairly complicated one: if you type `?lm`, the help files will reveal that there are a lot of arguments that you can specify, and most of them won't make a lot of sense to you. At this stage however, there's really only two of them that you care about, and as it turns out you've seen them before:

- **formula.** A formula that specifies the regression model. For the simple linear regression models that we've talked about so far, in which you have a single predictor variable as well as an intercept term, this formula is of the form `outcome ~ predictor`. However, more complicated formulas are allowed, and we'll discuss them later.
- **data.** The data frame containing the variables.

The output of the `lm()` function is a fairly complicated object, with quite a lot of technical information buried under the hood. Because this technical information is used by other functions, it's generally a good idea to create a variable that stores the results of your regression. With this in mind, to run my linear regression, the command I want to use is this:

```
regression.1 <- lm( formula = dan.grump ~ dan.sleep,
                     data = parenthood )
```

Note that I used `dan.grump ~ dan.sleep` as the formula: in the model that I'm trying to estimate, `dan.grump` is the *outcome* variable, and `dan.sleep` is the predictor variable. It's always a good idea to remember which one is which!

⁹Or at least, I'm assuming that it doesn't help most people. But on the off chance that someone reading this is a proper kung fu master of linear algebra (and to be fair, I always have a few of these people in my intro stats class), it *will* help *you* to know that the solution to the estimation problem turns out to be $\hat{b} = (X^T X)^{-1} X^T y$, where \hat{b} is a vector containing the estimated regression coefficients, X is the "design matrix" that contains the predictor variables (plus an additional column containing all ones; strictly X is a matrix of the regressors, but I haven't discussed the distinction yet), and y is a vector containing the outcome variable. For everyone else, this isn't exactly helpful, and can be downright scary. However, since quite a few things in linear regression can be written in linear algebra terms, you'll see a bunch of footnotes like this one in this chapter. If you can follow the maths in them, great. If not, ignore it.

Anyway, what this does is create an “`lm` object” (i.e., a variable whose class is “`lm`”) called `regression.1`. Let’s have a look at what happens when we `print()` it out:

```
print( regression.1 )

##
## Call:
## lm(formula = dan.grump ~ dan.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)  dan.sleep
##      125.956     -8.937
```

This looks promising. There’s two separate pieces of information here. Firstly, R is politely reminding us what the command was that we used to specify the model in the first place, which can be helpful. More importantly from our perspective, however, is the second part, in which R gives us the intercept $\hat{b}_0 = 125.96$ and the slope $\hat{b}_1 = -8.94$. In other words, the best-fitting regression line that I plotted in Figure 8.8 has this formula:

$$\hat{Y}_i = -8.94 X_i + 125.96$$

8.6.2 Interpreting the estimated model

The most important thing to be able to understand is how to interpret these coefficients. Let’s start with \hat{b}_1 , the slope. If we remember the definition of the slope, a regression coefficient of $\hat{b}_1 = -8.94$ means that if I increase X_i by 1, then I’m decreasing Y_i by 8.94. That is, each additional hour of sleep that I gain will improve my mood, reducing my grumpiness by 8.94 grumpiness points. What about the intercept? Well, since \hat{b}_0 corresponds to “the expected value of Y_i when X_i equals 0”, it’s pretty straightforward. It implies that if I get zero hours of sleep ($X_i = 0$) then my grumpiness will go off the scale, to an insane value of ($Y_i = 125.96$). Best to be avoided, I think.

8.7 Multiple linear regression

The simple linear regression model that we’ve discussed up to this point assumes that there’s a single predictor variable that you’re interested in, in this case `dan.sleep`. In fact, up to this point, *every* statistical tool that we’ve talked about has assumed that your analysis uses one predictor variable and one outcome variable. However, in many (perhaps most) research projects you actually have multiple predictors that you want to examine. If so, it would be nice to

be able to extend the linear regression framework to be able to include multiple predictors. Perhaps some kind of ***multiple regression*** model would be in order?

Multiple regression is conceptually very simple. All we do is add more terms to our regression equation. Let's suppose that we've got two variables that we're interested in; perhaps we want to use both `dan.sleep` and `baby.sleep` to predict the `dan.grump` variable. As before, we let Y_i refer to my grumpiness on the i -th day. But now we have two X variables: the first corresponding to the amount of sleep I got and the second corresponding to the amount of sleep my son got. So we'll let X_{i1} refer to the hours I slept on the i -th day, and X_{i2} refers to the hours that the baby slept on that day. If so, then we can write our regression model like this:

$$Y_i = b_2 X_{i2} + b_1 X_{i1} + b_0 + \epsilon_i$$

As before, ϵ_i is the residual associated with the i -th observation, $\epsilon_i = Y_i - \hat{Y}_i$. In this model, we now have three coefficients that need to be estimated: b_0 is the intercept, b_1 is the coefficient associated with my sleep, and b_2 is the coefficient associated with my son's sleep. However, although the number of coefficients that need to be estimated has changed, the basic idea of how the estimation works is unchanged: our estimated coefficients \hat{b}_0 , \hat{b}_1 and \hat{b}_2 are those that minimise the sum squared residuals.

8.7.1 Doing it in R

Multiple regression in R is no different to simple regression: all we have to do is specify a more complicated `formula` when using the `lm()` function. For example, if we want to use both `dan.sleep` and `baby.sleep` as predictors in our attempt to explain why I'm so grumpy, then the formula we need is this:

```
dan.grump ~ dan.sleep + baby.sleep
```

Notice that, just like last time, I haven't explicitly included any reference to the intercept term in this formula; only the two predictor variables and the outcome. By default, the `lm()` function assumes that the model should include an intercept (though you can get rid of it if you want). In any case, I can create a new regression model – which I'll call `regression.2` – using the following command:

```
regression.2 <- lm( formula = dan.grump ~ dan.sleep + baby.sleep,
                     data = parenthood )
```

And just like last time, if we `print()` out this regression model we can see what the estimated regression coefficients are:

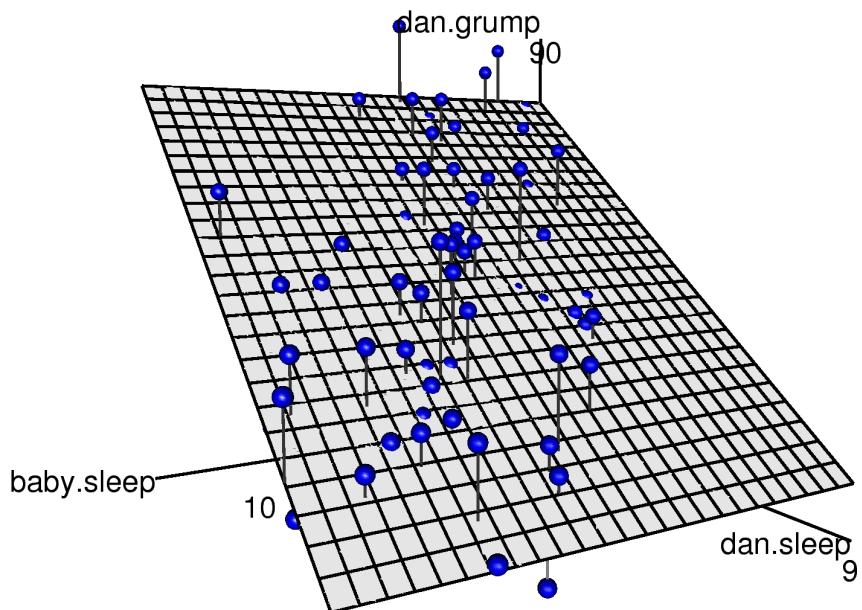


Figure 8.12: A 3D visualisation of a multiple regression model. There are two predictors in the model, `dan.sleep` and `baby.sleep`; the outcome variable is `dan.grump`. Together, these three variables form a 3D space: each observation (blue dots) is a point in this space. In much the same way that a simple linear regression model forms a line in 2D space, this multiple regression model forms a plane in 3D space. When we estimate the regression coefficients, what we're trying to do is find a plane that is as close to all the blue dots as possible.

```
print( regression.2 )

##
## Call:
## lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)  dan.sleep  baby.sleep
## 125.96557    -8.95025     0.01052
```

The coefficient associated with `dan.sleep` is quite large, suggesting that every hour of sleep I lose makes me a lot grumpier. However, the coefficient for `baby.sleep` is very small, suggesting that it doesn't really matter how much sleep my son gets; not really. What matters as far as my grumpiness goes is how much sleep *I* get. To get a sense of what this multiple regression model looks like, Figure 8.12 shows a 3D plot that plots all three variables, along with the regression model itself.

8.7.2 Formula for the general case

The equation that I gave above shows you what a multiple regression model looks like when you include two predictors. Not surprisingly, then, if you want more than two predictors all you have to do is add more X terms and more b coefficients. In other words, if you have K predictor variables in the model then the regression equation looks like this:

$$Y_i = \left(\sum_{k=1}^K b_k X_{ik} \right) + b_0 + \epsilon_i$$

8.8 Quantifying the fit of the regression model

So we now know how to estimate the coefficients of a linear regression model. The problem is, we don't yet know if this regression model is any good. For example, the `regression.1` model *claims* that every hour of sleep will improve my mood by quite a lot, but it might just be rubbish. Remember, the regression model only produces a prediction \hat{Y}_i about what my mood is like: my actual mood is Y_i . If these two are very close, then the regression model has done a good job. If they are very different, then it has done a bad job.

8.8.1 The R^2 value

Once again, let's wrap a little bit of mathematics around this. Firstly, we've got the sum of the squared residuals:

$$\text{SS}_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

which we would hope to be pretty small. Specifically, what we'd like is for it to be very small in comparison to the total variability in the outcome variable,

$$\text{SS}_{tot} = \sum_i (Y_i - \bar{Y})^2$$

While we're here, let's calculate these values in R. Firstly, in order to make my R commands look a bit more similar to the mathematical equations, I'll create variables X and Y:

```
X <- parenthood$dan.sleep # the predictor
Y <- parenthood$dan.grump # the outcome
```

Now that we've done this, let's calculate the \hat{Y} values and store them in a variable called Y.pred. For the simple model that uses only a single predictor, regression.1, we would do the following:

```
Y.pred <- -8.94 * X + 125.97
```

Okay, now that we've got a variable which stores the regression model predictions for how grumpy I will be on any given day, let's calculate our sum of squared residuals. We would do that using the following command:

```
SS.resid <- sum( (Y - Y.pred)^2 )
print( SS.resid )
```

```
## [1] 1838.722
```

Wonderful. A big number that doesn't mean very much. Still, let's forge boldly onwards anyway, and calculate the total sum of squares as well. That's also pretty simple:

```
SS.tot <- sum( (Y - mean(Y))^2 )
print( SS.tot )
```

```
## [1] 9998.59
```

Hm. Well, it's a much bigger number than the last one, so this does suggest that our regression model was making good predictions. But it's not very interpretable.

Perhaps we can fix this. What we'd like to do is to convert these two fairly meaningless numbers into one number. A nice, interpretable number, which for no particular reason we'll call R^2 . What we would like is for the value of R^2 to be equal to 1 if the regression model makes no errors in predicting the data. In other words, if it turns out that the residual errors are zero – that is, if $SS_{res} = 0$ – then we expect $R^2 = 1$. Similarly, if the model is completely useless, we would like R^2 to be equal to 0. What do I mean by “useless”? Tempting as it is demand that the regression model move out of the house, cut its hair and get a real job, I'm probably going to have to pick a more practical definition: in this case, all I mean is that the residual sum of squares is no smaller than the total sum of squares, $SS_{res} = SS_{tot}$. Wait, why don't we do exactly that? The formula that provides us with out R^2 value is pretty simple to write down,

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

and equally simple to calculate in R:

```
R.squared <- 1 - (SS.resid / SS.tot)
print( R.squared )
## [1] 0.8161018
```

The R^2 value, sometimes called the *coefficient of determination*¹⁰ has a simple interpretation: it is the *proportion* of the variance in the outcome variable that can be accounted for by the predictor. So in this case, the fact that we have obtained $R^2 = .816$ means that the predictor (`my.sleep`) explains 81.6% of the variance in the outcome (`my.grump`).

Naturally, you don't actually need to type in all these commands yourself if you want to obtain the R^2 value for your regression model. As we'll see later on in Section 8.9.3, all you need to do is use the `summary()` function. However, let's put that to one side for the moment. There's another property of R^2 that I want to point out.

8.8.2 The relationship between regression and correlation

At this point we can revisit my earlier claim that regression, in this very simple form that I've discussed so far, is basically the same thing as a correlation. Previously, we used the symbol r to denote a Pearson correlation. Might there

¹⁰ And by “sometimes” I mean “almost never”. In practice everyone just calls it “ R -squared”.

be some relationship between the value of the correlation coefficient r and the R^2 value from linear regression? Of course there is: the squared correlation r^2 is identical to the R^2 value for a linear regression with only a single predictor. To illustrate this, here's the squared correlation:

```
r <- cor(X, Y) # calculate the correlation
print( r^2 )      # print the squared correlation
```

```
## [1] 0.8161027
```

Yep, same number. In other words, running a Pearson correlation is more or less equivalent to running a linear regression model that uses only one predictor variable.

8.8.3 The adjusted R^2 value

One final thing to point out before moving on. It's quite common for people to report a slightly different measure of model performance, known as "adjusted R^2 ". The motivation behind calculating the adjusted R^2 value is the observation that adding more predictors into the model will *always* call the R^2 value to increase (or at least not decrease). The adjusted R^2 value introduces a slight change to the calculation, as follows. For a regression model with K predictors, fit to a data set containing N observations, the adjusted R^2 is:

$$\text{adj. } R^2 = 1 - \left(\frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}} \times \frac{N - 1}{N - K - 1} \right)$$

This adjustment is an attempt to take the degrees of freedom into account. The big advantage of the adjusted R^2 value is that when you add more predictors to the model, the adjusted R^2 value will only increase if the new variables improve the model performance more than you'd expect by chance. The big disadvantage is that the adjusted R^2 value *can't* be interpreted in the elegant way that R^2 can. R^2 has a simple interpretation as the proportion of variance in the outcome variable that is explained by the regression model; to my knowledge, no equivalent interpretation exists for adjusted R^2 .

An obvious question then, is whether you should report R^2 or adjusted R^2 . This is probably a matter of personal preference. If you care more about interpretability, then R^2 is better. If you care more about correcting for bias, then adjusted R^2 is probably better. Speaking just for myself, I prefer R^2 : my feeling is that it's more important to be able to interpret your measure of model performance. Besides, as we'll see in Section 8.9, if you're worried that the improvement in R^2 that you get by adding a predictor is just due to chance and not because it's a better model, well, we've got hypothesis tests for that.

8.9 Hypothesis tests for regression models

So far we've talked about what a regression model is, how the coefficients of a regression model are estimated, and how we quantify the performance of the model (the last of these, incidentally, is basically our measure of effect size). The next thing we need to talk about is hypothesis tests. There are two different (but related) kinds of hypothesis tests that we need to talk about: those in which we test whether the regression model as a whole is performing significantly better than a null model; and those in which we test whether a particular regression coefficient is significantly different from zero.

At this point, you're probably groaning internally, thinking that I'm going to introduce a whole new collection of tests...Me too. I'm so sick of hypothesis tests that I'm going to shamelessly reuse the F -test from ANOVA and the t -test. In fact, all I'm going to do in this section is show you how those tests are imported wholesale into the regression framework.

8.9.1 Testing the model as a whole: The omnibus test

Okay, suppose you've estimated your regression model. The first hypothesis test you might want to try is one in which the null hypothesis that there is *no relationship* between the predictors and the outcome, and the alternative hypothesis is that *the data are distributed in exactly the way that the regression model predicts*. Formally, our "null model" corresponds to the fairly trivial "regression" model in which we include 0 predictors, and only include the intercept term b_0

$$H_0 : Y_i = b_0 + \epsilon_i$$

If our regression model has K predictors, the "alternative model" is described using the usual formula for a multiple regression model:

$$H_1 : Y_i = \left(\sum_{k=1}^K b_k X_{ik} \right) + b_0 + \epsilon_i$$

How can we test these two hypotheses against each other? The trick is to understand that just like we did with ANOVA, it's possible to divide up the total variance SS_{tot} into the sum of the residual variance SS_{res} and the regression model variance SS_{mod} . I'll skip over the technicalities, since we will cover most of them in a future ANOVA chapter, and just note that the sum of squares for the model is equal to the total sum of squares minus sum of squares for the residual:

$$SS_{mod} = SS_{tot} - SS_{res}$$

And, just like we will do with the ANOVA, we can convert the sums of squares

in to mean squares by dividing by the degrees of freedom.

$$\text{MS}_{mod} = \frac{\text{SS}_{mod}}{df_{mod}}$$

$$\text{MS}_{res} = \frac{\text{SS}_{res}}{df_{res}}$$

So, how many degrees of freedom do we have? As you might expect, the df associated with the model is closely tied to the number of predictors that we've included. In fact, it turns out that $df_{mod} = K$. For the residuals, the total degrees of freedom is $df_{res} = N - K - 1$.

Now that we've got our mean square values, you're probably going to be entirely unsurprised to discover that we can calculate an F -statistic like this:

$$F = \frac{\text{MS}_{mod}}{\text{MS}_{res}}$$

and the degrees of freedom associated with this are K and $N - K - 1$. This F statistic has exactly the same interpretation as for ANOVA. Large F values indicate that the null hypothesis is performing poorly in comparison to the alternative hypothesis. And since we already did some tedious “do it the long way” calculations back then, I won't waste your time repeating them. In a moment I'll show you how to do the test in R the easy way, but first, let's have a look at the tests for the individual regression coefficients.

8.9.2 Tests for individual coefficients

The F -test that we've just introduced is useful for checking that the model as a whole is performing better than chance. This is important: if your regression model doesn't produce a significant result for the F -test then you probably don't have a very good regression model (or, quite possibly, you don't have very good data). However, while failing this test is a pretty strong indicator that the model has problems, *passing* the test (i.e., rejecting the null) doesn't imply that the model is good! Why is that, you might be wondering? The answer to that can be found by looking at the coefficients for the `regression.2` model:

```
print( regression.2 )

##
## Call:
## lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)    dan.sleep    baby.sleep
## 125.96557     -8.95025      0.01052
```

I can't help but notice that the estimated regression coefficient for the `baby.sleep` variable is tiny (0.01), relative to the value that we get for `dan.sleep` (-8.95). Given that these two variables are absolutely on the same scale (they're both measured in "hours slept"), I find this suspicious. In fact, I'm beginning to suspect that it's really only the amount of sleep that *I* get that matters in order to predict my grumpiness.

Once again, we can reuse a hypothesis test that we discussed earlier, this time the *t*-test. The test that we're interested has a null hypothesis that the true regression coefficient is zero ($b = 0$), which is to be tested against the alternative hypothesis that it isn't ($b \neq 0$). That is:

$$\begin{aligned} H_0 : & b = 0 \\ H_1 : & b \neq 0 \end{aligned}$$

How can we test this? Well, if the central limit theorem is kind to us, we might be able to guess that the sampling distribution of \hat{b} , the estimated regression coefficient, is a normal distribution with mean centred on b . What that would mean is that if the null hypothesis were true, then the sampling distribution of \hat{b} has mean zero and unknown standard deviation. Assuming that we can come up with a good estimate for the standard error of the regression coefficient, $\text{SE}(\hat{b})$, then we're in luck. That's *exactly* the situation for which we introduced the one-sample *t* way back in Chapter ???. So let's define a *t*-statistic like this,

$$t = \frac{\hat{b}}{\text{SE}(\hat{b})}$$

I'll skip over the reasons why, but our degrees of freedom in this case are $df = N - K - 1$. Irritatingly, the estimate of the standard error of the regression coefficient, $\text{SE}(\hat{b})$, is not as easy to calculate as the standard error of the mean that we used for the simpler *t*-tests and *z*-tests. In fact, the formula is somewhat ugly, and not terribly helpful to look at. For our purposes it's sufficient to point out that the standard error of the estimated regression coefficient depends on both the predictor and outcome variables, and is somewhat sensitive to violations of the homogeneity of variance assumption (discussed shortly).

In any case, this *t*-statistic can be interpreted in the same way as any *t*-statistic. Assuming that you have a two-sided alternative (i.e., you don't really care if $b > 0$ or $b < 0$), then it's the extreme values of *t* (i.e., a lot less than zero or a lot greater than zero) that suggest that you should reject the null hypothesis.

8.9.3 Running the hypothesis tests in R

To compute all of the quantities that we have talked about so far, all you need to do is ask for a `summary()` of your regression model. Since I've been using `regression.2` as my example, let's do that:

```
summary( regression.2 )

##
## Call:
## lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0345  -2.2198  -0.4016   2.6775  11.7496
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 125.96557    3.04095 41.423 <2e-16 ***
## dan.sleep    -8.95025    0.55346 -16.172 <2e-16 ***
## baby.sleep     0.01052    0.27106   0.039    0.969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.354 on 97 degrees of freedom
## Multiple R-squared:  0.8161, Adjusted R-squared:  0.8123
## F-statistic: 215.2 on 2 and 97 DF,  p-value: < 2.2e-16
```

The output that this command produces is pretty dense, but we've already discussed everything of interest in it, so what I'll do is go through it line by line. The first line reminds us of what the actual regression model is:

```
Call:
lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)
```

You can see why this is handy, since it was a little while back when we actually created the `regression.2` model, and so it's nice to be reminded of what it was we were doing. The next part provides a quick summary of the residuals (i.e., the ϵ_i values),

```
Residuals:
      Min       1Q   Median       3Q      Max
-11.0345  -2.2198  -0.4016   2.6775  11.7496
```

which can be convenient as a quick and dirty check that the model is okay. Remember, we did assume that these residuals were normally distributed, with mean 0. In particular it's worth quickly checking to see if the median is close to zero, and to see if the first quartile is about the same size as the third quartile. If they look badly off, there's a good chance that the assumptions of regression are

violated. These ones look pretty nice to me, so let's move on to the interesting stuff. The next part of the R output looks at the coefficients of the regression model:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	125.96557	3.04095	41.423	<2e-16 ***
dan.sleep	-8.95025	0.55346	-16.172	<2e-16 ***
baby.sleep	0.01052	0.27106	0.039	0.969
<hr/>				
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Each row in this table refers to one of the coefficients in the regression model. The first row is the intercept term, and the later ones look at each of the predictors. The columns give you all of the relevant information. The first column is the actual estimate of b (e.g., 125.96 for the intercept, and -8.9 for the `dan.sleep` predictor). The second column is the standard error estimate $\hat{\sigma}_b$. The third column gives you the t -statistic, and it's worth noticing that in this table $t = \hat{b}/\text{SE}(\hat{b})$ every time. Finally, the fourth column gives you the actual p value for each of these tests.¹¹ The only thing that the table itself doesn't list is the degrees of freedom used in the t -test, which is always $N - K - 1$ and is listed immediately below, in this line:

Residual standard error: 4.354 on 97 degrees of freedom

The value of $df = 97$ is equal to $N - K - 1$, so that's what we use for our t -tests. In the final part of the output we have the F -test and the R^2 values which assess the performance of the model as a whole

Residual standard error: 4.354 on 97 degrees of freedom
 Multiple R-squared: 0.8161, Adjusted R-squared: 0.8123
 F-statistic: 215.2 on 2 and 97 DF, p-value: < 2.2e-16

So in this case, the model performs significantly better than you'd expect by chance ($F(2, 97) = 215.2$, $p < .001$), which isn't all that surprising: the $R^2 = .812$ value indicate that the regression model accounts for 81.2% of the variability in the outcome measure. However, when we look back up at the t -tests for each of the individual coefficients, we lack evidence that the `baby.sleep` variable has a significant effect; all the work could be done by the `dan.sleep` variable. Taken together, these results suggest that `regression.2` is actually the wrong model for the data: you'd probably be better off dropping the `baby.sleep` predictor entirely. In other words, the `regression.1` model that we started with is the better model.

¹¹Note that, although R has done multiple tests here, it hasn't done a Bonferroni correction or anything. These are standard one-sample t -tests with a two-sided alternative. If you want to make corrections for multiple tests, you need to do that yourself.

8.10 Testing the significance of a correlation

8.10.1 Hypothesis tests for a single correlation

I don't want to spend too much time on this, but it's worth very briefly returning to the point I made earlier, that Pearson correlations are basically the same thing as linear regressions with only a single predictor added to the model. What this means is that the hypothesis tests that I just described in a regression context can also be applied to correlation coefficients. To see this, let's take a `summary()` of the `regression.1` model:

```
summary( regression.1 )

##
## Call:
## lm(formula = dan.grump ~ dan.sleep, data = parenthood)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -11.025 -2.213 -0.399  2.681 11.750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 125.9563    3.0161   41.76 <2e-16 ***
## dan.sleep    -8.9368    0.4285  -20.85 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.332 on 98 degrees of freedom
## Multiple R-squared:  0.8161, Adjusted R-squared:  0.8142
## F-statistic: 434.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

The important thing to note here is the t test associated with the predictor, in which we get a result of $t(98) = -20.85$, $p < .001$. Now let's compare this to the output of a different function, which goes by the name of `cor.test()`. As you might expect, this function runs a hypothesis test to see if the observed correlation between two variables is significantly different from 0. Let's have a look:

```
cor.test( x = parenthood$dan.sleep, y = parenthood$dan.grump )

##
## Pearson's product-moment correlation
##
```

```
## data: parenthood$dan.sleep and parenthood$dan.grump
## t = -20.854, df = 98, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.9340614 -0.8594714
## sample estimates:
##       cor
## -0.903384
```

Again, the key thing to note is the line that reports the hypothesis test itself, which seems to be saying that $t(98) = -20.85$, $p < .001$. Hm. Looks like it's exactly the same test, doesn't it? And that's exactly what it is. The test for the significance of a correlation is identical to the t test that we run on a coefficient in a regression model.

8.10.2 Hypothesis tests for all pairwise correlations

Okay, one more digression before I return to regression properly. In the previous section I talked about the `cor.test()` function, which lets you run a hypothesis test on a single correlation. The `cor.test()` function is (obviously) an extension of the `cor()` function, which we talked about in Section 8.4. However, the `cor()` function isn't restricted to computing a single correlation: you can use it to compute *all* pairwise correlations among the variables in your data set. This leads people to the natural question: can the `cor.test()` function do the same thing? Can we use `cor.test()` to run hypothesis tests for all possible pairwise correlations among the variables in a data frame?

The answer is no, and there's a very good reason for this. Testing a single correlation is fine: if you've got some reason to be asking "is A related to B?", then you should absolutely run a test to see if there's a significant correlation. But if you've got variables A, B, C, D and E and you're thinking about testing the correlations among all possible pairs of these, a statistician would want to ask: what's your hypothesis? If you're in the position of wanting to test all possible pairs of variables, then you're pretty clearly on a fishing expedition, hunting around in search of significant effects when you don't actually have a clear research hypothesis in mind. This is *dangerous*, and the authors of `cor.test()` obviously felt that they didn't want to support that kind of behaviour.

On the other hand... a somewhat less hardline view might be to argue we encounter this situation when we talk about *post hoc tests* in ANOVA. When running post hoc tests, we didn't have any specific comparisons in mind, so what we do is apply a correction (e.g., Bonferroni, Holm, etc) in order to avoid the possibility of an inflated Type I error rate. From this perspective, it's okay to run hypothesis tests on all your pairwise correlations, but you must treat them as post hoc analyses, and if so you need to apply a correction for multiple comparisons. That's what the `correlate()` function in the `lsr` package

does. When we use the `correlate()` function in Section 8.4 all it did was print out the correlation matrix. But you can get it to output the results of all the pairwise tests as well by specifying `test=TRUE`. Here's what happens with the `parenthood` data:

```
library(lsr)
correlate(parenthood, test=TRUE)

##
## CORRELATIONS
## =====
## - correlation type: pearson
## - correlations shown only when both variables are numeric
##
##      dan.sleep   baby.sleep   dan.grump     day
## dan.sleep       .        0.628***    -0.903*** -0.098
## baby.sleep     0.628***       .        -0.566*** -0.010
## dan.grump     -0.903***    -0.566***       .        0.076
## day           -0.098       -0.010        0.076       .
##
## ---
## Signif. codes: . = p < .1, * = p<.05, ** = p<.01, *** = p<.001
##
##
## p-VALUES
## =====
## - total number of tests run: 6
## - correction for multiple testing: holm
##
##      dan.sleep   baby.sleep   dan.grump     day
## dan.sleep       .        0.000        0.000  0.990
## baby.sleep     0.000       .        0.000  0.990
## dan.grump     0.000       0.000       .        0.990
## day            0.990       0.990        0.990       .
##
##
## SAMPLE SIZES
## =====
##      dan.sleep   baby.sleep   dan.grump     day
## dan.sleep      100        100        100  100
## baby.sleep     100        100        100  100
## dan.grump     100        100        100  100
## day            100        100        100  100
```

The output here contains three matrices. First it prints out the correlation

matrix. Second it prints out a matrix of p -values, using the Holm method¹² to correct for multiple comparisons. Finally, it prints out a matrix indicating the sample size (number of pairwise complete cases) that contributed to each correlation.

So there you have it. If you really desperately want to do pairwise hypothesis tests on your correlations, the `correlate()` function will let you do it. But please, **please** be careful. I can't count the number of times I've had a student panicking in my office because they've run these pairwise correlation tests, and they get one or two significant results that don't make any sense. For some reason, the moment people see those little significance stars appear, they feel compelled to throw away all common sense and assume that the results must correspond to something real that requires an explanation. In most such cases, my experience has been that the right answer is "it's a Type I error".

8.11 Regarding regression coefficients

Before moving on to discuss the assumptions underlying linear regression and what you can do to check if they're being met, there's two more topics I want to briefly discuss, both of which relate to the regression coefficients. The first thing to talk about is calculating confidence intervals for the coefficients; after that, I'll discuss the somewhat murky question of how to determine which of predictor is most important.

8.11.1 Confidence intervals for the coefficients

Like any population parameter, the regression coefficients b cannot be estimated with complete precision from a sample of data; that's part of why we need hypothesis tests. Given this, it's quite useful to be able to report confidence intervals that capture our uncertainty about the true value of b . This is especially useful when the research question focuses heavily on an attempt to find out *how* strongly variable X is related to variable Y , since in those situations the interest is primarily in the regression weight b . Fortunately, confidence intervals for the regression weights can be constructed in the usual fashion,

$$\text{CI}(b) = \hat{b} \pm (t_{crit} \times \text{SE}(\hat{b}))$$

where $\text{SE}(\hat{b})$ is the standard error of the regression coefficient, and t_{crit} is the relevant critical value of the appropriate t distribution. For instance, if it's a 95% confidence interval that we want, then the critical value is the 97.5th quantile of a t distribution with $N - K - 1$ degrees of freedom. In other words, this is basically the same approach to calculating confidence intervals that we've

¹²You can change the kind of correction it applies by specifying the `p.adjust.method` argument.

used throughout. To do this in R we can use the `confint()` function. There arguments to this function are

- `object`. The regression model (`lm` object) for which confidence intervals are required.
- `parm`. A vector indicating which coefficients we should calculate intervals for. This can be either a vector of numbers or (more usefully) a character vector containing variable names. By default, all coefficients are included, so usually you don't bother specifying this argument.
- `level`. A number indicating the confidence level that should be used. As is usually the case, the default value is 0.95, so you wouldn't usually need to specify this argument.

So, suppose I want 99% confidence intervals for the coefficients in the `regression.2` model. I could do this using the following command:

```
confint( object = regression.2,
         level = .99)

##           0.5 %      99.5 %
## (Intercept) 117.9755724 133.9555593
## dan.sleep   -10.4044419  -7.4960575
## baby.sleep   -0.7016868   0.7227357
```

Simple enough.

8.11.2 Calculating standardised regression coefficients

One more thing that you might want to do is to calculate “standardised” regression coefficients, often denoted β . The rationale behind standardised coefficients goes like this. In a lot of situations, your variables are on fundamentally different scales. Suppose, for example, my regression model aims to predict people’s IQ scores, using their educational attainment (number of years of education) and their income as predictors. Obviously, educational attainment and income are not on the same scales: the number of years of schooling can only vary by 10s of years, whereas income would vary by 10,000s of dollars (or more). The units of measurement have a big influence on the regression coefficients: the b coefficients only make sense when interpreted in light of the units, both of the predictor variables and the outcome variable. This makes it very difficult to compare the coefficients of different predictors. Yet there are situations where you really do want to make comparisons between different coefficients. Specifically, you might want some kind of standard measure of which predictors have the strongest relationship to the outcome. This is what ***standardised coefficients*** aim to do.

The basic idea is quite simple: the standardised coefficients are the coefficients that you would have obtained if you'd converted all the variables to z -scores before running the regression.¹³ The idea here is that, by converting all the predictors to z -scores, they all go into the regression on the same scale, thereby removing the problem of having variables on different scales. Regardless of what the original variables were, a β value of 1 means that an increase in the predictor of 1 standard deviation will produce a corresponding 1 standard deviation increase in the outcome variable. Therefore, if variable A has a larger absolute value of β than variable B, it is deemed to have a stronger relationship with the outcome. Or at least that's the idea: it's worth being a little cautious here, since this does rely very heavily on the assumption that "a 1 standard deviation change" is fundamentally the same kind of thing for all variables. It's not always obvious that this is true.

Leaving aside the interpretation issues, let's look at how it's calculated. What you could do is standardise all the variables yourself and then run a regression, but there's a much simpler way to do it. As it turns out, the β coefficient for a predictor X and outcome Y has a very simple formula, namely

$$\beta_X = b_X \times \frac{\sigma_X}{\sigma_Y}$$

where σ_X is the standard deviation of the predictor, and σ_Y is the standard deviation of the outcome variable Y . This makes matters a lot simpler. To make things even simpler, the **lsr** package includes a function **standardCoefs()** that computes the β coefficients.

```
standardCoefs( regression.2 )
```

```
##                   b      beta
## dan.sleep -8.95024973 -0.90474809
## baby.sleep  0.01052447  0.00217223
```

This clearly shows that the **dan.sleep** variable has a much stronger effect than the **baby.sleep** variable. However, this is a perfect example of a situation where it would probably make sense to use the original coefficients b rather than the standardised coefficients β . After all, my sleep and the baby's sleep are *already* on the same scale: number of hours slept. Why complicate matters by converting these to z -scores?

¹³Strictly, you standardise all the *regressors*: that is, every "thing" that has a regression coefficient associated with it in the model. For the regression models that I've talked about so far, each predictor variable maps onto exactly one regressor, and vice versa. However, that's not actually true in general: we'll see some examples of this in Chapter ???. But for now, we don't need to care too much about this distinction.

8.12 Assumptions of regression

The linear regression model that I've been discussing relies on several assumptions. In Section 8.13 we'll talk a lot more about how to check that these assumptions are being met, but first, let's have a look at each of them.

- *Normality.* Like half the models in statistics, standard linear regression relies on an assumption of normality. Specifically, it assumes that the *residuals* are normally distributed. It's actually okay if the predictors X and the outcome Y are non-normal, so long as the residuals ϵ are normal. See Section 8.13.3.
- *Linearity.* A pretty fundamental assumption of the linear regression model is that relationship between X and Y actually be linear! Regardless of whether it's a simple regression or a multiple regression, we assume that the relationships involved are linear. See Section 8.13.4.
- *Homogeneity of variance.* Strictly speaking, the regression model assumes that each residual ϵ_i is generated from a normal distribution with mean 0, and (more importantly for the current purposes) with a standard deviation σ that is the same for every single residual. In practice, it's impossible to test the assumption that every residual is identically distributed. Instead, what we care about is that the standard deviation of the residual is the same for all values of \hat{Y} , and (if we're being especially paranoid) all values of every predictor X in the model. See Section 8.13.5.
- *Uncorrelated predictors.* The idea here is that, in a multiple regression model, you don't want your predictors to be too strongly correlated with each other. This isn't "technically" an assumption of the regression model, but in practice it's required. Predictors that are too strongly correlated with each other (referred to as "collinearity") can cause problems when evaluating the model. See Section 8.13.6.
- *Residuals are independent of each other.* This is really just a "catch all" assumption, to the effect that "there's nothing else funny going on in the residuals". If there is something weird (e.g., the residuals all depend heavily on some other unmeasured variable) going on, it might screw things up.
- *No "bad" outliers.* Again, not actually a technical assumption of the model (or rather, it's sort of implied by all the others), but there is an implicit assumption that your regression model isn't being too strongly influenced by one or two anomalous data points; since this raises questions about the adequacy of the model, and the trustworthiness of the data in some cases. See Section 8.13.2.

8.13 Model checking

The main focus of this section is *regression diagnostics*, a term that refers to the art of checking that the assumptions of your regression model have been met, figuring out how to fix the model if the assumptions are violated, and generally to check that nothing “funny” is going on. I refer to this as the “art” of model checking with good reason: it’s not easy, and while there are a lot of fairly standardised tools that you can use to diagnose and maybe even cure the problems that ail your model (if there are any, that is!), you really do need to exercise a certain amount of judgment when doing this. It’s easy to get lost in all the details of checking this thing or that thing, and it’s quite exhausting to try to remember what all the different things are. This has the very nasty side effect that a lot of people get frustrated when trying to learn *all* the tools, so instead they decide not to do *any* model checking. This is a bit of a worry!

In this section, I describe several different things you can do to check that your regression model is doing what it’s supposed to. It doesn’t cover the full space of things you could do, but it’s still much more detailed than what I see a lot of people doing in practice; and I don’t usually cover all of this in my intro stats class myself. However, I do think it’s important that you get a sense of what tools are at your disposal, so I’ll try to introduce a bunch of them here. Finally, I should note that this section draws quite heavily from the Fox and Weisberg (2011) text, the book associated with the `car` package. The `car` package is notable for providing some excellent tools for regression diagnostics, and the book itself talks about them in an admirably clear fashion. I don’t want to sound too gushy about it, but I do think that Fox and Weisberg (2011) is well worth reading.

8.13.1 Three kinds of residuals

The majority of regression diagnostics revolve around looking at the residuals, and by now you’ve probably formed a sufficiently pessimistic theory of statistics to be able to guess that – precisely *because* of the fact that we care a lot about the residuals – there are several different kinds of residual that we might consider. In particular, the following three kinds of residual are referred to in this section: “ordinary residuals”, “standardised residuals”, and “Studentised residuals”. There is a fourth kind that you’ll see referred to in some of the Figures, and that’s the “Pearson residual”: however, for the models that we’re talking about in this chapter, the Pearson residual is identical to the ordinary residual.

The first and simplest kind of residuals that we care about are *ordinary residuals*. These are the actual, raw residuals that I’ve been talking about throughout this chapter. The ordinary residual is just the difference between the fitted value \hat{Y}_i and the observed value Y_i . I’ve been using the notation ϵ_i to refer to the i -th ordinary residual, and by gum I’m going to stick to it. With this in

mind, we have the very simple equation

$$\epsilon_i = Y_i - \hat{Y}_i$$

This is of course what we saw earlier, and unless I specifically refer to some other kind of residual, this is the one I'm talking about. So there's nothing new here: I just wanted to repeat myself. In any case, you can get R to output a vector of ordinary residuals, you can use a command like this:

```
residuals( object = regression.2 )
```

```
##          1          2          3          4          5          6
## -2.1403095  4.7081942  1.9553640 -2.0602806  0.7194888 -0.4066133
##          7          8          9         10         11         12
##  0.2269987 -1.7003077  0.2025039  3.8524589  3.9986291 -4.9120150
##         13         14         15         16         17         18
##  1.2060134  0.4946578 -2.6579276 -0.3966805  3.3538613  1.7261225
##         19         20         21         22         23         24
## -0.4922551 -5.6405941 -0.4660764  2.7238389  9.3653697  0.2841513
##         25         26         27         28         29         30
## -0.5037668 -1.4941146  8.1328623  1.9787316 -1.5126726  3.5171148
##         31         32         33         34         35         36
## -8.9256951 -2.8282946  6.1030349 -7.5460717  4.5572128 -10.6510836
##         37         38         39         40         41         42
## -5.6931846  6.3096506 -2.1082466 -0.5044253  0.1875576  4.8094841
##         43         44         45         46         47         48
## -5.4135163 -6.2292842 -4.5725232 -5.3354601  3.9950111  2.1718745
##         49         50         51         52         53         54
## -3.4766440  0.4834367  6.2839790  2.0109396 -1.5846631 -2.2166613
##         55         56         57         58         59         60
##  2.2033140  1.9328736 -1.8301204 -1.5401430  2.5298509 -3.3705782
##         61         62         63         64         65         66
## -2.9380806  0.6590736 -0.5917559 -8.6131971  5.9781035  5.9332979
##         67         68         69         70         71         72
## -1.2341956  3.0047669 -1.0802468  6.5174672 -3.0155469  2.1176720
##         73         74         75         76         77         78
##  0.6058757 -2.7237421 -2.2291472 -1.4053822  4.7461491 11.7495569
##         79         80         81         82         83         84
##  4.7634141  2.6620908 -11.0345292 -0.7588667  1.4558227 -0.4745727
##         85         86         87         88         89         90
##  8.9091201 -1.1409777  0.7555223 -0.4107130  0.8797237 -1.4095586
##         91         92         93         94         95         96
##  3.1571385 -3.4205757 -5.7228699 -2.2033958 -3.8647891  0.4982711
##         97         98         99        100
## -5.5249495  4.1134221 -8.2038533  5.6800859
```

One drawback to using ordinary residuals is that they're always on a different scale, depending on what the outcome variable is and how good the regression model is. That is, Unless you've decided to run a regression model without an intercept term, the ordinary residuals will have mean 0; but the variance is different for every regression. In a lot of contexts, especially where you're only interested in the *pattern* of the residuals and not their actual values, it's convenient to estimate the ***standardised residuals***, which are normalised in such a way as to have standard deviation 1. The way we calculate these is to divide the ordinary residual by an estimate of the (population) standard deviation of these residuals. For technical reasons, mumble mumble, the formula for this is:

$$\epsilon'_i = \frac{\epsilon_i}{\hat{\sigma} \sqrt{1 - h_i}}$$

where $\hat{\sigma}$ in this context is the estimated population standard deviation of the ordinary residuals, and h_i is the “hat value” of the i th observation. I haven't explained hat values to you yet (but have no fear,¹⁴ it's coming shortly), so this won't make a lot of sense. For now, it's enough to interpret the standardised residuals as if we'd converted the ordinary residuals to z -scores. In fact, that is more or less the truth, it's just that we're being a bit fancier. To get the standardised residuals, the command you want is this:

```
rstandard( model = regression.2 )
```

```
##          1          2          3          4          5          6
## -0.49675845  1.10430571  0.46361264 -0.47725357  0.16756281 -0.09488969
##          7          8          9         10         11         12
##  0.05286626 -0.39260381  0.04739691  0.89033990  0.95851248 -1.13898701
##         13         14         15         16         17         18
##  0.28047841  0.11519184 -0.61657092 -0.09191865  0.77692937  0.40403495
##         19         20         21         22         23         24
## -0.11552373 -1.31540412 -0.10819238  0.62951824  2.17129803  0.06586227
##         25         26         27         28         29         30
## -0.11980449 -0.34704024  1.91121833  0.45686516 -0.34986350  0.81233165
##         31         32         33         34         35         36
## -2.08659993 -0.66317843  1.42930082 -1.77763064  1.07452436 -2.47385780
##         37         38         39         40         41         42
## -1.32715114  1.49419658 -0.49115639 -0.11674947  0.04401233  1.11881912
##         43         44         45         46         47         48
## -1.27081641 -1.46422595 -1.06943700 -1.24659673  0.94152881  0.51069809
##         49         50         51         52         53         54
## -0.81373349  0.11412178  1.47938594  0.46437962 -0.37157009 -0.51609949
##         55         56         57         58         59         60
##  0.51800753  0.44813204 -0.42662358 -0.35575611  0.58403297 -0.78022677
##         61         62         63         64         65         66
```

¹⁴Or have no hope, as the case may be.

```

## -0.67833325  0.15484699 -0.13760574 -2.05662232  1.40238029  1.37505125
##      67          68          69          70          71          72
## -0.28964989  0.69497632 -0.24945316  1.50709623 -0.69864682  0.49071427
##      73          74          75          76          77          78
##  0.14267297 -0.63246560 -0.51972828 -0.32509811  1.10842574  2.72171671
##      79          80          81          82          83          84
##  1.09975101  0.62057080 -2.55172097 -0.17584803  0.34340064 -0.11158952
##      85          86          87          88          89          90
##  2.10863391 -0.26386516  0.17624445 -0.09504416  0.20450884 -0.32730740
##      91          92          93          94          95          96
##  0.73475640 -0.79400855 -1.32768248 -0.51940736 -0.91512580  0.11661226
##      97          98          99         100
## -1.28069115  0.96332849 -1.90290258  1.31368144

```

Note that this function uses a different name for the input argument, but it's still just a linear regression object that the function wants to take as its input here.

The third kind of residuals are ***Studentised residuals*** (also called “jackknifed residuals”) and they're even fancier than standardised residuals. Again, the idea is to take the ordinary residual and divide it by some quantity in order to estimate some standardised notion of the residual, but the formula for doing the calculations this time is subtly different:

$$\epsilon_i^* = \frac{\epsilon_i}{\hat{\sigma}_{(-i)} \sqrt{1 - h_i}}$$

Notice that our estimate of the standard deviation here is written $\hat{\sigma}_{(-i)}$. What this corresponds to is the estimate of the residual standard deviation that you *would have obtained*, if you just deleted the i th observation from the data set. This sounds like the sort of thing that would be a nightmare to calculate, since it seems to be saying that you have to run N new regression models (even a modern computer might grumble a bit at that, especially if you've got a large data set). Fortunately, some terribly clever person has shown that this standard deviation estimate is actually given by the following equation:

$$\hat{\sigma}_{(-i)} = \hat{\sigma} \sqrt{\frac{N - K - 1 - \epsilon_i'^2}{N - K - 2}}$$

Isn't that a pip? Anyway, the command that you would use if you wanted to pull out the Studentised residuals for our regression model is

```

rstudent( model = regression.2 )

##           1           2           3           4           5           6
## -0.49482102  1.10557030  0.46172854 -0.47534555  0.16672097 -0.09440368

```

```

##      7       8       9      10      11      12
## 0.05259381 -0.39088553 0.04715251 0.88938019 0.95810710 -1.14075472
## 13       14       15       16       17       18
## 0.27914212 0.11460437 -0.61459001 -0.09144760 0.77533036 0.40228555
## 19       20       21       22       23       24
## -0.11493461 -1.32043609 -0.10763974 0.62754813 2.21456485 0.06552336
## 25       26       27       28       29       30
## -0.11919416 -0.34546127 1.93818473 0.45499388 -0.34827522 0.81089646
## 31       32       33       34       35       36
## -2.12403286 -0.66125192 1.43712830 -1.79797263 1.07539064 -2.54258876
## 37       38       39       40       41       42
## -1.33244515 1.50388257 -0.48922682 -0.11615428 0.04378531 1.12028904
## 43       44       45       46       47       48
## -1.27490649 -1.47302872 -1.07023828 -1.25020935 0.94097261 0.50874322
## 49       50       51       52       53       54
## -0.81230544 0.11353962 1.48863006 0.46249410 -0.36991317 -0.51413868
## 55       56       57       58       59       60
## 0.51604474 0.44627831 -0.42481754 -0.35414868 0.58203894 -0.77864171
## 61       62       63       64       65       66
## -0.67643392 0.15406579 -0.13690795 -2.09211556 1.40949469 1.38147541
## 67       68       69       70       71       72
## -0.28827768 0.69311245 -0.24824363 1.51717578 -0.69679156 0.48878534
## 73       74       75       76       77       78
## 0.14195054 -0.63049841 -0.51776374 -0.32359434 1.10974786 2.81736616
## 79       80       81       82       83       84
## 1.10095270 0.61859288 -2.62827967 -0.17496714 0.34183379 -0.11101996
## 85       86       87       88       89       90
## 2.14753375 -0.26259576 0.17536170 -0.09455738 0.20349582 -0.32579584
## 91       92       93       94       95       96
## 0.73300184 -0.79248469 -1.33298848 -0.51744314 -0.91435205 0.11601774
## 97       98       99       100
## -1.28498273 0.96296745 -1.92942389 1.31867548

```

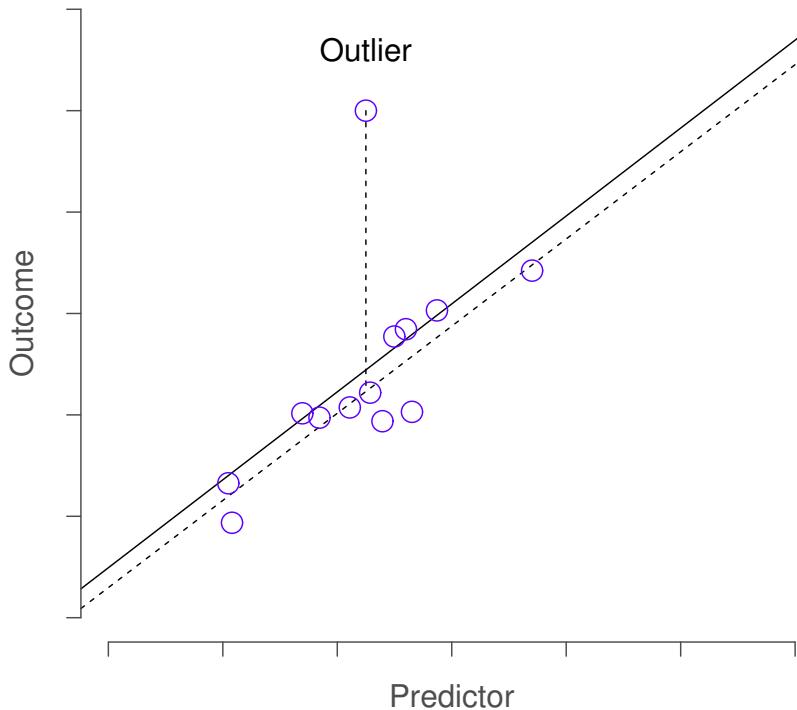
Before moving on, I should point out that you don't often need to manually extract these residuals yourself, even though they are at the heart of almost all regression diagnostics. That is, the `residuals()`, `rstandard()` and `rstudent()` functions are all useful to *know* about, but most of the time the various functions that run the diagnostics will take care of these calculations for you. Even so, it's always nice to know how to actually get hold of these things yourself in case you ever need to do something non-standard.

8.13.2 Three kinds of anomalous data

One danger that you can run into with linear regression models is that your analysis might be disproportionately sensitive to a smallish number of "unusual"

or “anomalous” observations. I discussed this idea previously in Section 3.9.3.2 in the context of discussing the outliers that get automatically identified by the `boxplot()` function, but this time we need to be much more precise. In the context of linear regression, there are three conceptually distinct ways in which an observation might be called “anomalous”. All three are interesting, but they have rather different implications for your analysis.

The first kind of unusual observation is an ***outlier***. The definition of an outlier (in this context) is an observation that is very different from what the regression model predicts. An example is shown in Figure ???. In practice, we operationalise this concept by saying that an outlier is an observation that has a very large Studentised residual, ϵ_i^* . Outliers are interesting: a big outlier *might* correspond to junk data – e.g., the variables might have been entered incorrectly, or some other defect may be detectable. Note that you shouldn’t throw an observation away just because it’s an outlier. But the fact that it’s an outlier is often a cue to look more closely at that case, and try to find out why it’s so different.



The second way in which an observation can be unusual is if it has high ***leverage***.

age: this happens when the observation is very different from all the other observations. This doesn't necessarily have to correspond to a large residual: if the observation happens to be unusual on all variables in precisely the same way, it can actually lie very close to the regression line. An example of this is shown in Figure 8.13. The leverage of an observation is operationalised in terms of its *hat value*, usually written h_i . The formula for the hat value is rather complicated¹⁵ but its interpretation is not: h_i is a measure of the extent to which the i -th observation is “in control” of where the regression line ends up going. You can extract the hat values using the following command:

```
hatvalues( model = regression.2 )
```

```
##      1       2       3       4       5       6
## 0.02067452 0.04105320 0.06155445 0.01685226 0.02734865 0.03129943
##      7       8       9      10      11      12
## 0.02735579 0.01051224 0.03698976 0.01229155 0.08189763 0.01882551
##     13      14      15      16      17      18
## 0.02462902 0.02718388 0.01964210 0.01748592 0.01691392 0.03712530
##     19      20      21      22      23      24
## 0.04213891 0.02994643 0.02099435 0.01233280 0.01853370 0.01804801
##     25      26      27      28      29      30
## 0.06722392 0.02214927 0.04472007 0.01039447 0.01381812 0.01105817
##     31      32      33      34      35      36
## 0.03468260 0.04048248 0.03814670 0.04934440 0.05107803 0.02208177
##     37      38      39      40      41      42
## 0.02919013 0.05928178 0.02799695 0.01519967 0.04195751 0.02514137
##     43      44      45      46      47      48
## 0.04267879 0.04517340 0.03558080 0.03360160 0.05019778 0.04587468
##     49      50      51      52      53      54
## 0.03701290 0.05331282 0.04814477 0.01072699 0.04047386 0.02681315
##     55      56      57      58      59      60
## 0.04556787 0.01856997 0.02919045 0.01126069 0.01012683 0.01546412
##     61      62      63      64      65      66
## 0.01029534 0.04428870 0.02438944 0.07469673 0.04135090 0.01775697
##     67      68      69      70      71      72
## 0.04217616 0.01384321 0.01069005 0.01340216 0.01716361 0.01751844
##     73      74      75      76      77      78
## 0.04863314 0.02158623 0.02951418 0.01411915 0.03276064 0.01684599
##     79      80      81      82      83      84
## 0.01028001 0.02920514 0.01348051 0.01752758 0.05184527 0.04583604
```

¹⁵Again, for the linear algebra fanatics: the “hat matrix” is defined to be that matrix H that converts the vector of observed values y into a vector of fitted values \hat{y} , such that $\hat{y} = Hy$. The name comes from the fact that this is the matrix that “puts a hat on y ”. The hat value of the i -th observation is the i -th diagonal element of this matrix (so technically I should be writing it as h_{ii} rather than h_i). Oh, and in case you care, here's how it's calculated: $H = X(X^T X)^{-1} X^T$. Pretty, isn't it?

```
##      85      86      87      88      89      90
## 0.05825858 0.01359644 0.03054414 0.01487724 0.02381348 0.02159418
##      91      92      93      94      95      96
## 0.02598661 0.02093288 0.01982480 0.05063492 0.05907629 0.03682026
##      97      98      99     100
## 0.01817919 0.03811718 0.01945603 0.01373394
```

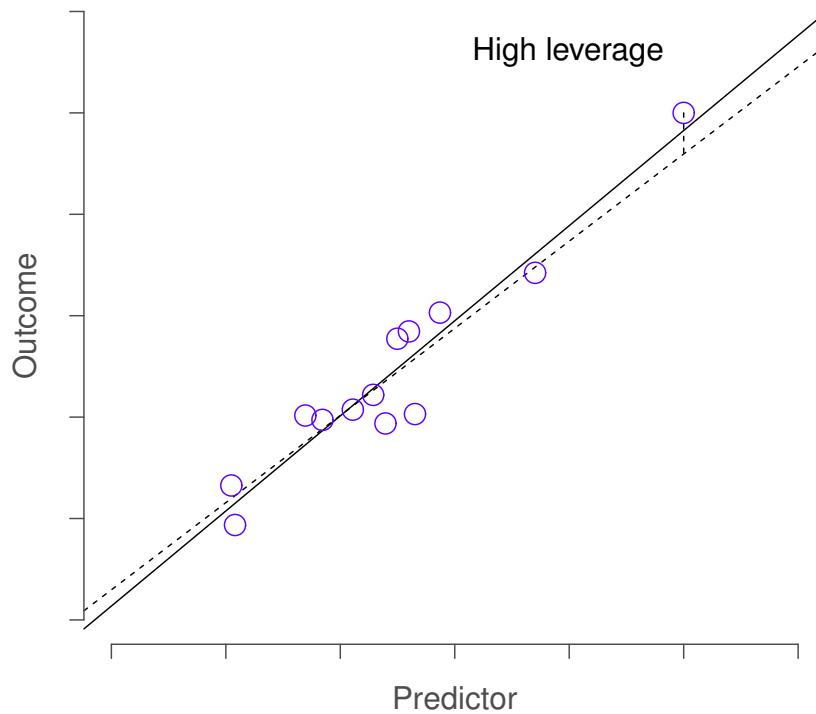


Figure 8.13: An illustration of high leverage points. The anomalous observation in this case is unusual both in terms of the predictor (x axis) and the outcome (y axis), but this unusualness is highly consistent with the pattern of correlations that exists among the other observations; as a consequence, the observation falls very close to the regression line and does not distort it.

In general, if an observation lies far away from the other ones in terms of the predictor variables, it will have a large hat value (as a rough guide, high leverage

is when the hat value is more than 2-3 times the average; and note that the sum of the hat values is constrained to be equal to $K + 1$). High leverage points are also worth looking at in more detail, but they're much less likely to be a cause for concern unless they are also outliers. % guide from Venables and Ripley.

This brings us to our third measure of unusualness, the *influence* of an observation. A high influence observation is an outlier that has high leverage. That is, it is an observation that is very different to all the other ones in some respect, and also lies a long way from the regression line. This is illustrated in Figure 8.14. Notice the contrast to the previous two figures: outliers don't move the regression line much, and neither do high leverage points. But something that is an outlier and has high leverage... that has a big effect on the regression line.

That's why we call these points high influence; and it's why they're the biggest worry. We operationalise influence in terms of a measure known as *Cook's distance*,

$$D_i = \frac{\epsilon_i^*{}^2}{K + 1} \times \frac{h_i}{1 - h_i}$$

Notice that this is a multiplication of something that measures the outlier-ness of the observation (the bit on the left), and something that measures the leverage of the observation (the bit on the right). In other words, in order to have a large Cook's distance, an observation must be a fairly substantial outlier *and* have high leverage. In a stunning turn of events, you can obtain these values using the following command:

```
cooks.distance( model = regression.2 )
```

	1	2	3	4	5
##	1.736512e-03	1.740243e-02	4.699370e-03	1.301417e-03	2.631557e-04
##	6	7	8	9	10
##	9.697585e-05	2.620181e-05	5.458491e-04	2.876269e-05	3.288277e-03
##	11	12	13	14	15
##	2.731835e-02	8.296919e-03	6.621479e-04	1.235956e-04	2.538915e-03
##	16	17	18	19	20
##	5.012283e-05	3.461742e-03	2.098055e-03	1.957050e-04	1.780519e-02
##	21	22	23	24	25
##	8.367377e-05	1.649478e-03	2.967594e-02	2.657610e-05	3.448032e-04
##	26	27	28	29	30
##	9.093379e-04	5.699951e-02	7.307943e-04	5.716998e-04	2.459564e-03
##	31	32	33	34	35
##	5.214331e-02	6.185200e-03	2.700686e-02	5.467345e-02	2.071643e-02
##	36	37	38	39	40
##	4.606378e-02	1.765312e-02	4.689817e-02	2.316122e-03	7.012530e-05
##	41	42	43	44	45
##	2.827824e-05	1.076083e-02	2.399931e-02	3.381062e-02	1.406498e-02
##	46	47	48	49	50

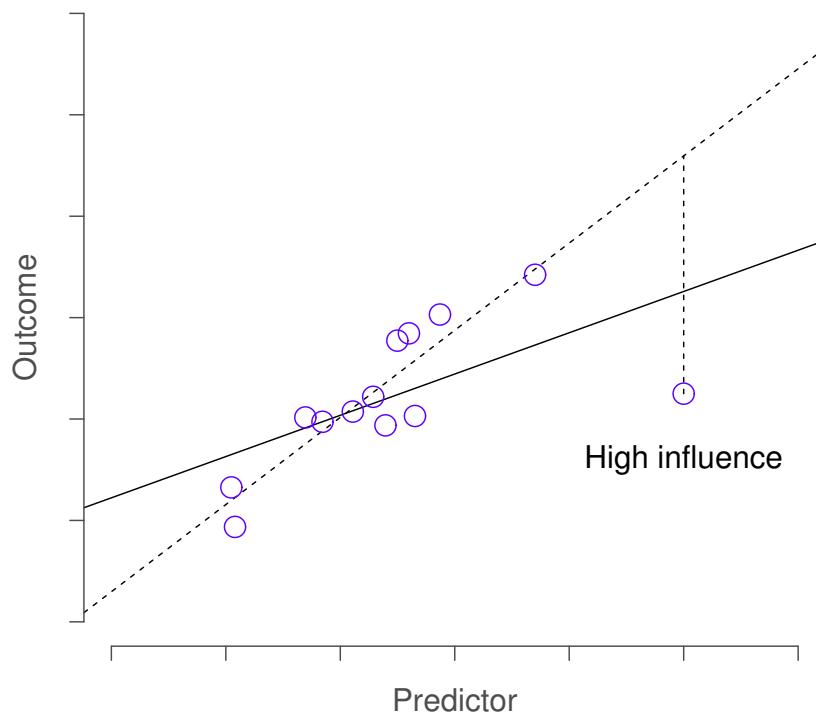


Figure 8.14: An illustration of high influence points. In this case, the anomalous observation is highly unusual on the predictor variable (x axis), and falls a long way from the regression line. As a consequence, the regression line is highly distorted, even though (in this case) the anomalous observation is entirely typical in terms of the outcome variable (y axis).

```

## 1.801086e-02 1.561699e-02 4.179986e-03 8.483514e-03 2.444787e-04
##      51          52          53          54          55
## 3.689946e-02 7.794472e-04 1.941235e-03 2.446230e-03 4.270361e-03
##      56          57          58          59          60
## 1.266609e-03 1.824212e-03 4.804705e-04 1.163181e-03 3.187235e-03
##      61          62          63          64          65
## 1.595512e-03 3.703826e-04 1.577892e-04 1.138165e-01 2.827715e-02
##      66          67          68          69          70
## 1.139374e-02 1.231422e-03 2.260006e-03 2.241322e-04 1.028479e-02
##      71          72          73          74          75
## 2.841329e-03 1.431223e-03 3.468538e-04 2.941757e-03 2.738249e-03
##      76          77          78          79          80
## 5.045357e-04 1.387108e-02 4.230966e-02 4.187440e-03 3.861831e-03
##      81          82          83          84          85
## 2.965826e-02 1.838888e-04 2.149369e-03 1.993929e-04 9.168733e-02
##      86          87          88          89          90
## 3.198994e-04 3.262192e-04 4.547383e-05 3.400893e-04 7.881487e-04
##      91          92          93          94          95
## 4.801204e-03 4.493095e-03 1.188427e-02 4.796360e-03 1.752666e-02
##      96          97          98          99         100
## 1.732793e-04 1.012302e-02 1.225818e-02 2.394964e-02 8.010508e-03

```

As a rough guide, Cook's distance greater than 1 is often considered large (that's what I typically use as a quick and dirty rule), though a quick scan of the internet and a few papers suggests that $4/N$ has also been suggested as a possible rule of thumb.

As hinted above, you don't usually need to make use of these functions, since you can have R automatically draw the critical plots.¹⁶ For the `regression.2` model, these are the plots showing Cook's distance (Figure 8.15) and the more detailed breakdown showing the scatter plot of the Studentised residual against leverage (Figure 8.16). To draw these, we can use the `plot()` function. When the main argument `x` to this function is a linear model object, it will draw one of six different plots, each of which is quite useful for doing regression diagnostics. You specify which one you want using the `which` argument (a number between 1 and 6). If you don't do this then R will draw all six. The two plots of interest to us in this context are generated using the following commands:

```
plot(x = regression.2, which = 4)
```

¹⁶Though special mention should be made of the `influenceIndexPlot()` and `influencePlot()` functions in the `car` package. These produce somewhat more detailed pictures than the default plots that I've shown here. There's also an `outlierTest()` function that tests to see if any of the Studentised residuals are significantly larger than would be expected by chance.

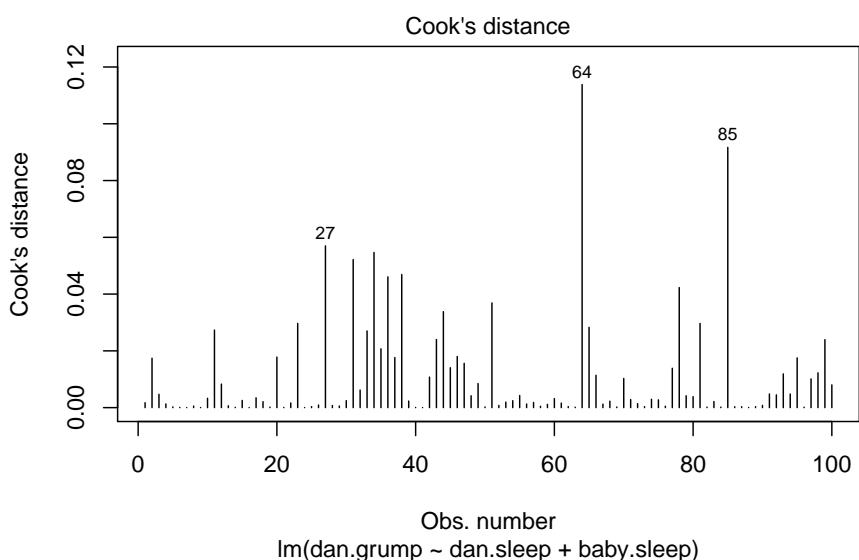


Figure 8.15: Cook's distance for every observation. This is one of the standard regression plots produced by the `plot()` function when the input is a linear regression object. It is obtained by setting `which=4`

```
plot(x = regression.2, which = 5)
```

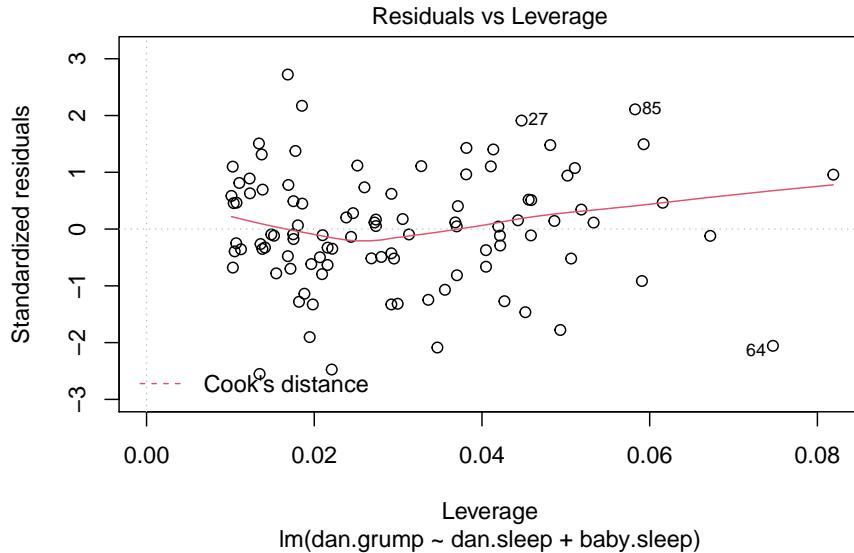


Figure 8.16: Residuals versus leverage. This is one of the standard regression plots produced by the `plot()` function when the input is a linear regression object. It is obtained by setting `which=5`.

An obvious question to ask next is, if you do have large values of Cook’s distance, what should you do? As always, there’s no hard and fast rules. Probably the first thing to do is to try running the regression with that point excluded and see what happens to the model performance and to the regression coefficients. If they really are substantially different, it’s time to start digging into your data set and your notes that you no doubt were scribbling as you ran your study; try to figure out *why* the point is so different. If you start to become convinced that this one data point is badly distorting your results, you might consider excluding it, but that’s less than ideal unless you have a solid explanation for why this particular case is qualitatively different from the others and therefore deserves to be handled separately.¹⁷ To give an example, let’s delete the observation from day 64, the observation with the largest Cook’s distance for the `regression.2` model. We can do this using the `subset` argument:

¹⁷An alternative is to run a “robust regression”; I’ll discuss robust regression in a later version of this book.

```

lm( formula = dan.grump ~ dan.sleep + baby.sleep, # same formula
    data = parenthood,          # same data frame...
    subset = -64                # ...but observation 64 is deleted
)

## 
## Call:
## lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood,
##      subset = -64)
##
## Coefficients:
## (Intercept)  dan.sleep  baby.sleep
##       126.3553     -8.8283      -0.1319

```

As you can see, those regression coefficients have barely changed in comparison to the values we got earlier. In other words, we really don't have any problem as far as anomalous data are concerned.

8.13.3 Checking the normality of the residuals

Like many of the statistical tools we've discussed in this book, regression models rely on a normality assumption. In this case, we assume that the residuals are normally distributed. Firstly, I firmly believe that it never hurts to draw an old fashioned histogram. The command I use might be something like this:

```

hist( x = residuals( regression.2 ),      # data are the residuals
      xlab = "Value of residual",        # x-axis label
      main = "",                      # no title
      breaks = 20                     # lots of breaks
)

```

The resulting plot is shown in Figure 8.17, and as you can see the plot looks pretty damn close to normal, almost unnaturally so.

I could also run a Shapiro-Wilk test to check, using the `shapiro.test()` function; the W value of .99, at this sample size, is non-significant ($p = .84$), again suggesting that the normality assumption isn't in any danger here. As a third measure, we might also want to draw a QQ-plot using the `qqnorm()` function. The QQ plot is an excellent one to draw, and so you might not be surprised to discover that it's one of the regression plots that we can produce using the `plot()` function:

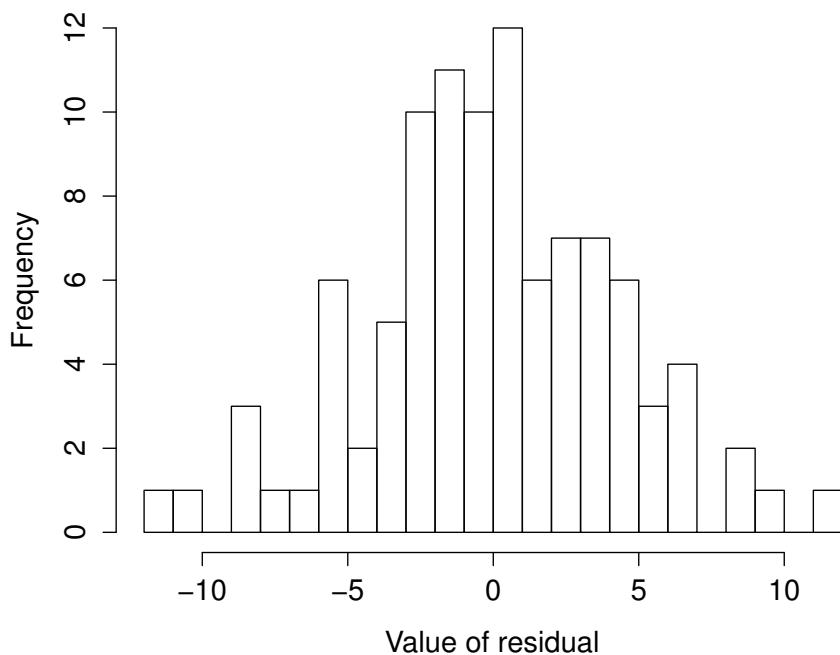


Figure 8.17: A histogram of the (ordinary) residuals in the `regression.2` model. These residuals look very close to being normally distributed, much moreso than is typically seen with real data. This shouldn't surprise you... they aren't real data, and they aren't real residuals!

```
plot( x = regression.2, which = 2 ) # Figure @ref{fig:regressionplot2}
```

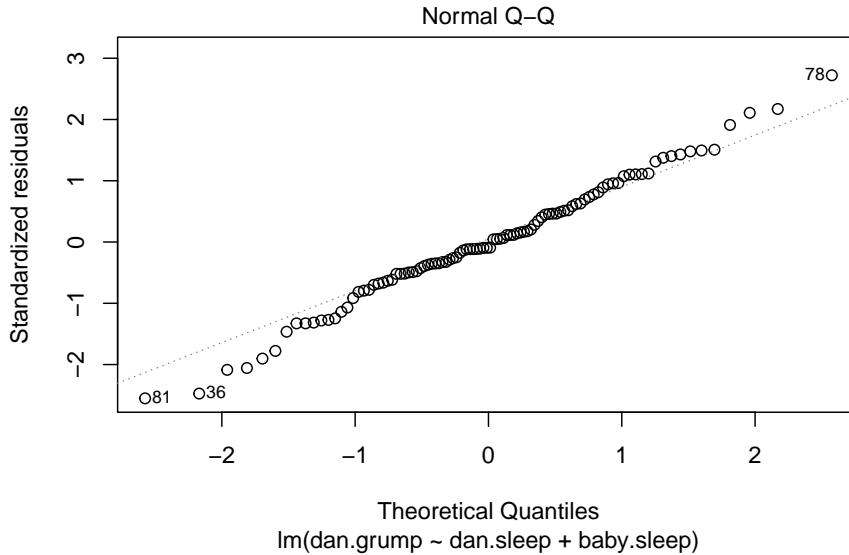


Figure 8.18: Plot of the theoretical quantiles according to the model, against the quantiles of the standardised residuals. This is one of the standard regression plots produced by the `plot()` function when the input is a linear regression object. It is obtained by setting `which=2`.

The output is shown in Figure 8.18, showing the standardised residuals plotted as a function of their theoretical quantiles according to the regression model. The fact that the output appends the model specification to the picture is nice.

8.13.4 Checking the linearity of the relationship

The third thing we might want to test is the linearity of the relationships between the predictors and the outcomes. There's a few different things that you might want to do in order to check this. Firstly, it never hurts to just plot the relationship between the fitted values \hat{Y}_i and the observed values Y_i for the outcome variable, as illustrated in Figure 8.19. To draw this we could use the `fitted.values()` function to extract the \hat{Y}_i values in much the same way that we used the `residuals()` function to extract the ϵ_i values. So the commands to draw this figure might look like this:

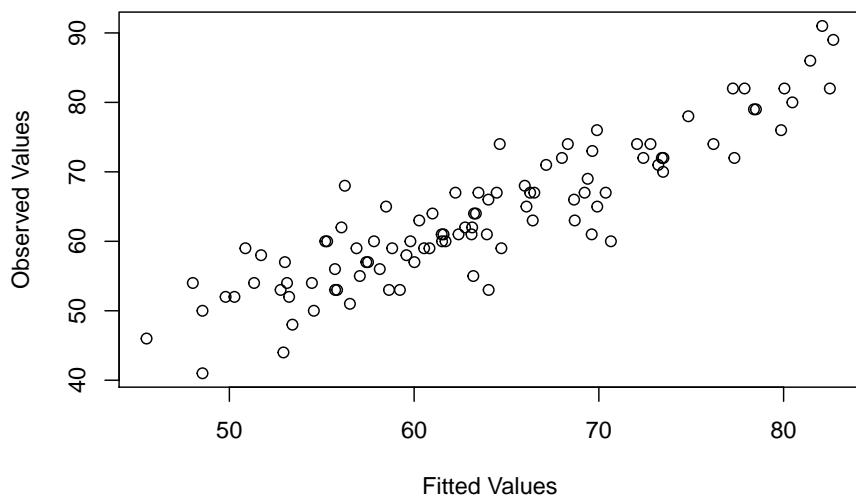


Figure 8.19: Plot of the fitted values against the observed values of the outcome variable. A straight line is what we're hoping to see here. This looks pretty good, suggesting that there's nothing grossly wrong, but there could be hidden subtle issues.

```

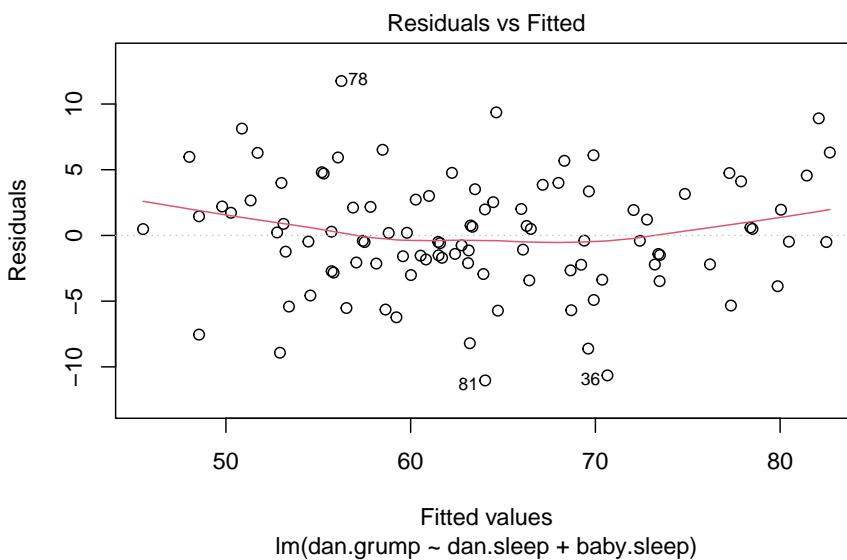
yhat.2 <- fitted.values( object = regression.2 )
plot( x = yhat.2,
      y = parenthood$dan.grump,
      xlab = "Fitted Values",
      ylab = "Observed Values"
)

```

One of the reasons I like to draw these plots is that they give you a kind of “big picture view”. If this plot looks approximately linear, then we’re probably not doing too badly (though that’s not to say that there aren’t problems). However, if you can see big departures from linearity here, then it strongly suggests that you need to make some changes.

In any case, in order to get a more detailed picture it’s often more informative to look at the relationship between the fitted values and the residuals themselves. Again, we could draw this plot using low level commands, but there’s an easier way. Just `plot()` the regression model, and select `which = 1`:

```
plot(x = regression.2, which = 1)
```



The output is shown in Figure ???. As you can see, not only does it draw the scatterplot showing the fitted value against the residuals, it also plots a line through the data that shows the relationship between the two. Ideally, this should be a straight, perfectly horizontal line. There’s some hint of curvature here, but it’s not clear whether or not we be concerned.

A somewhat more advanced version of the same plot is produced by the `residualPlots()` function in the `car` package. This function not only draws plots comparing the fitted values to the residuals, it does so for each individual predictor. The command is and the resulting plots are shown in Figure 8.20.

```
residualPlots( model = regression.2 )
```

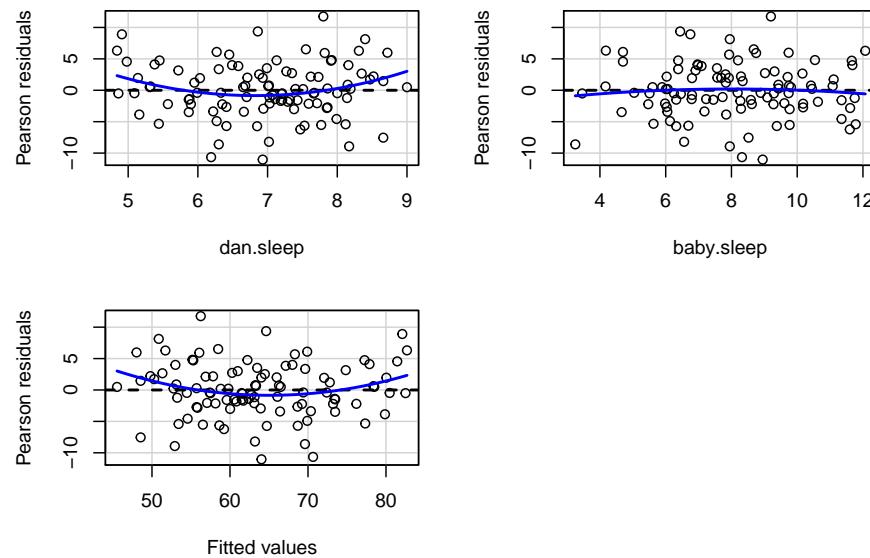


Figure 8.20: Plot of the fitted values against the residuals for `regression.2`, along with similar plots for the two predictors individually. This plot is produced by the `residualPlots()` function in the `car` package. Note that it refers to the residuals as “Pearson residuals”, but in this context these are the same as ordinary residuals.

```
##           Test stat Pr(>|Test stat|)
## dan.sleep     2.1604      0.03323 *
## baby.sleep    -0.5445      0.58733
## Tukey test     2.1615      0.03066 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that this function also reports the results of a bunch of *curvature tests*. For a predictor variable X in some regression model, this test is equivalent to adding a new predictor to the model corresponding to X^2 , and running the t -test

on the b coefficient associated with this new predictor. If it comes up significant, it implies that there is some nonlinear relationship between the variable and the residuals.

The third line here is the **Tukey test**, which is basically the same test, except that instead of squaring one of the predictors and adding it to the model, you square the fitted-value. In any case, the fact that the curvature tests have come up significant is hinting that the curvature that we can see in Figures ?? and 8.20 is genuine;¹⁸ although it still bears remembering that the pattern in Figure 8.19 is pretty damn straight: in other words the deviations from linearity are pretty small, and probably not worth worrying about.

In a lot of cases, the solution to this problem (and many others) is to transform one or more of the variables.

Dave note: In my graduate training, I was taught to be conservative with variable transformations because they improve the statistical performance at the cost of interpretability. A significant relationship between a log-transformed Y and X needs to be interpreted that way; you can no longer say you found a relationship between Y and X. For this reason, it may be preferable to suffer a decrease in statistical performance (possibly increasing the Type II error probability) rather than end up with a significant model that is challenging to interpret. In case you need it (and because Navarro (2018) said it was common), I will leave the following mention of the Box-Cox transform.

We discussed the basics of variable transformation in Sections ?? and ??mathfunc), but I do want to make special note of one additional possibility that I didn't mention earlier: the Box-Cox transform. The Box-Cox function is a fairly simple one, but it's very widely used

$$f(x, \lambda) = \frac{x^\lambda - 1}{\lambda}$$

for all values of λ except $\lambda = 0$. When $\lambda = 0$ we just take the natural logarithm (i.e., $\ln(x)$). You can calculate it using the `boxCox()` function in the `car` package. Better yet, if what you're trying to do is convert a data to normal, or as normal as possible, there's the `powerTransform()` function in the `car` package that can estimate the best value of λ . Variable transformation is another topic that deserves a fairly detailed treatment, but (again) due to deadline constraints, it will have to wait until a future version of this book.

8.13.5 Checking the homogeneity of variance

The regression models that we've talked about all make a homogeneity of variance assumption: the variance of the residuals is assumed to be constant. The

¹⁸And, if you take the time to check the `residualPlots()` for `regression.1`, it's pretty clear that this isn't some wacky distortion being caused by the fact that `baby.sleep` is a useless predictor variable. It's an actual nonlinearity in the relationship between `dan.sleep` and `dan.grump`.

“default” plot that R provides to help with doing this (`which = 3` when using `plot()`) shows a plot of the square root of the size of the residual $\sqrt{|\epsilon_i|}$, as a function of the fitted value \hat{Y}_i . We can produce the plot using the following command,

```
plot(x = regression.2, which = 3)
```

and the resulting plot is shown in Figure 8.21. Note that this plot actually uses the standardised residuals (i.e., converted to z scores) rather than the raw ones, but it’s immaterial from our point of view. What we’re looking to see here is a straight, horizontal line running through the middle of the plot.

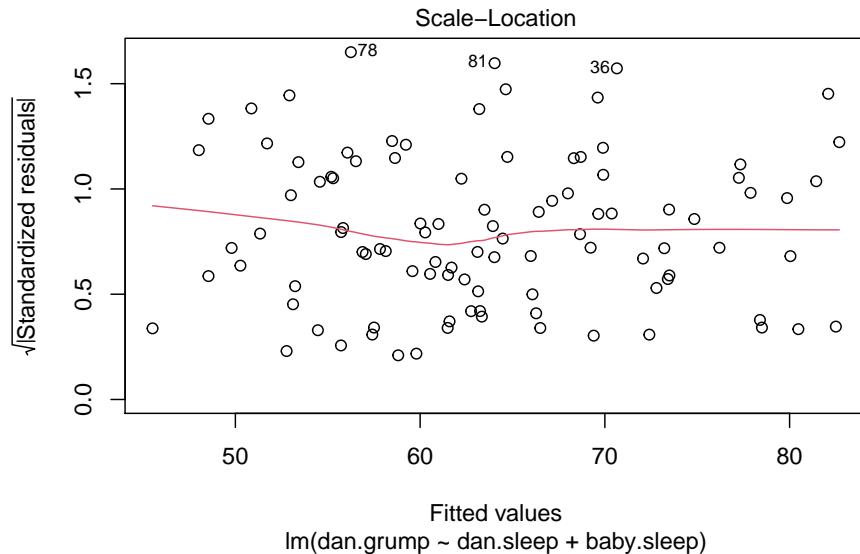


Figure 8.21: Plot of the fitted values (model predictions) against the square root of the abs standardised residuals. This plot is used to diagnose violations of homogeneity of variance. If the variance is really constant, then the line through the middle should be horizontal and flat. This is one of the standard regression plots produced by the `plot()` function when the input is a linear regression object. It is obtained by setting `which=3`.

A slightly more formal approach is to run hypothesis tests. The `car` package provides a function called `ncvTest()` (*non-constant variance test*) that can be used for this purpose (Cook and Weisberg, 1983). I won’t explain the details of how it works, other than to say that the idea is that what you do is run a regression to see if there is a relationship between the squared residuals ϵ_i

and the fitted values \hat{Y}_i , or possibly to run a regression using all of the original predictors instead of just \hat{Y}_i .¹⁹ Using the default settings, the `ncvTest()` looks for a relationship between \hat{Y}_i and the variance of the residuals, making it a straightforward analogue of Figure 8.21. So if we run it for our model,

```
ncvTest( regression.2 )

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.09317511, Df = 1, p = 0.76018
```

We see that our original impression was right: there's no violations of homogeneity of variance in this data.

Dave note: According to Cohen (2013), ANOVA is robust to violations of homogeneity of variance so long as sample sizes are roughly equal. If your model is comparing groups, and those groups are unequal such that one group is 1.5x or more of another, then you may need to address your violation of homogeneity of variance. In a design with balanced groups, I would not go to the trouble of the correction described next. I leave it included here in case you ever need it.

It's a bit beyond the scope of this chapter to talk too much about how to deal with violations of homogeneity of variance, but I'll give you a quick sense of what you need to consider. The *main* thing to worry about, if homogeneity of variance is violated, is that the standard error estimates associated with the regression coefficients are no longer entirely reliable, and so your t tests for the coefficients aren't quite right either. A simple fix to the problem is to make use of a "heteroscedasticity corrected covariance matrix" when estimating the standard errors. These are often called **sandwich estimators**, for reasons that only make sense if you understand the maths at a low level²⁰ have implemented as the default in the `hccm()` function is a tweak on this, proposed by Long and Ervin (2000). This version uses $\Sigma = \text{diag}(\epsilon_i^2 / (1 - h_i^2))$, where h_i is the i th hat value. Gosh, regression is *fun*, isn't it?] You don't need to understand what this means (not for an introductory class), but it might help to note that there's a `hccm()` function in the `car()` package that does it. Better yet, you don't even need to use it. You can use the `coeftest()` function in the `lmtest` package, but you need the `car` package loaded:

¹⁹Note that the underlying mechanics of the test aren't the same as the ones I've described for regressions; the goodness of fit is assessed using what's known as a score-test not an F -test, and the test statistic is (approximately) χ^2 distributed if there's no relationship

²⁰Again, a footnote that should be read only by the two readers of this book that love linear algebra (mmmm... I love the smell of matrix computations in the morning; smells like... nerd). In these estimators, the covariance matrix for b is given by $(X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}$. See, it's a "sandwich"? Assuming you think that $(X^T X)^{-1}$ = "bread" and $X^T \Sigma X$ = "filling", that is. Which of course everyone does, right? In any case, the usual estimator is what you get when you set $\Sigma = \hat{\sigma}^2 I$. The corrected version that I learned originally uses $\Sigma = \text{diag}(\epsilon_i^2)$ (White, 1980). However, the version that Fox and Weisberg (2011)

```

library(lmtest)
library(car)
coeftest( regression.2, vcov= hccm )

##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) 125.965566   3.247285 38.7910 <2e-16 ***
## dan.sleep    -8.950250   0.615820 -14.5339 <2e-16 ***
## baby.sleep     0.010524   0.291565   0.0361   0.9713
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Not surprisingly, these t tests are pretty much identical to the ones that we saw when we used the `summary(regression.2)` command earlier; because the homogeneity of variance assumption wasn't violated. But if it had been, we might have seen some more substantial differences.

8.13.6 Checking for collinearity

The last kind of regression diagnostic that I'm going to discuss in this chapter is the use of *variance inflation factors* (VIFs), which are useful for determining whether or not the predictors in your regression model are too highly correlated with each other. There is a variance inflation factor associated with each predictor X_k in the model, and the formula for the k -th VIF is:

$$\text{VIF}_k = \frac{1}{1 - R_{(-k)}^2}$$

where $R_{(-k)}^2$ refers to R -squared value you would get if you ran a regression using X_k as the outcome variable, and all the other X variables as the predictors. The idea here is that $R_{(-k)}^2$ is a very good measure of the extent to which X_k is correlated with all the other variables in the model. Better yet, the square root of the VIF is pretty interpretable: it tells you how much wider the confidence interval for the corresponding coefficient b_k is, relative to what you would have expected if the predictors are all nice and uncorrelated with one another. If you've only got two predictors, the VIF values are always going to be the same, as we can see if we use the `vif()` function (`car` package)...

```

vif( mod = regression.2 )

##
##  dan.sleep  baby.sleep
## 1.651038   1.651038

```

And since the square root of 1.65 is 1.28, we see that the correlation between our two predictors isn't causing much of a problem.

To give a sense of how we could end up with a model that has bigger collinearity problems, suppose I were to run a much less interesting regression model, in which I tried to predict the `day` on which the data were collected, as a function of all the other variables in the data set. To see why this would be a bit of a problem, let's have a look at the correlation matrix for all four variables:

```
cor( parenthood )
```

```
##          dan.sleep  baby.sleep  dan.grump      day
## dan.sleep   1.0000000  0.62794934 -0.90338404 -0.09840768
## baby.sleep   0.62794934  1.00000000 -0.56596373 -0.01043394
## dan.grump   -0.90338404 -0.56596373  1.00000000  0.07647926
## day        -0.09840768 -0.01043394  0.07647926  1.00000000
```

We have some fairly large correlations between some of our predictor variables! When we run the regression model and look at the VIF values, we see that the collinearity is causing a lot of uncertainty about the coefficients. First, run the regression...

```
regression.3 <- lm( day ~ baby.sleep + dan.sleep + dan.grump, parenthood )
```

and second, look at the VIFs...

```
vif( regression.3 )
```

```
## baby.sleep  dan.sleep  dan.grump
##    1.651064   6.102337   5.437903
```

Yep, that's some mighty fine collinearity you've got there.

Dave note: A VIF cutoff of 4 is commonly used, where $VIF \leq 4$ is not concerning, but $VIF > 4$ suggests redundancy of variables in the model. The statistical solution would be to eliminate redundant variables from the model.

8.14 Model selection

One fairly major problem that remains is the problem of “model selection”. That is, if we have a data set that contains several variables, which ones should we include as predictors, and which ones should we not include? In other words, we have a problem of ***variable selection***. In general, model selection is a

complex business, but it's made somewhat simpler if we restrict ourselves to the problem of choosing a subset of the variables that ought to be included in the model. Nevertheless, I'm not going to try covering even this reduced topic in a lot of detail. Instead, I'll talk about two broad principles that you need to think about; and then discuss one concrete tool that R provides to help you select a subset of variables to include in your model. Firstly, the two principles:

- It's nice to have an actual substantive basis for your choices. That is, in a lot of situations you the researcher have good reasons to pick out a smallish number of possible regression models that are of theoretical interest; these models will have a sensible interpretation in the context of your field. Never discount the importance of this. Statistics serves the scientific process, not the other way around.
- To the extent that your choices rely on statistical inference, there is a trade off between simplicity and goodness of fit. As you add more predictors to the model, you make it more complex; each predictor adds a new free parameter (i.e., a new regression coefficient), and each new parameter increases the model's capacity to "absorb" random variations. So the goodness of fit (e.g., R^2) continues to rise as you add more predictors no matter what. If you want your model to be able to generalise well to new observations, you need to avoid throwing in too many variables.

This latter principle is often referred to as ***Ockham's razor***, and is often summarised in terms of the following pithy saying: *do not multiply entities beyond necessity*. We also calls this **parsimony**. In this context, it means: don't chuck in a bunch of largely irrelevant predictors just to boost your R^2 . Hm. Yeah, the original was better.

In any case, what we need is an actual mathematical criterion that will implement the qualitative principle behind Ockham's razor in the context of selecting a regression model. As it turns out there are several possibilities. The one that I'll talk about is the **Akaike information criterion** [AIC; Akaike (1974)] simply because it's the default one used in the R function `step()`. In the context of a linear regression model (and ignoring terms that don't depend on the model in any way!), the AIC for a model that has K predictor variables plus an intercept is:²¹

$$\text{AIC} = \frac{\text{SS}_{\text{res}}}{\hat{\sigma}} + 2K$$

²¹Note, however, that the `step()` function computes the full version of AIC, including the irrelevant constants that I've dropped here. As a consequence this equation won't correctly describe the AIC values that you see in the outputs here. However, if you calculate the AIC values using my formula for two different regression models and take the difference between them, this will be the same as the differences between AIC values that `step()` reports. In practice, this is all you care about: the actual value of an AIC statistic isn't very informative, but the differences between two AIC values *are* useful, since these provide a measure of the extent to which one model outperforms another.

The smaller the AIC value, the better the model performance is. If we ignore the low level details, it's fairly obvious what the AIC does: on the left we have a term that increases as the model predictions get worse; on the right we have a term that increases as the model complexity increases. The best model is the one that fits the data well (low residuals; left hand side) using as few predictors as possible (low K ; right hand side). In short, this is a simple implementation of Ockham's razor.

Dave note: Please consider the following sections to have a big "Warning!" sign next to them when used in confirmatory research. They are inherently exploratory techniques. One assumption common to all automated model selection strategies is that they assume that your sample data are perfect representations of their population distributions. The computer is effectively testing many possible models and picking the one that fits best. The resulting model is always exploratory, tentative, and highly dependent on the sample. It would need to be tested in a confirmatory study. It would be highly inappropriate and misleading to present a computer-generated model as if it was developed *a priori* or without explaining the process used to create the model.

8.14.1 Backward elimination

Okay, let's have a look at the `step()` function at work. In this example I'll keep it simple and use only the basic **backward elimination** approach. That is, start with the complete regression model, including all possible predictors. Then, at each "step" we try all possible ways of removing one of the variables, and whichever of these is best (in terms of lowest AIC value) is accepted. This becomes our new regression model; and we then try all possible deletions from the new model, again choosing the option with lowest AIC. This process continues until we end up with a model that has a lower AIC value than any of the other possible models that you could produce by deleting one of its predictors. Let's see this in action. First, I need to define the model from which the process starts.

```
full.model <- lm( formula = dan.grump ~ dan.sleep + baby.sleep + day,
                   data = parenthood
)
```

That's nothing terribly new: yet another regression. Booooring. Still, we do need to do it: the `object` argument to the `step()` function will be this regression model. With this in mind, I would call the `step()` function using the following command:

```
step( object = full.model,      # start at the full model
      direction = "backward"   # allow it remove predictors but not add them
)
```

```

## Start: AIC=299.08
## dan.grump ~ dan.sleep + baby.sleep + day
##
##          Df Sum of Sq    RSS    AIC
## - baby.sleep  1      0.1 1837.2 297.08
## - day         1      1.6 1838.7 297.16
## <none>           1837.1 299.08
## - dan.sleep   1    4909.0 6746.1 427.15
##
## Step: AIC=297.08
## dan.grump ~ dan.sleep + day
##
##          Df Sum of Sq    RSS    AIC
## - day        1      1.6 1838.7 295.17
## <none>           1837.2 297.08
## - dan.sleep  1    8103.0 9940.1 463.92
##
## Step: AIC=295.17
## dan.grump ~ dan.sleep
##
##          Df Sum of Sq    RSS    AIC
## <none>           1838.7 295.17
## - dan.sleep  1    8159.9 9998.6 462.50

##
## Call:
## lm(formula = dan.grump ~ dan.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)  dan.sleep
##       125.956     -8.937

```

although in practice I didn't need to specify `direction` because "backward" is the default. The output is somewhat lengthy, so I'll go through it slowly. Firstly, the output reports the AIC value for the current best model:

```

Start: AIC=299.08
dan.grump ~ dan.sleep + baby.sleep + day

```

That's our starting point. Since small AIC values are good, we want to see if we can get a value smaller than 299.08 by deleting one of those three predictors. So what R does is try all three possibilities, calculate the AIC values for each one, and then print out a short table with the results:

Df	Sum of Sq	RSS	AIC
----	-----------	-----	-----

```

- baby.sleep 1      0.1 1837.2 297.08
- day         1      1.6 1838.7 297.16
<none>          1837.1 299.08
- dan.sleep  1     4909.0 6746.1 427.15

```

To read this table, it helps to note that the text in the left hand column is telling you what *change* R made to the regression model. So the line that reads <none> is the actual model we started with, and you can see on the right hand side that this still corresponds to an AIC value of 299.08 (obviously). The other three rows in the table correspond to the other three models that it looked at: it tried removing the `baby.sleep` variable, which is indicated by - `baby.sleep`, and this produced an AIC value of 297.08. That was the best of the three moves, so it's at the top of the table. So, this move is accepted, and now we start again. There are two predictors left in the model, `dan.sleep` and `day`, so it tries deleting those:

```

Step: AIC=297.08
dan.grump ~ dan.sleep + day

```

	Df	Sum of Sq	RSS	AIC
- day	1	1.6	1838.7	295.17
<none>			1837.2	297.08
- dan.sleep	1	8103.0	9940.1	463.92

Okay, so what we can see is that removing the `day` variable lowers the AIC value from 297.08 to 295.17. So R decides to keep that change too, and moves on:

```

Step: AIC=295.17
dan.grump ~ dan.sleep

```

	Df	Sum of Sq	RSS	AIC
<none>			1838.7	295.17
- dan.sleep	1	8159.9	9998.6	462.50

This time around, there's no further deletions that can actually improve the AIC value. So the `step()` function stops, and prints out the result of the best regression model it could find:

```

Call:
lm(formula = dan.grump ~ dan.sleep, data = parenthood)

Coefficients:
(Intercept)    dan.sleep
125.956       -8.937

```

which is (perhaps not all that surprisingly) the `regression.1` model that we started with at the beginning of the chapter.

8.14.2 Forward selection

As an alternative, you can also try *forward selection*. This time around we start with the smallest possible model as our start point, and only consider the possible additions to the model. However, there's one complication: you also need to tell `step()` what the largest possible model you're willing to entertain is, using the `scope` argument. The simplest usage is like this:

```
null.model <- lm( dan.grump ~ 1, parenthood ) # intercept only.
step( object = null.model,          # start with null.model
      direction = "forward",       # only consider "addition" moves
      scope = dan.grump ~ dan.sleep + baby.sleep + day # largest model allowed
)

## Start: AIC=462.5
## dan.grump ~ 1
##
##           Df Sum of Sq   RSS   AIC
## + dan.sleep    1     8159.9 1838.7 295.17
## + baby.sleep   1     3202.7 6795.9 425.89
## <none>          9998.6 462.50
## + day         1      58.5 9940.1 463.92
##
## Step: AIC=295.17
## dan.grump ~ dan.sleep
##
##           Df Sum of Sq   RSS   AIC
## <none>          1838.7 295.17
## + day         1     1.55760 1837.2 297.08
## + baby.sleep  1     0.02858 1838.7 297.16

##
## Call:
## lm(formula = dan.grump ~ dan.sleep, data = parenthood)
##
## Coefficients:
## (Intercept)  dan.sleep
##       125.956      -8.937
```

If I do this, the output takes on a similar form, but now it only considers addition (+) moves rather than deletion (-) moves:

```
Start: AIC=462.5
dan.grump ~ 1
```

```

Df Sum of Sq    RSS    AIC
+ dan.sleep   1     8159.9 1838.7 295.17
+ baby.sleep   1     3202.7 6795.9 425.89
<none>                   9998.6 462.50
+ day         1      58.5 9940.1 463.92

Step: AIC=295.17
dan.grump ~ dan.sleep

Df Sum of Sq    RSS    AIC
<none>                   1838.7 295.17
+ day         1     1.55760 1837.2 297.08
+ baby.sleep   1     0.02858 1838.7 297.16

Call:
lm(formula = dan.grump ~ dan.sleep, data = parenthood)

Coefficients:
(Intercept)  dan.sleep
        125.956       -8.937

```

As you can see, it's found the same model. In general though, forward and backward selection don't always have to end up in the same place.

8.14.3 A caveat

Automated variable selection methods are seductive things, especially when they're bundled up in (fairly) simple functions like `step()`. They provide an element of objectivity to your model selection, and that's kind of nice. Unfortunately, they're sometimes used as an excuse for thoughtlessness. No longer do you have to think carefully about which predictors to add to the model and what the theoretical basis for their inclusion might be... everything is solved by the magic of AIC. And if we start throwing around phrases like Ockham's razor, well, it sounds like everything is wrapped up in a nice neat little package that no-one can argue with.

Or, perhaps not. Firstly, there's very little agreement on what counts as an appropriate model selection criterion. When I was taught backward elimination as an undergraduate, we used *F*-tests to do it, because that was the default method used by the software. The default in the `step()` function is AIC, and since this is an introductory text that's the only method I've described, but the AIC is hardly the Word of the Gods of Statistics. It's an approximation, derived under certain assumptions, and it's guaranteed to work only for large samples when those assumptions are met. Alter those assumptions and you get

a different criterion, like the BIC for instance. Take a different approach again and you get the NML criterion. Decide that you're a Bayesian and you get model selection based on posterior odds ratios. Then there are a bunch of regression specific tools that I haven't mentioned. And so on. All of these different methods have strengths and weaknesses, and some are easier to calculate than others (AIC is probably the easiest of the lot, which might account for its popularity). Almost all of them produce the same answers when the answer is "obvious" but there's a fair amount of disagreement when the model selection problem becomes hard.

What does this mean in practice? Well, you *could* go and spend several years teaching yourself the theory of model selection, learning all the ins and outs of it; so that you could finally decide on what you personally think the right thing to do is. Speaking as someone who actually did that, I wouldn't recommend it: you'll probably come out the other side even more confused than when you started. A better strategy is to show a bit of common sense... if you're staring at the results of a `step()` procedure, and the model that makes sense is close to having the smallest AIC, but is narrowly defeated by a model that doesn't make any sense... trust your instincts. Statistical model selection is an inexact tool, and as I said at the beginning, *interpretability matters*.

8.14.4 Comparing two regression models

An alternative to using automated model selection procedures is for the researcher to explicitly select two or more regression models to compare to each other. You can do this in a few different ways, depending on what research question you're trying to answer. Suppose we want to know whether or not the amount of sleep that my son got has any relationship to my grumpiness, over and above what we might expect from the amount of sleep that I got. We also want to make sure that the day on which we took the measurement has no influence on the relationship. That is, we're interested in the relationship between `baby.sleep` and `dan.grump`, and from that perspective `dan.sleep` and `day` are nuisance variable or **covariates** that we want to control for. In this situation, what we would like to know is whether `dan.grump ~ dan.sleep + day + baby.sleep` (which I'll call Model 1, or M1) is a better regression model for these data than `dan.grump ~ dan.sleep + day` (which I'll call Model 0, or M0). There are two different ways we can compare these two models, one based on a model selection criterion like AIC, and the other based on an explicit hypothesis test. I'll show you the AIC based approach first because it's simpler, and follows naturally from the `step()` function that we saw in the last section. The first thing I need to do is actually run the regressions:

```
M0 <- lm( dan.grump ~ dan.sleep + day, parenthood )
M1 <- lm( dan.grump ~ dan.sleep + day + baby.sleep, parenthood )
```

Now that I have my regression models, I could use the `summary()` function to run various hypothesis tests and other useful statistics, just as we have discussed throughout this chapter. However, since the current focus on model comparison, I'll skip this step and go straight to the AIC calculations. Conveniently, the `AIC()` function in R lets you input several regression models, and it will spit out the AIC values for each of them:²²

```
AIC( M0, M1 )

##      df      AIC
## M0    4  582.8681
## M1    5  584.8646
```

Since Model 0 has the smaller AIC value, it is judged to be the better model for these data.

A somewhat different approach to the problem comes out of the hypothesis testing framework. Suppose you have two regression models, where one of them (Model 0) contains a *subset* of the predictors from the other one (Model 1). That is, Model 1 contains all of the predictors included in Model 0, plus one or more additional predictors. When this happens we say that Model 0 is ***nested*** within Model 1, or possibly that Model 0 is a ***submodel*** of Model 1. Regardless of the terminology what this means is that we can think of Model 0 as a null hypothesis and Model 1 as an alternative hypothesis. And in fact we can construct an F test for this in a fairly straightforward fashion. We can fit both models to the data and obtain a residual sum of squares for both models. I'll denote these as $\text{SS}_{\text{res}}^{(0)}$ and $\text{SS}_{\text{res}}^{(1)}$ respectively. The superscripting here just indicates which model we're talking about. Then our F statistic is

$$F = \frac{(\text{SS}_{\text{res}}^{(0)} - \text{SS}_{\text{res}}^{(1)})/k}{(\text{SS}_{\text{res}}^{(1)})/(N - p - 1)}$$

where N is the number of observations, p is the number of predictors in the full model (not including the intercept), and k is the difference in the number of parameters between the two models.²³ The degrees of freedom here are k and

²²While I'm on this topic I should point out that there is also a function called `BIC()` which computes the Bayesian information criterion (BIC) for the models. So you could type `BIC(M0,M1)` and get a very similar output. In fact, while I'm not particularly impressed with either AIC or BIC as model selection methods, if you do find yourself using one of these two, the empirical evidence suggests that BIC is the better criterion of the two. In most simulation studies that I've seen, BIC does a much better job of selecting the correct model.

²³It's worth noting in passing that this same F statistic can be used to test a much broader range of hypotheses than those that I'm mentioning here. Very briefly: notice that the nested model $M0$ corresponds to the full model $M1$ when we constrain some of the regression coefficients to zero. It is sometimes useful to construct submodels by placing other kinds of constraints on the regression coefficients. For instance, maybe two different coefficients might have to sum to zero, or something like that. You can construct hypothesis tests for those

$N - p - 1$. Note that it's often more convenient to think about the difference between those two SS values as a sum of squares in its own right. That is:

$$\text{SS}_\Delta = \text{SS}_{res}^{(0)} - \text{SS}_{res}^{(1)}$$

The reason why this is helpful is that we can express SS_Δ a measure of the extent to which the two models make different predictions about the outcome variable. Specifically:

$$\text{SS}_\Delta = \sum_i (\hat{y}_i^{(1)} - \hat{y}_i^{(0)})^2$$

where $\hat{y}_i^{(0)}$ is the fitted value for y_i according to model M_0 and $\hat{y}_i^{(1)}$ is the fitted value for y_i according to model M_1 .

Okay, so that's the hypothesis test that we use to compare two regression models to one another. Now, how do we do it in R? The answer is to use the `anova()` function. All we have to do is input the two models that we want to compare (null model first):

```
anova( M0, M1 )
```

```
## Analysis of Variance Table
##
## Model 1: dan.grump ~ dan.sleep + day
## Model 2: dan.grump ~ dan.sleep + day + baby.sleep
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     97 1837.2
## 2     96 1837.1  1  0.063688 0.0033 0.9541
```

Note that, just like we saw with the output from the `step()` function, R has used the acronym RSS to refer to the residual sum of squares from each model. That is, RSS in this output corresponds to SS_{res} in the formula above. Since we have $p > .05$ we retain the null hypothesis (M_0). This approach to regression, in which we add all of our covariates into a null model, and then *add* the variables of interest into an alternative model, and then compare the two models in hypothesis testing framework, is often referred to as *hierarchical regression*.

8.15 Practical Issues in Correlation and Regression

Text by David Schuster

kind of constraints too, but it is somewhat more complicated and the sampling distribution for F can end up being something known as the non-central F distribution, which is waaaaay beyond the scope of this book! All I want to do is alert you to this possibility.

8.15.1 Correlation is not causation

Remember that causality can be inferred when the cause comes before the effect, a relationship is shown between cause and effect, and we can find no other plausible explanations for the effect other than the cause. Our statistical analysis (regardless of which GLM analysis is used) can help with the middle criterion; we can use statistics to provide evidence of a relationship between two variables. However, no statistical analysis can, by itself, allow us to make causal inferences. Our best tool for making causal inferences is an experimental design. Designs that are quasi-experimental or non-experimental will not let us make inferences about causality.

In this statement, correlation refers to the evidence of a relationship between two variables. Therefore, it would be a bit clearer if it was, “quasi- and non-experimental research designs are not causation.” but I suppose the original version is catchier.

8.15.2 Interpreting NHST in Big Data

Sample size must be sufficiently large to detect a linear relationship. Larger samples increase power. Power is especially important when your measures are not reliable or when there is a weak linear relationship you are trying to find evidence for. You can also have too large of a sample. Extremely large samples with small effect sizes will be significant because it is assumed that your sample is an almost perfect representation of the population. Thus, even the tiniest effects will be significant. Take care to examine effect size and sample size when conducting NHST for correlations. Besides issues of power, a sufficiently large sample size is important to increase the stability of your estimates of the strength and direction of the correlation.

8.15.3 Outliers

If either of your variables includes an outlier, it will reduce the strength of your correlation. Because of this, you want to consider whether the extreme score belongs as part of your data set. The most common cause of an outlier in real world data is a mistake during data entry. As previously stated, being an outlier is not generally, itself, enough criteria for a researcher to exclude a case from a sample without concern that the sample will be biased.

8.15.4 Restriction of Range

When scores that are very high or very low on the IV, range restriction can occur. Range restriction lowers the strength of the correlation. **Floor and ceiling effects** can cause range restriction. Imagine a very easy math test as a measure

of math skill. Students at the highest levels of math ability will score 100% on the test. Because the test does not measure past 100%, some information is lost about individuals who could have scored higher than 100% if it were possible. This is an example of a ceiling effect. Similarly, a very challenging test can result in 0% scores with the test similarly lacking diagnosticity between low-scorers at the bottom of the distribution.

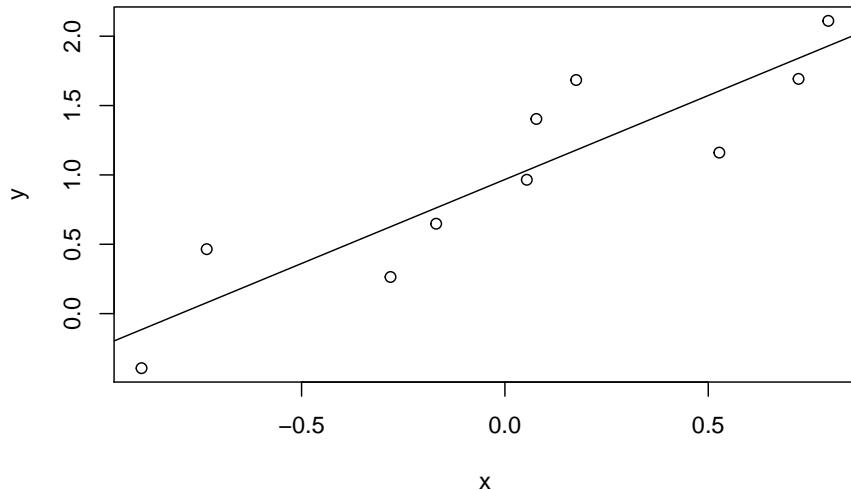
This can be illustrated easily by drawing scatterplots. First, we start with two variables with a high correlation:

```
# Two variables with a known correlation of r = .9.
x <- c(-0.169062295, 0.053822267, 0.527217327, 0.07727407, 0.722027061, -0.28132826, -0.8934
y <- c(0.648162835, 0.964985264, 1.160828115, 1.403311403, 1.692166183, 0.26407813, -0.3931

cor(x, y) # r = .9

## [1] 0.9

plot(x,y,xlab="x",ylab="y") # generate a scatterplot and label the axes
abline(reg = lm(y ~ x)) # draw the line of best fit for the regression equation
```



We will create a floor effect by saying that 0 is now the lowest possible score on this measure. Any negative values become zero in our new variable, `x_floor`. Notice the change in the line, fit, and r .

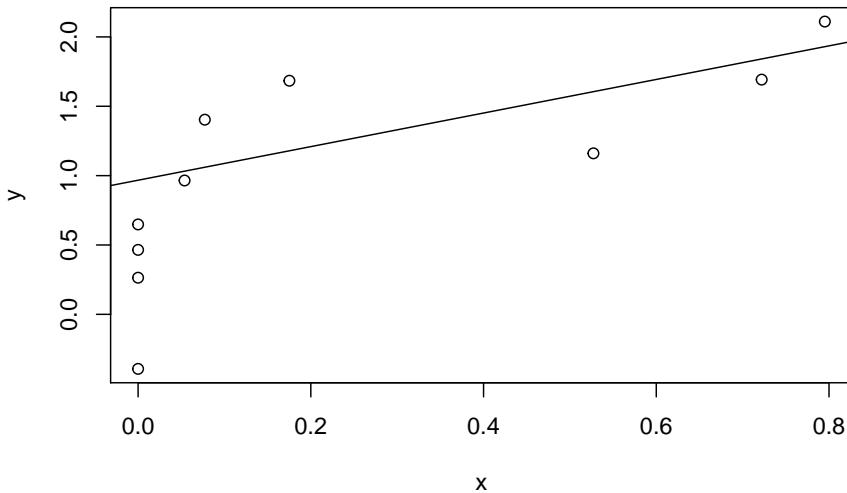
```
# Two variables with a known correlation of r = .9, changing all scores where x < 0 to 0.

x_floor <- c(0,0.053822267,0.527217327,0.07727407,0.722027061,0,0,0.79516792,0.175189753,0)
y <- c(0.648162835, 0.964985264,1.160828115,1.403311403,1.692166183,0.26407813,-0.393116812,2.110632623,1.684217475)

cor(x_floor, y) # Correlation r = .74

## [1] 0.7353465

plot(x_floor,y,xlab="x",ylab="y") # generate a scatterplot and label the axes
abline(reg = lm(y ~ x)) # draw the line of best fit for the regression equation
```



A floor/ceiling effect can cause this problem but is not the only cause. Simply excluding low scores can have this effect. Next, we will eliminate any case with a negative value of x instead of removing it. Notice that although no scores are changed, the correlation is weaker. Whenever we start with a linear relationship and “zoom in” to a portion of it, the subset of the data will be a weaker correlation than the overall relationship. Therefore, restriction of range can result in a weaker correlation.

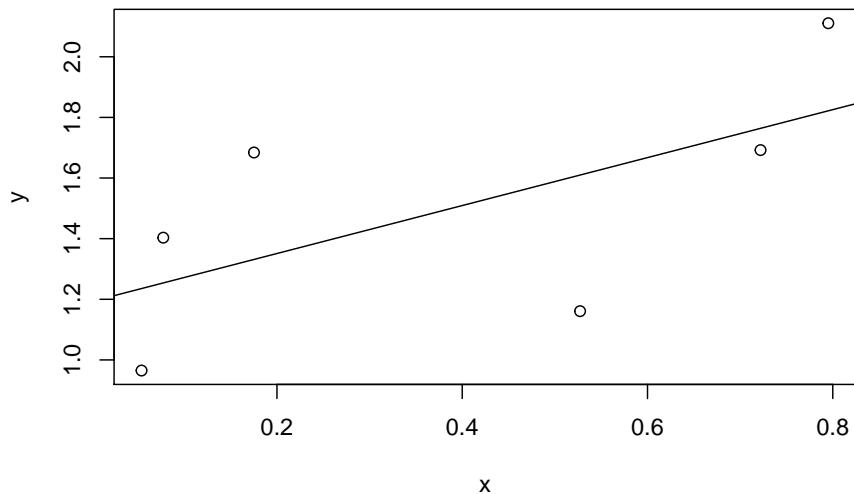
```
# Two variables with a known correlation of r = .9, excluding all scores where x < 0.

x <- c(0.053822267,0.527217327,0.07727407,0.722027061,0.79516792,0.175189753)
y <- c(0.964985264,1.160828115,1.403311403,1.692166183,2.110632623,1.684217475)

cor(x, y) # r = .9
```

```
## [1] 0.6347461

plot(x,y,xlab="x",ylab="y") # generate a scatterplot and label the axes
abline(reg = lm(y ~ x)) # draw the line of best fit for the regression equation
```



8.15.5 Regression Toward the Mean

Galton (1886) came up with the concept of regression toward mediocrity which he applied to heredity. This phrase is designed to be pejorative (suggesting that inheritance is losing something of value) and Galton's explanation for why it mattered was eugenics (see also the Video: Eugenics and Francis Galton - 13 min on this topic). We now understand that it has nothing to do with heredity and more to do with the properties of the mean. In present use, regression toward the mean refers to mistaking an extreme score as evidence of an effect. Imagine you are playing a challenging level of a video game. After a lot of practice, your score tends to hover around 10,000 points. Sometimes you score a bit more, sometimes a bit less, but if you were to have all the data available, you would find a mean of 10,000 points. Regression toward the mean would occur if you played this level one day and scored 20,000 points, then concluded that you were a video game superstar, left school, and went to be a pro gamer. Playing the level again, you are dismayed that your score has fallen back to around 10,000 points. What happened? How could you gain and lose such an ability? What is happening is that your performance has variability, and it is unlikely

but possible to observe an extreme score. Researchers need to be mindful that extreme events will tend to be less extreme with repeated sampling.

8.15.6 Report Effect Size

Interpretation of the correlation coefficient r (or R^2). The statistic r measures both strength (i.e., it is the measure of effect size) and direction of a correlation. A strong correlation means that the data points of the scatterplot lie close to the line. When you report a correlation, give the proper interpretation of its effect size. Also note that, depending on the research question, small effect sizes can be important and meaningful.

How do you interpret the direction of an effect in multiple regression? R^2 will always be positive. The answer is to look at the direction of the coefficients. If they are positive, it suggests a positive relationship with the DV. If negative, it suggests a negative relationship with the DV.

8.15.6.1 Interpretation of η^2 and r^2 (Cohen, 1988)

These are reference points, not firm cutoffs. For example, .056 is a medium effect size.

Effect Size	Interpretation
$\eta^2 = r^2 = .01$	Small effect
$\eta^2 = r^2 = .06$	Medium effect
$\eta^2 = r^2 = .14$	Large effect

8.15.6.2 Interpretation of r (Cohen, 1988; Note: This is not r^2)

These are reference points, not firm cutoffs. For example, .056 is a medium effect size.

Effect Size	Interpretation
$r \pm .10$ Small effe	ct
$r \pm .30$ Medium e	ffect
$r \pm .50$ Large effe	ct

8.15.7 What are Degrees of Freedom, again?

This is more of a footnote. The calculation of variance is a fraction. The numerator of the fraction is a sum of squared deviations. The denominator of

the fraction is degrees of freedom. I find it helpful to think about degrees of freedom in any of three ways (pick which makes the most sense to you):

1. Degrees of freedom (df), are the number of values free to vary given some constraint. The calculation of a mean has $N - 1$ degrees of freedom because if the mean is known and $N - 1$ of the values are known, then the last value is defined. Thus, only $N - 1$ values are free to vary. This is my least favorite explanation.
2. Degrees of freedom (df) are the number of values that provide information about variability. If you know only your score on the last exam, you could use that as an estimate of the class mean, but you would have no way of estimating variability of test scores in class ($df = 0$). If you know your score and another classmate's, you can now estimate variability, albeit crudely ($df = 1$). With your score and seven others, you can make a better estimate of the variability in the class ($df = 7$). This is my favorite definition and it comes from Cohen (2013).
3. Because degrees of freedom (df) are directly related to sample size, think of degrees of freedom as sample size. I think of the sum of squared deviations as adding up all of the variation that has been observed, and degrees of freedom as being the sample size. Therefore, I think of variance as being roughly the variation per unit, or the average variation. This is the least precise.

8.16 Summary

- Basic ideas in linear regression and how regression models are estimated (Sections 8.5 and 8.6).
- Multiple linear regression (Section 8.7).
- Measuring the overall performance of a regression model using R^2 (Section 8.8)
- Hypothesis tests for regression models (Section 8.9)
- Calculating confidence intervals for regression coefficients, and standardised coefficients (Section 8.11)
- The assumptions of regression (Section 8.12) and how to check them (Section 8.13)
- Selecting a regression model (Section 8.14)

Chapter 9

Statistics Reference

9.1 One-Sample z -Test

Video: The one-sample Z-test

9.1.1 Definition

The one-sample z-test tests the null hypothesis that a mean is equivalent to the mean of a known population.

9.1.2 Test Statistic

The test statistic is z , which measures the distance between two means. In this case, one mean is from our sample and the other mean is a known constant. The sampling distribution of z is the normal distribution with a standard deviation defined by the formula for standard error ($\sigma_{\bar{X}}$).

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_{hyp}}{\frac{\sigma}{\sqrt{N}}}$$

9.1.3 Assumptions & Required Data

- 1 variable measured using a quantitative, continuous scale
- The variable was measured for a sample that was taken randomly, with replacement, from a population
- The normality assumption, meaning at least one of these:
 - $N \geq 30$
 - The variable is normally distributed in the population

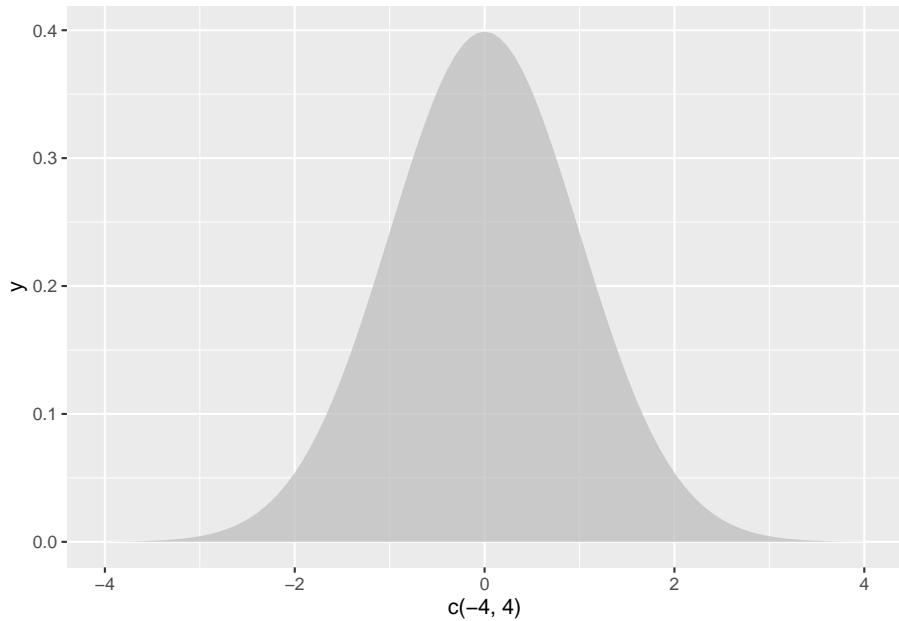


Figure 9.1: The null hypothesis distribution of z

- The population mean, μ , is known.
- The population standard deviation, σ , is known

9.1.4 When to use it

Use a z -test when you are comparing a single sample mean to a known population parameter and can meet the assumptions.

If the population standard deviation is unknown, or if the normality assumption cannot be met, consider a t -test.

9.1.5 Example

Imagine a high school has a graduation test with $M = .80$ with a standard deviation (σ) of $\sigma = .10$. A random sample of $N = 35$ students at the high school participate in an after-school program aimed at increasing performance on the graduation test.

9.1.5.1 Data

The data are test scores from 35 students.

```

## [1] 1.00 0.77 0.66 0.65 1.05 0.97 0.90 0.71 0.88 1.00 0.75 0.67 0.68 0.88
## [15] 0.92 0.87 0.94 0.78 0.98 0.93 0.93 1.00 0.97 0.95 0.85 1.07 0.87 0.89
## [29] 0.89 0.89 1.06 1.02 0.69 0.93 0.96

## [1] 0.8845714

```

The students in the program took the test and performed higher than the population average ($M = \text{print}(\text{mean}(\text{sample}))$). Is there evidence that the after school program is effective?

9.1.5.2 Hypotheses

Because researchers are interested in detecting higher performance on the test, a one-tailed test is used to increase statistical power. If, instead, researchers wanted to see if the sample had higher or lower performance, a two-tailed test should be used.

```

$H_0=\mu\leq .80$

$H_a=\mu>.80$

```

9.1.5.3 Analysis

Set the alpha level. By convention, an alpha level of $\alpha = .05$ will be used.

Assume the null hypothesis is true. Assuming the null hypothesis is true means that we need to determine the probability of obtaining a sample mean this distance from the population mean. We will determine this using the sampling distribution of the null hypothesis for z (the normal distribution).

Unlike later statistical tests, R does not provide a built-in z -test. This is actually a feature, as it lets us demonstrate the steps in more detail.

The most challenging part is the function `pnorm()`, which gives the area to the left of a score on a standard normal distribution. By using the argument `'lower.tail = FALSE'`, the function will give the area to the right of the score.

```

mu <- .80
sigma <- .10
n <- length(data)
z <- (mean(sample) - mu) / (sigma / sqrt(n))
z

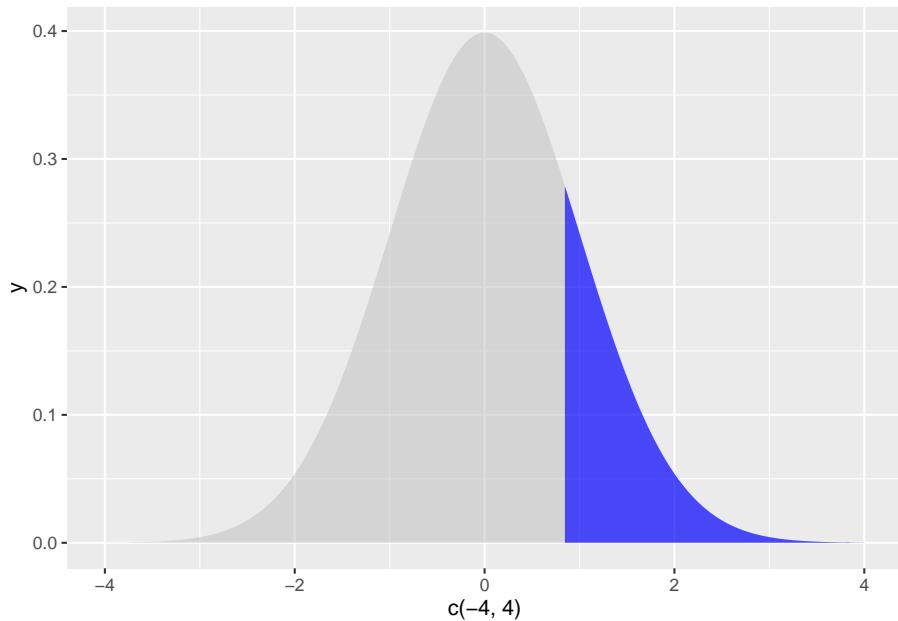
## [1] 0.8457143

```

```
p_value <- pnorm(z, mean = 0, sd = 1, lower.tail = FALSE) # gives area to the right of
p_value
## [1] 0.1988561
```

To visualize this result, we can graph the location of the test statistic in the sampling distribution, shading everything beyond the test statistic in one tail:

```
library(ggplot2)
ggplot(NULL, aes(c(-4,4))) +
  geom_area(stat = "function", fun = dnorm, xlim = c(-4, z), alpha = 0.5, fill=alpha("grey"))
  geom_area(stat = "function", fun = dnorm, fill="blue", xlim = c(z, 4), alpha = 0.7)
```



The shaded area is well over 5 percent, showing visually that $p > \alpha$.

9.1.5.4 Decision

Because $p > \alpha$, the null hypothesis is retained and the results are inconclusive. These data do not provide evidence of effectiveness of the program.

9.1.5.5 Variations

- This was a one-tailed test on the right side of the distribution. The use of `rnorm()` would need to be adapted if the one-tailed test was on the left

side of the distribution (to detect if scores were lower than the population). Simply omit `lower.tail = FALSE` to have `rnorm()` calculate from the left side (lower tail).

- In a two-tailed test, the shading would need to be repeated on the left side, and the shaded area on both sides would need to be added together. You can save a step by knowing that each tail is always the same area. To convert this one-tailed p-value into a two-tailed p-value, you would need to double it, giving you a two-tailed p-value of `{r, echo=FALSE} print(p_value*2)`. When doing a two-tailed test, check to make sure you are calculating in the correct tail; if your two-tailed test had a sample mean lower than the population mean, you would want to shade/calculate to the left.
- If $p < \alpha$, you would have rejected the null hypothesis and concluded that there was a difference between your sample mean and the population.

9.2 Correlation

9.2.1 Definition

A correlation analysis measures the strength and direction of a relationship between two variables. The hypothesis test for a correlation tests the null hypothesis that there is no linear relationship between two variables. This is also called a bivariate correlation (because it involves two variables) and the Pearson correlation coefficient.

9.2.2 Test Statistic

The test statistic is actually a t distribution calculated from the observed value of r , which measures the strength and direction of the relationship. The statistic r has values $-1 \leq r \leq 1$, with -1 indicating a perfect negative relationship and +1 indicating a perfect positive relationship. $r = 0$ indicates no relationship between the variables and is rarely observed to be exactly 0 in practice. To conduct a hypothesis test, r is converted to a value of t because this function of the sampling distribution of r follows a t -distribution:

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

This function can be reversed as $r = \frac{t}{\sqrt{n-2+t^2}}$

The t distribution is actually a family of distributions defined by degrees of freedom. **Degrees of freedom** is a concept that can be interpreted multiple ways. For now, it is sufficient to say that it is based on sample size. The value of degrees of freedom grows by 1 with each additional unit increased in the sample.

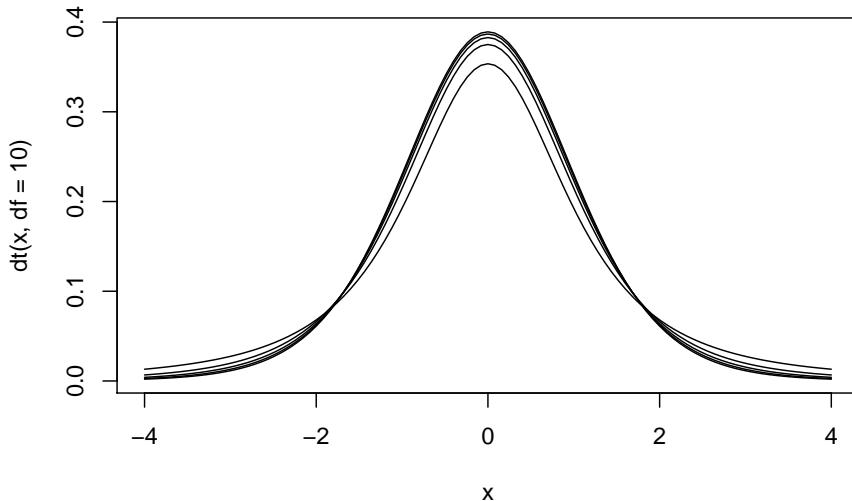


Figure 9.2: The null hypothesis distribution of t with values of df between 2 and 10. Notice how the curve is starting to converge at the higher values of df .

In other words, the specific sampling distribution used in the hypothesis test depends on the sample size (and degrees of freedom).

9.2.3 Assumptions & Required Data

For more detail, see 8.12.

- 2 quantitative variables (interval or ratio), or one quantitative variable with a dichotomous (two possible values) variable. The latter version is called a point-biserial correlation and is mathematically the same as the Pearson correlation coefficient. Note that this procedure is not appropriate for ordinal variables, and Spearman's rank correlation coefficient should be used instead.
- Normality, meaning normal distribution of residuals.
- Linearity
- Homogeneity of variance (also called homoscedasiticty and equality of variance)
- No outliers

9.2.4 When to use it

Use a correlation when you want to detect a linear relationship between two variables.

9.2.5 Example

Navarro (2018) wanted to see if there was a relationship between hours slept in a night and their rating of grumpiness the next day.

9.2.5.1 Data

The data are two variables, one indicating sleep and the other indicating grumpiness. Data were collected for 100 nights.

```
##      dan.sleep baby.sleep dan.grump day
## 1      7.59     10.18      56   1
## 2      7.91     11.66      60   2
## 3      5.14      7.92      82   3
## 4      7.71      9.61      55   4
## 5      6.68      9.75      67   5
## 6      5.99      5.04      72   6
## 7      8.19     10.45      53   7
## 8      7.19      8.27      60   8
## 9      7.40      6.06      60   9
## 10     6.58      7.09      71  10
## 11     6.49     11.68      72  11
## 12     6.27      6.13      65  12
## 13     5.95      7.83      74  13
## 14     6.65      5.60      67  14
## 15     6.41      6.03      66  15
## 16     6.33      8.19      69  16
## 17     6.30      6.38      73  17
## 18     8.47     11.11      52  18
## 19     7.21      5.51      61  19
## 20     7.53      6.69      53  20
## 21     8.00      9.74      54  21
## 22     7.35      9.02      63  22
## 23     6.86      6.44      74  23
## 24     7.86      9.43      56  24
## 25     4.86      3.46      82  25
## 26     5.87      6.32      72  26
## 27     8.40      7.95      59  27
## 28     6.93      7.69      66  28
```

## 29	7.21	7.45	60	29
## 30	6.99	7.56	67	30
## 31	8.17	7.95	44	31
## 32	7.85	11.61	53	32
## 33	6.27	4.70	76	33
## 34	8.66	8.52	41	34
## 35	4.98	4.70	86	35
## 36	6.19	8.32	60	36
## 37	6.41	9.38	63	37
## 38	4.84	4.18	89	38
## 39	7.03	5.98	61	39
## 40	7.66	9.29	57	40
## 41	7.51	6.01	59	41
## 42	7.92	10.54	60	42
## 43	8.12	11.78	48	43
## 44	7.47	11.60	53	44
## 45	7.99	11.35	50	45
## 46	5.44	5.63	72	46
## 47	8.16	6.98	57	47
## 48	7.62	6.03	60	48
## 49	5.87	4.66	70	49
## 50	9.00	9.81	46	50
## 51	8.31	12.07	58	51
## 52	6.71	7.57	68	52
## 53	7.43	11.35	58	53
## 54	5.90	5.47	71	54
## 55	8.52	8.29	52	55
## 56	6.03	6.80	74	56
## 57	7.29	10.63	59	57
## 58	7.32	8.59	59	58
## 59	6.88	7.82	67	59
## 60	6.22	7.18	67	60
## 61	6.94	8.29	61	61
## 62	7.01	11.08	64	62
## 63	7.20	6.46	61	63
## 64	6.30	3.25	61	64
## 65	8.72	9.74	54	65
## 66	7.82	8.75	62	66
## 67	8.14	11.75	52	67
## 68	7.27	9.31	64	68
## 69	6.70	7.73	65	69
## 70	7.55	8.68	65	70
## 71	7.38	9.77	57	71
## 72	7.73	9.71	59	72
## 73	5.32	4.17	79	73
## 74	7.86	10.18	53	74

## 75	6.35	9.28	67	75
## 76	7.11	7.23	61	76
## 77	5.45	6.38	82	77
## 78	7.80	9.20	68	78
## 79	7.13	8.20	67	79
## 80	8.35	10.16	54	80
## 81	6.93	8.95	53	81
## 82	7.07	6.80	62	82
## 83	8.66	8.34	50	83
## 84	5.09	6.25	80	84
## 85	4.91	6.75	91	85
## 86	7.03	9.09	62	86
## 87	7.02	10.42	64	87
## 88	7.67	8.89	57	88
## 89	8.15	9.43	54	89
## 90	5.88	6.79	72	90
## 91	5.72	6.91	78	91
## 92	6.66	6.05	63	92
## 93	6.85	6.32	59	93
## 94	5.57	8.62	74	94
## 95	5.16	7.84	76	95
## 96	5.31	5.89	79	96
## 97	7.77	9.77	51	97
## 98	5.38	6.97	82	98
## 99	7.02	6.56	55	99
## 100	6.45	7.93	74	100

9.2.5.2 Hypotheses

Given their often-exploratory use, correlations are typically conducted as two-tailed tests. However, a one-tailed correlation could be conducted if researchers predict a direction for the effect.

Hypotheses are always written with population parameters (since we are making hypotheses about truth in the population, not what we have observed in our sample). The population parameter corresponding to the test statistic r is ρ (rho). The null is that there is no relationship. The alternative hypothesis is that there is a relationship.

$$H_0 = \rho = 0$$

$$H_a = \rho \neq 0$$

9.2.5.3 Analysis

Set the alpha level. By convention, an alpha level of $\alpha = .05$ will be used.

Assume the null hypothesis is true. Assuming the null hypothesis is true means that we need to determine the probability of obtaining an effect size (r) this strong at our sample size through random sampling from a population with effect size $\rho = 0$. We will determine this using the sampling distribution of the null hypothesis for t .

There are several ways to generate correlations in R:

- `cor()` will output just the correlation coefficient
- `cor.test()` will perform NHST and give a p-value

```
cor( x = parenthood$dan.sleep, y = parenthood$dan.grump )
```

```
## [1] -0.903384
```

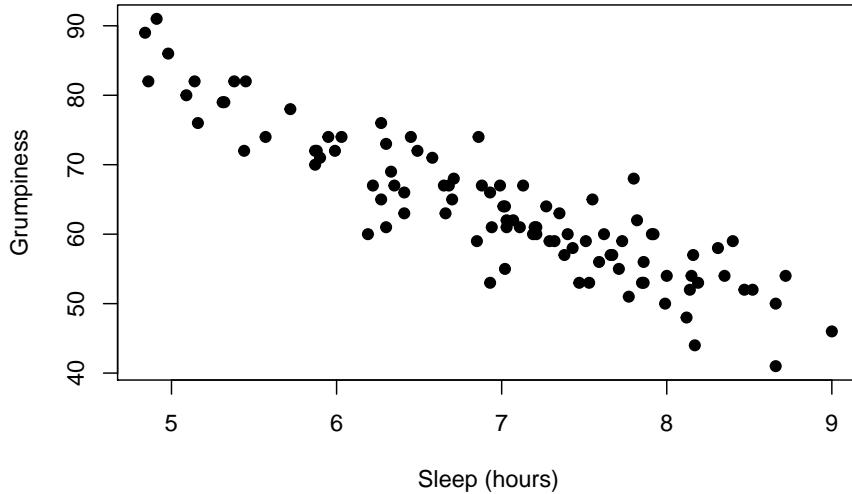
```
cor.test(x = parenthood$dan.sleep, y = parenthood$dan.grump)
```

```
##
##  Pearson's product-moment correlation
##
## data:  parenthood$dan.sleep and parenthood$dan.grump
## t = -20.854, df = 98, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.9340614 -0.8594714
## sample estimates:
##      cor
## -0.903384
```

To visualize this result, we can graph a scatterplot, with observed values of one variable plotted against the observed value of the second variable:

```
oneCorPlot <- function(x,y,...) {
  plot(x,y,pch=19,col="black"),...
}

oneCorPlot( parenthood$dan.sleep, parenthood$dan.grump,
  xlab="Sleep (hours)", ylab="Grumpiness"
)
```



The magnitude of r (it's value) indicates the strength of the relationship. The closer the value of r to ± 1 , the more closely the points hug a straight line. The line will have a positive slope if the relationship is positive and a negative slope if the relationship is negative.

9.2.5.4 Decision

Because $p < \alpha$, the null hypothesis is rejected and we conclude that there is a relationship between sleep and grumpiness. Further, because the value of r is negative, the relationship between grumpiness and sleep is that higher amounts of sleep are associated with lower levels of grumpiness.

9.2.5.5 Variations

- If data are ordinal, Spearman's rank order correlation can be calculated by adding the argument method = "separman" using the following syntax:

```
cor( x, y, method = "spearman")
```

- In a two-tailed test, the shading would need to be repeated on the left side, and the shaded area on both sides would need to be added together. You can save a step by knowing that each tail is always the same area. To convert this one-tailed p-value into a two-tailed p-value, you would

need to double it, giving you a two-tailed p-value of `{r, echo=FALSE}`
`print(p_value*2)`. When doing a two-tailed test, check to make sure
you are calculating in the correct tail; if your two-tailed test had a sample
mean lower than the population mean, you would want to shade/calculate
to the left.

- If $p \geq \alpha$, you would have retained the null hypothesis and made no conclusion.
- You can calculate a **correlation matrix** that shows all possible bivariate correlations from a dataframe. Simply include the entire dataframe as the argument instead of two variables, like this:

```
cor( x = parenthood)
```

```
##                  dan.sleep  baby.sleep  dan.grump      day
## dan.sleep     1.00000000  0.62794934 -0.90338404 -0.09840768
## baby.sleep    0.62794934  1.00000000 -0.56596373 -0.01043394
## dan.grump    -0.90338404 -0.56596373  1.00000000  0.07647926
## day          -0.09840768 -0.01043394  0.07647926  1.00000000
```

When using a correlation matrix, p-values are not interpretable because the probability of a type I error on one or more of these correlations is higher than .05 (because the alpha level of .05 is used on each one, and many tests are being conducted).

9.3 Linear Regression

9.3.1 Definition

Linear regression creates a linear model that can be used to predict an outcome variable from one or more predictor variables. The hypothesis test for a linear regression occurs in two stages. The first stage, called the **omnibus test**, is a hypothesis test of the entire model. If this test is significant, researchers can conclude that the collection of variables significantly predicts the outcome. Assuming significance, the second stage follows in which each predictor variable is tested with a against the null hypothesis that its regression coefficient is zero.

9.3.2 Test Statistic

The omnibus test uses an F -test to test the null hypothesis that there is no linear relationship between the outcome variable and the linear combination of the predictor variables. Or, that the variance explained by the predictor

variables is no more than the variance explained by the mean. Tests of each coefficient use t -tests. The statistics F and t are related such that $t = \sqrt{F}$ and $F = t^2$.

$$F = \frac{\text{MS}_{model}}{\text{MS}_{residual}}$$

$$t = \frac{\hat{b}}{\text{SE}(\hat{b})}$$

As with t , F is a family of distributions defined around two degrees of freedom.

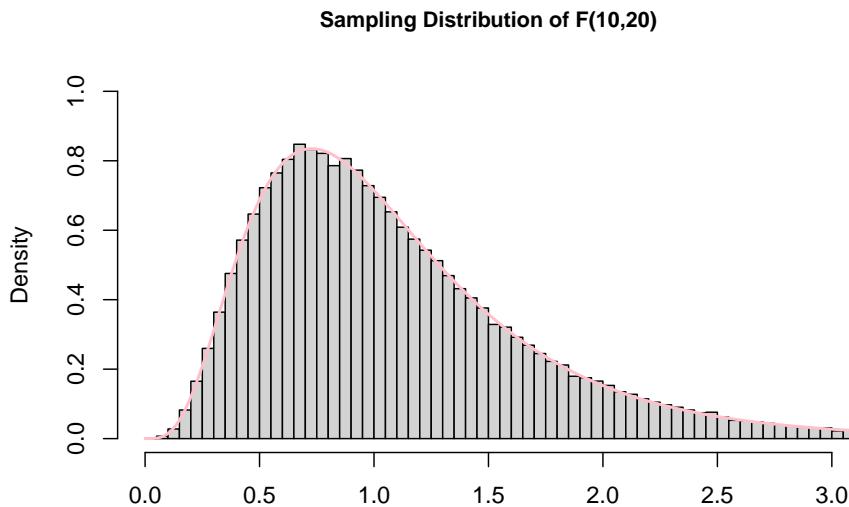


Figure 9.3: Graph of the sampling distribution of F used under Creative Commons License from Hartmann, K., Krois, J., Waske, B. (2018): E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin.

9.3.3 Assumptions & Required Data

For more detail, see 8.12.

- A single quantitative outcome variable (interval or ratio), and one or more quantitative or dichotomous (two possible values) variables. Note that this procedure is not appropriate for ordinal variables.

- Normality of residuals
- Linearity
- Homogeneity of variance (also called homoscedasticity and equality of variance)
- Uncorrelated predictors
- Residuals that are independent of each other
- No outliers

9.3.4 When to use it

Use a regression when you want to formulate and test a linear model, including mediating and moderating variables. Regression can also be used to run ANOVAs.

9.3.5 Example

The same data from the correlation guide will be used. Navarro (2018) wanted to see if hours slept in a night and hours the baby slept in a night could predict their rating of grumpiness the next day.

9.3.5.1 Data

The data are three variables, one indicating hours slept (predictor 1), one indicating the hours slept by the baby (predictor 2), and the other (the outcome) indicating grumpiness. Data were collected for 100 nights.

```
##      dan.sleep baby.sleep dan.grump day
## 1        7.59     10.18       56   1
## 2        7.91     11.66       60   2
## 3        5.14      7.92       82   3
## 4        7.71      9.61       55   4
## 5        6.68      9.75       67   5
## 6        5.99      5.04       72   6
## 7        8.19     10.45       53   7
## 8        7.19      8.27       60   8
## 9        7.40      6.06       60   9
## 10       6.58      7.09       71  10
## 11       6.49     11.68       72  11
## 12       6.27      6.13       65  12
## 13       5.95      7.83       74  13
## 14       6.65      5.60       67  14
## 15       6.41      6.03       66  15
## 16       6.33      8.19       69  16
```

## 17	6.30	6.38	73	17
## 18	8.47	11.11	52	18
## 19	7.21	5.51	61	19
## 20	7.53	6.69	53	20
## 21	8.00	9.74	54	21
## 22	7.35	9.02	63	22
## 23	6.86	6.44	74	23
## 24	7.86	9.43	56	24
## 25	4.86	3.46	82	25
## 26	5.87	6.32	72	26
## 27	8.40	7.95	59	27
## 28	6.93	7.69	66	28
## 29	7.21	7.45	60	29
## 30	6.99	7.56	67	30
## 31	8.17	7.95	44	31
## 32	7.85	11.61	53	32
## 33	6.27	4.70	76	33
## 34	8.66	8.52	41	34
## 35	4.98	4.70	86	35
## 36	6.19	8.32	60	36
## 37	6.41	9.38	63	37
## 38	4.84	4.18	89	38
## 39	7.03	5.98	61	39
## 40	7.66	9.29	57	40
## 41	7.51	6.01	59	41
## 42	7.92	10.54	60	42
## 43	8.12	11.78	48	43
## 44	7.47	11.60	53	44
## 45	7.99	11.35	50	45
## 46	5.44	5.63	72	46
## 47	8.16	6.98	57	47
## 48	7.62	6.03	60	48
## 49	5.87	4.66	70	49
## 50	9.00	9.81	46	50
## 51	8.31	12.07	58	51
## 52	6.71	7.57	68	52
## 53	7.43	11.35	58	53
## 54	5.90	5.47	71	54
## 55	8.52	8.29	52	55
## 56	6.03	6.80	74	56
## 57	7.29	10.63	59	57
## 58	7.32	8.59	59	58
## 59	6.88	7.82	67	59
## 60	6.22	7.18	67	60
## 61	6.94	8.29	61	61
## 62	7.01	11.08	64	62

## 63	7.20	6.46	61	63
## 64	6.30	3.25	61	64
## 65	8.72	9.74	54	65
## 66	7.82	8.75	62	66
## 67	8.14	11.75	52	67
## 68	7.27	9.31	64	68
## 69	6.70	7.73	65	69
## 70	7.55	8.68	65	70
## 71	7.38	9.77	57	71
## 72	7.73	9.71	59	72
## 73	5.32	4.17	79	73
## 74	7.86	10.18	53	74
## 75	6.35	9.28	67	75
## 76	7.11	7.23	61	76
## 77	5.45	6.38	82	77
## 78	7.80	9.20	68	78
## 79	7.13	8.20	67	79
## 80	8.35	10.16	54	80
## 81	6.93	8.95	53	81
## 82	7.07	6.80	62	82
## 83	8.66	8.34	50	83
## 84	5.09	6.25	80	84
## 85	4.91	6.75	91	85
## 86	7.03	9.09	62	86
## 87	7.02	10.42	64	87
## 88	7.67	8.89	57	88
## 89	8.15	9.43	54	89
## 90	5.88	6.79	72	90
## 91	5.72	6.91	78	91
## 92	6.66	6.05	63	92
## 93	6.85	6.32	59	93
## 94	5.57	8.62	74	94
## 95	5.16	7.84	76	95
## 96	5.31	5.89	79	96
## 97	7.77	9.77	51	97
## 98	5.38	6.97	82	98
## 99	7.02	6.56	55	99
## 100	6.45	7.93	74	100

9.3.5.2 Hypotheses

All F -tests have two-tailed, directionless hypotheses. Because the omnibus test uses the F -distribution, all regression omnibus tests are two-tailed.

The null is that the mean (which is the y -intercept, or b_0) predicts the outcome variable (Y_i) as well as the model. The alternative hypothesis is that the

predictors add explained variance.

$$H_0 : Y_i = b_0 + \epsilon_i$$

$$H_a : Y_i = \left(\sum_{k=1}^K b_k X_{ik} \right) + b_0 + \epsilon_i$$

9.3.5.3 Analysis

Set the alpha level. By convention, an alpha level of $\alpha = .05$ will be used.

Assume the null hypothesis is true. Assuming the null hypothesis is true means that we need to determine the probability of obtaining an effect size (in this case, the predictive power of our model over the null model with only the mean, measured using R^2) this strong at our sample size through random sampling from a population with no effect (a null model). We will determine this using an analysis of variance using the sampling distribution of the null hypothesis for F .

```
regression.2 <- lm( formula = dan.grump ~ dan.sleep + baby.sleep,
                     data = parenthood )
summary(regression.2)

##
## Call:
## lm(formula = dan.grump ~ dan.sleep + baby.sleep, data = parenthood)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -11.0345 -2.2198 -0.4016  2.6775 11.7496 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 125.96557   3.04095  41.423   <2e-16 ***
## dan.sleep    -8.95025   0.55346 -16.172   <2e-16 ***
## baby.sleep     0.01052   0.27106   0.039    0.969   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.354 on 97 degrees of freedom
## Multiple R-squared:  0.8161, Adjusted R-squared:  0.8123 
## F-statistic: 215.2 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
confint(regression.2, level = 0.95)

##                 2.5 %      97.5 %
## (Intercept) 119.930125 132.0010063
## dan.sleep   -10.048710  -7.8517895
## baby.sleep    -0.527462   0.5485109
```

The omnibus test can be found in the last row as F-statistic. Here, the omnibus test was significant, $F(2, 97) = 215.2, p < .001$, $R^2 = 0.82$, suggesting that the model significantly predicts grumpiness.

While a multiple regression with more than one predictor cannot be visualized easily, you could generate added variable plots for each predictor.

9.3.5.4 Decision

9.3.5.4.1 Omnibus Test Because $p < \alpha$, the null hypothesis is rejected and we conclude that grumpiness can be predicted by the linear combination of hours slept and hours the baby slept. However, we do not know which variable(s) significantly explain variance. Because we have a significant omnibus test, we are now justified in testing the coefficients.

9.3.5.4.2 Test of Coefficients t -tests are conducted for each coefficient in the `summary()` table. The hours slept coefficient was a significant predictor, $t(97) = -16.17, p < .001$, 95% CI [-10.05, -7.85]. As hours slept increased, grumpiness decreased. The hours the baby slept was not a significant predictor, $t(97) = -16.17, p < .001$, 95% CI [-0.53, 0.55].

Note that the degrees of freedom are the residual degrees of freedom (the second number). F tests are always reported with the model degrees of freedom first (2) and the residual degrees of freedom second (97).

9.3.5.5 Variations

This guide will be updated to add assumption checking and diagnostics procedures for regression.

Bibliography

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.
- Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, 27:17–21.
- Assocation, A. P. (2020). *Publication Manual of the American Psychological Association*. 7th edition.
- Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187:398–404.
- Campbell, D. T. and Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin, Boston, MA.
- Cohen, B. H. (2013). *Explaining Psychological Statistics*. John Wiley and Sons.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum, 2nd edition.
- Cook, R. D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, 70:1–10.
- Cook, T. D., Campbell, D. T., and Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA.
- Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, Cambridge, UK.
- Ellman, M. (2002). Soviet repression statistics: some comments. *Europe-Asia Studies*, 54(7):1151–1172.
- Fisher, R. A. (1922). On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222:309–368.
- Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage, Los Angeles, 2nd edition.

- Galton, F. (1886). Regression towards mediocrity in hereditary stature.
- Gelman, A. and Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, 60:328–331.
- Keynes, J. M. (1923). *A Tract on Monetary Reform*. Macmillan and Company, London.
- Kozma, A. and Stones, M. J. (1983). Predictors of happiness. *Journal of Gerontology*, 38.
- Lehmann, E. L. (2011). *Fisher, Neyman, and the Creation of Classical Statistics*. Springer.
- Long, J. and Ervin, L. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54:217–224.
- Meehl, P. H. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34:103–115.
- Navarro, D. (2018). *Learning statistics with R: A tutorial for psychology students and other beginners (version 0.6)*.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103:677–680.
- Stigler, S. M. (1986). *The History of Statistics*. Harvard University Press, Cambridge, MA.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrika*, 48:817–838.