LEARNING STATISTICS WITH JAMOVI: A TUTORIAL FOR BEGINNERS IN STATISTICAL ANALYSIS

DANIELLE NAVARRO
AND DAVID FOXCROFT

LEARNING STATISTICS WITH JAMOVI

Learning Statistics with jamovi

A Tutorial for Beginners in Statistical Analysis

Danielle Navarro and David Foxcroft

All external links were active at the time of publication unless otherwise stated and have been archived via the Internet Archive Wayback Machine at
https://archive.org/web

Digital material and resources associated with this volume are available at
https://doi.org/10.11647/OBP.0333#resources

ISBN Paperback: 978-1-80064-937-8

ISBN Hardback: 978-1-80064-938-5

ISBN Digital (PDF): 978-1-80064-939-2

DOI: 10.11647/OBP.0333

This textbook covers the contents of an introductory statistics class, as typically taught to undergraduate psychology, health or social science students. The book covers how to get started in jamovi as well as giving an introduction to data manipulation. From a statistical perspective, the book discusses descriptive statistics and graphing first, followed by chapters on probability theory, sampling and estimation, and null hypothesis testing. After introducing the theory, the book covers the analysis of contingency tables, correlation, *t*-tests, regression, ANOVA and factor analysis. Bayesian statistics are touched on at the end of the book.

Data sets used in the book are freely available for use in jamovi. All the data files you need can be accessed within jamovi via an add-on module in the jamovi library. Or you can download the files from https://www.learnstatswithjamovi.com.

Citation: Danielle Navarro and David Foxcroft, *Learning statistics with jamovi: A tutorial for beginners in statistical analysis*. Cambridge, UK: Open Book Publishers, 2025, https://doi.org/10.11647/OBP.0333

Table of contents

Preface to Version 1.00

This book is an adaptation of DJ Navarro (2018). Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.6). https://learningstatisticswithr.com/.

The jamovi version of this book was first released in 2018, as version 0.65. Versions 0.70 and 0.75 were released in subsequent years with corrections and additions. In that time, many people have contacted us asking for a hard copy version of the book. To achieve this, and to preserve the open source attributes of the book and materials, we have worked with Open Book Publishers in Cambridge, UK, to release this updated version. Open Book Publishers are the leading independent open access publisher of academic research in the Humanities and Social Sciences in the UK. They are award-winning, not-for-profit, run by scholars, and committed to making high-quality research freely available to readers around the world.

David Foxcroft
January 1st, 2025

Part I

Beginnings

Chapter 1

Why do we learn statistics

"Thou shalt not answer questionnaires
Or quizzes upon World Affairs,
Nor with compliance
Take any test. Thou shalt not sit
With statisticians nor commit
A social science"
– W.H. Auden[1]

1.1 On the psychology of statistics

To the surprise of many students, statistics is a fairly significant part of a psychological education. To the surprise of no-one, statistics is very rarely the *favourite* part of one's psychological education. After all, if you really loved the idea of doing statistics, you'd probably be enrolled in a statistics class right now, not a psychology class. So, not surprisingly, there's a pretty large proportion of the student base that isn't happy about the fact that psychology has so much statistics in it. In view of this, I thought that the right place to start might be to answer some of the more common questions that people have about stats.

A big part of this issue at hand relates to the very idea of statistics. What is it? What's it there for? And why are scientists so bloody obsessed with it? These are all good questions, when you think about it. So let's start with the last one. As a group, scientists seem to be bizarrely fixated on running statistical tests on everything. In fact, we use statistics so often that we sometimes forget to explain to people why we do. It's a kind of article of faith among scientists – and especially social scientists – that your findings can't be trusted until you've done some stats. Undergraduate students might

[1] The quote comes from Auden's 1946 poem *Under Which Lyre: A Reactionary Tract for the Times*, delivered as part of a commencement address at Harvard University. The history of the poem is kind of interesting, see Adam Kirsch's analysis in the *Harvard Magazine*

be forgiven for thinking that we're all completely mad, because no-one takes the time to answer one very simple question:

Why do you do statistics? Why don't scientists just use common sense?

It's a naive question in some ways, but most good questions are. There's a lot of good answers to it,[2] but for my money, the best answer is a really simple one: we don't trust ourselves enough. We worry that we're human, and susceptible to all of the biases, temptations and frailties that humans suffer from. Much of statistics is basically a safeguard. Using "common sense" to evaluate evidence means trusting gut instincts, relying on verbal arguments and on using the raw power of human reason to come up with the right answer. Most scientists don't think this approach is likely to work.

In fact, come to think of it, this sounds a lot like a psychological question to me, and since I do work in a psychology department, it seems like a good idea to dig a little deeper here. Is it really plausible to think that this "common sense" approach is very trustworthy? Verbal arguments have to be constructed in language, and all languages have biases – some things are harder to say than others, and not necessarily because they're false (e.g., quantum electrodynamics is a good theory, but hard to explain in words). The instincts of our "gut" aren't designed to solve scientific problems, they're designed to handle day to day inferences – and given that biological evolution is slower than cultural change, we should say that they're designed to solve the day-to-day problems for a *different world* than the one we live in. Most fundamentally, reasoning sensibly requires people to engage in "induction", making wise guesses and going beyond the immediate evidence of the senses to make generalisations about the world. If you think that you can do that without being influenced by various distractors, well, I have a bridge in London I'd like to sell you. Heck, as the next section shows, we can't even solve "deductive" problems (ones where no guessing is required) without being influenced by our pre-existing biases.

1.1.1 The curse of belief bias

People are mostly pretty smart. We're smarter than the other species that we share the planet with (though many people might disagree). Our minds are quite amazing things, and we seem to be capable of the most incredible feats of thought and reason. That doesn't make us perfect though. And among the many things that psychologists have shown over the years is that we really do find it hard to be neutral, to evaluate evidence impartially and without being swayed by pre-existing biases. A good example of this is the **belief bias effect** in logical reasoning: if you ask people to decide whether a particular argument is logically valid (i.e., the conclusion would be true if the premises were true), we tend to be influenced by the believability of the conclusion, even when we shouldn't. For instance, here's a valid argument where the conclusion is believable:

> All cigarettes are expensive (Premise 1)
> Some addictive things are inexpensive (Premise 2)
> Therefore, some addictive things are not cigarettes (Conclusion)

And here's a valid argument where the conclusion is not believable:

[2] Including the suggestion that common sense is in short supply among scientists.

All addictive things are expensive (Premise 1)
Some cigarettes are inexpensive (Premise 2)
Therefore, some cigarettes are not addictive (Conclusion)

The logical *structure* of argument #2 is identical to the structure of argument #1, and they're both valid. However, in the second argument, there are good reasons to think that premise 1 is incorrect, and as a result it's probably the case that the conclusion is also incorrect. But that's entirely irrelevant to the topic at hand; an argument is deductively valid if the conclusion is a logical consequence of the premises. That is, a valid argument doesn't have to involve true statements.

On the other hand, here's an invalid argument that has a believable conclusion:

All addictive things are expensive (Premise 1)
Some cigarettes are inexpensive (Premise 2)
Therefore, some addictive things are not cigarettes (Conclusion)

And finally, an invalid argument with an unbelievable conclusion:

All cigarettes are expensive (Premise 1)
Some addictive things are inexpensive (Premise 2)
Therefore, some cigarettes are not addictive (Conclusion)

Now, suppose that people really are perfectly able to set aside their pre-existing biases about what is true and what isn't, and purely evaluate an argument on its logical merits. We'd expect 100% of people to say that the valid arguments are valid, and 0% of people to say that the invalid arguments are valid. So if you ran an experiment looking at this, you'd expect to see data as in Table 1.1.

Table 1.1: Validity of arguments

	conclusion feels true	**conclusion feels false**
argument is valid	100% say "valid"	100% say "valid"
argument is invalid	0% say "valid"	0% say "valid"

If the psychological data looked like this (or even a good approximation to this), we might feel safe in just trusting our gut instincts. That is, it'd be perfectly okay just to let scientists evaluate data based on their common sense, and not bother with all this murky statistics stuff. However, you guys have taken psych classes, and by now you probably know where this is going.

In a classic study, J. St. B. T. Evans et al. (1983) ran an experiment looking at exactly this. What they found is that when pre-existing biases (i.e., beliefs) were in agreement with the structure of the data, everything went the way you'd hope (Table 1.2).

Not perfect, but that's pretty good. But look what happens when our intuitive feelings about the truth of the conclusion run against the logical structure of the argument (Table 1.3):

Table 1.2: Pre-existing biases and argument validity

	conclusion feels true	**conclusion feels false**
argument is valid	92% say "valid"	
argument is invalid		8% say "valid"

Table 1.3: Intuition and argument validity

	conclusion feels true	**conclusion feels false**
argument is valid	92% say "valid"	46% say "valid"
argument is invalid	92% say "valid"	8% say "valid"

Oh dear, that's not as good. Apparently, when people are presented with a strong argument that contradicts our pre-existing beliefs, we find it pretty hard to even perceive it to be a strong argument (people only did so 46% of the time). Even worse, when people are presented with a weak argument that agrees with our pre-existing biases, almost no-one can see that the argument is weak (people got that one wrong 92% of the time!).[3]

If you think about it, it's not as if these data are horribly damning. Overall, people did do better than chance at compensating for their prior biases, since about 60% of people's judgements were correct (you'd expect 50% by chance). Even so, if you were a professional "evaluator of evidence", and someone came along and offered you a magic tool that improves your chances of making the right decision from 60% to (say) 95%, you'd probably jump at it, right? Of course you would. Thankfully, we actually do have a tool that can do this. But it's not magic, it's statistics. So that's reason #1 why scientists love statistics. It's just too easy for us to "believe what we want to believe". So instead, if we want to "believe in the data", we're going to need a bit of help to keep our personal biases under control. That's what statistics does, it helps keep us honest.

1.2 The cautionary tale of Simpson's paradox

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 postgraduate courses. Specifically, the thing that caused the problem was the gender breakdown of their admissions (Table 1.4).

Given this, they were worried about being sued![4] 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

[3] In my more cynical moments I feel like this fact alone explains 95% of what I read on the internet.

[4] Earlier versions of these notes incorrectly suggested that they actually were sued. But that's not true. There's a nice commentary on this by Alex Reinhart 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.

Table 1.4: Berkeley students by gender

	Number of applicants	Percent admitted
Males	8442	44%
Females	4321	35%

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 they told a rather different story (Bickel et al., 1975). 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 1.5 shows the admission figures for the six largest departments (with the names of the departments removed for privacy reasons):

Table 1.5: Berkeley students by gender for six largest departments

	Males		**Females**	
Department	Applicants	Percent admitted	Applicants	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%

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?

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., A, B) tended to admit a high percentage of the qualified applicants, whereas others (e.g., F) 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 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, counter-intuitive 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. 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 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

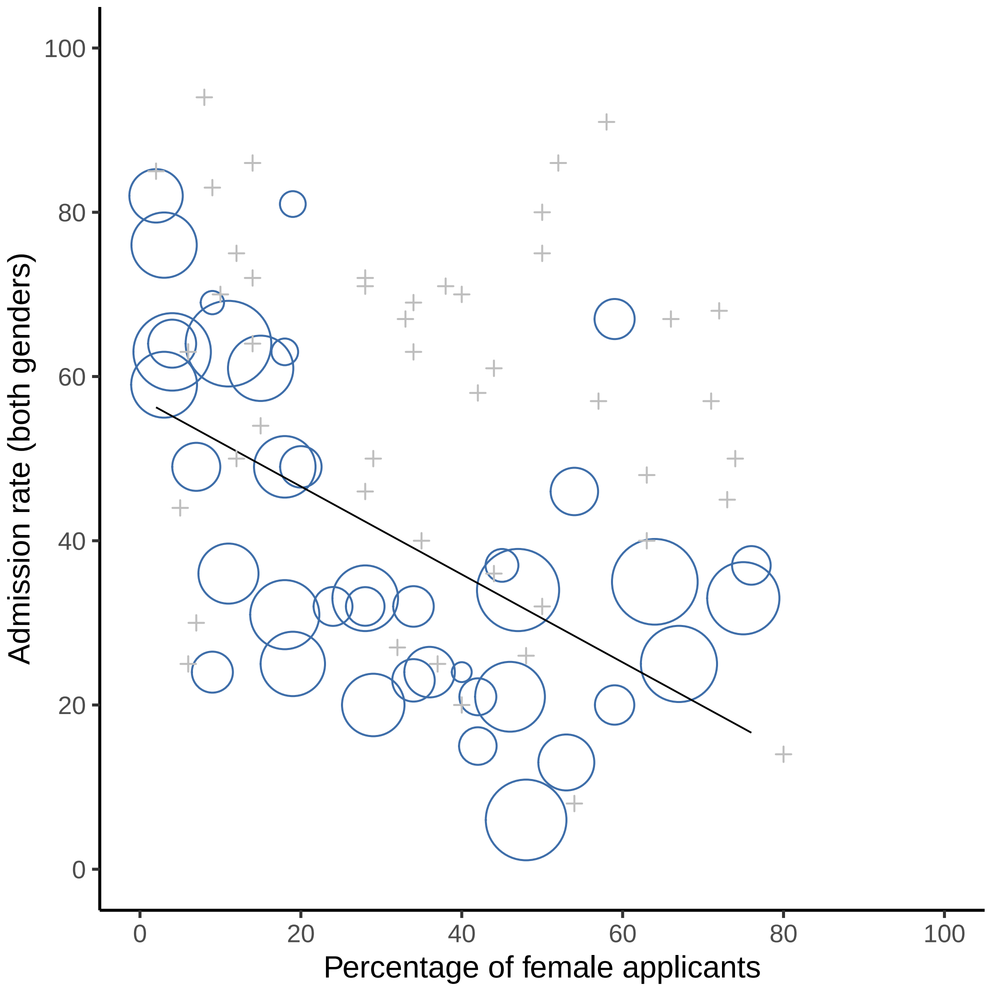


Figure 1.1: The Berkeley 1973 college admissions data. This figure plots the admission rate for the 85 departments that had at least one female applicant, as a function of the percentage of applicants that were female. The plot is a redrawing of Figure 1 from Bickel et al. (1975). Circles plot departments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

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.

1.3 Statistics in psychology

I hope that the discussion above helped explain why science in general is so focused on statistics. But I'm guessing that you have a lot more questions about what role statistics plays in psychology, and specifically why psychology classes always devote so many lectures to stats. So here's my attempt to answer a few of them…

Why does psychology have so much statistics?

To be perfectly honest, there's a few different reasons, some of which are better than others. The most important reason is that psychology is a statistical science. What I mean by that is that the "things" that we study are *people*. Real, complicated, gloriously messy, infuriatingly perverse people. The "things" of physics include objects like electrons, and while there are all sorts of complexities that arise in physics, electrons don't have minds of their own. They don't have opinions, they don't differ from each other in weird and arbitrary ways, they don't get bored in the middle of an experiment, and they don't get angry at the experimenter and then deliberately try to sabotage the data set (not that I've ever done that!). At a fundamental level psychology is harder than physics.[5] Basically, we teach statistics to you as psychologists because you need to be better at stats than physicists. There's actually a saying used sometimes in physics, to the effect that "if your experiment needs statistics, you should have done a better experiment". They have the luxury of being able to say that because their objects of study are pathetically simple in comparison to the vast mess that confronts social scientists. And it's not just psychology. Most social sciences are desperately reliant on statistics. Not because we're bad experimenters, but because we've picked a harder problem to solve. We teach you stats because you really, really need it.

Can't someone else do the statistics?

To some extent, but not completely. It's true that you don't need to become a fully trained statistician just to do psychology, but you do need to reach a certain level of statistical competence. In my view, there's three reasons that every psychological researcher ought to be able to do basic statistics:

- Firstly, there's the fundamental reason: statistics is deeply intertwined with research design. If you want to be good at designing psychological studies, you need to at the very least understand the basics of stats.
- Secondly, if you want to be good at the psychological side of the research, then you need to be able to understand the psychological literature, right? But almost every paper in the psychological literature reports the results of statistical analyses. So if you really want to understand the psychology, you need to be able to understand what other people did with their data. And that means understanding a certain amount of statistics.
- Thirdly, there's a big practical problem with being dependent on other people to do all your statistics: statistical analysis is *expensive*. If you ever get bored and

[5] Which might explain why physics is just a teensy bit further advanced as a science than we are.

> want to look up how much the Australian government charges for university fees, you'll notice something interesting: statistics is designated as a "national priority" category, and so the fees are much, much lower than for any other area of study. This is because there's a massive shortage of statisticians out there. So, from your perspective as a psychological researcher, the laws of supply and demand aren't exactly on your side here! As a result, in almost any real-life situation where you want to do psychological research, the cruel facts will be that you don't have enough money to afford a statistician. So the economics of the situation mean that you have to be pretty self-sufficient.

Note that a lot of these reasons generalise beyond researchers. If you want to be a practicing psychologist and stay on top of the field, it helps to be able to read the scientific literature, which relies pretty heavily on statistics.

I don't care about jobs, research or clinical work. Do I need statistics?

Okay, now you're just messing with me. Still, I think it should matter to you too. Statistics should matter to you in the same way that statistics should matter to *everyone*. We live in the 21st century, and data are *everywhere*. Frankly, given the world in which we live these days, a basic knowledge of statistics is pretty damn close to a survival tool! Which is the topic of the next section.

1.4 Statistics in everyday life

> *"We are drowning in information,*
> *but we are starved for knowledge"*
> - Various authors, original probably John Naisbitt

When I started writing up my lecture notes I took the 20 most recent news articles posted to the ABC News website. Of those 20 articles, it turned out that 8 of them involved a discussion of something that I would call a statistical topic and 6 of those made a mistake. The most common error, if you're curious, was failing to report baseline data (e.g., the article mentions that 5% of people in situation X have some characteristic Y, but doesn't say how common the characteristic is for everyone else!). The point I'm trying to make here isn't that journalists are bad at statistics (though they almost always are), it's that a basic knowledge of statistics is very helpful for trying to figure out when someone else is either making a mistake or even lying to you. In fact, one of the biggest things that a knowledge of statistics does to you is cause you to get angry at the newspaper or the internet on a far more frequent basis. You can find a good example of this in Section 4.1.5 in Chapter 4. In later versions of this book I'll try to include more anecdotes along those lines.

1.5 There's more to research methods than statistics

So far, most of what I've talked about is statistics, and so you'd be forgiven for thinking that statistics is all I care about in life. To be fair, you wouldn't be far wrong, but research methodology is a broader concept than statistics. So most research methods courses will cover a lot of topics that relate much more to the pragmatics of research design, and in particular the issues that you encounter when trying to do research with humans. However, about 99% of student fears relate to the statistics part of the course, so I've focused on the stats in this discussion, and hopefully I've convinced you that statistics matters, and more importantly, that it's not to be feared. That being said, it's pretty typical for introductory research methods classes to be very stats-heavy. This is not (usually) because the lecturers are evil people. Quite the contrary, in fact. Introductory classes focus a lot on the statistics because you almost always find yourself needing statistics before you need the other research methods training. Why? Because almost all of your assignments in other classes will rely on statistical training, to a much greater extent than they rely on other methodological tools. It's not common for undergraduate assignments to require you to design your own study from the ground up (in which case you would need to know a lot about research design), but it *is* common for assignments to ask you to analyse and interpret data that were collected in a study that someone else designed (in which case you need statistics). In that sense, from the perspective of allowing you to do well in all your other classes, the statistics is more urgent.

But note that "urgent" is different from "important" – they both matter. I really do want to stress that research design is just as important as data analysis, and this book does spend a fair amount of time on it. However, while statistics has a kind of universality, and provides a set of core tools that are useful for most types of psychological research, the research methods side isn't quite so universal. There are some general principles that everyone should think about, but a lot of research design is very idiosyncratic, and is specific to the area of research that you want to engage in. To the extent that it's the details that matter, those details don't usually show up in an introductory stats and research methods class.

Chapter 2

A brief introduction to research design

> *"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 Fisher[1]

In this chapter, we're going to start thinking about the basic ideas that go into designing a study, collecting data, checking whether your data collection works, and so on. It won't give you enough information to allow you to design studies of your own, but it will give you a lot of the basic tools that you need to assess the studies done by other people. However, since the focus of this book is much more on data analysis than on data collection, I'm only giving a very brief overview. Note that this chapter 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 & Stanley (1963) and Stevens (1946) for the discussion of scales of measurement.

2.1 Introduction to psychological measurement

The first thing to understand is data collection can be thought of as a kind of **measurement**. That is, what we're trying to do here is measure something about human behaviour or the human mind. What do I mean by "measurement"?

[1]Presidential Address to the First Indian Statistical Congress, 1938. Source: https://en.wikiquote.org/wiki/Ronald_Fisher

2.1.1 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 *female*.

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 *Klinfelter'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. 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, for example birth or death certificates. This is a time consuming and frustrating endeavour, but it has its uses (e.g., if the person is now dead).

2.1.2 Operationalisation: defining your measurement

All of the ideas discussed in the previous section 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 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, or hours? 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 specific options and instead let them answer in their own words? And if you open up the set of possible values to include all verbal response, 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.

2.2 Scales of measurement

As the previous section indicates, the outcome of a psychological measurement is called a variable. But not all variables are of the same qualitative type and so it's useful to understand what types there are. A very useful concept for distinguishing between different types of variables is what's known as **scales of measurement**.

2.2.1 Nominal scale

A **nominal scale** variable (also referred to as a **categorical** variable) is one in which there is no particular relationship between the different possibilities. For these kinds of variables it doesn't make any sense to say that one of them is "bigger" or "better" than any other one, and it absolutely doesn't make any sense to average them. The classic example for this is "eye colour". Eyes can be blue, green or brown, amongst other possibilities, but none of them is any "bigger" than any other one. As a result, it would feel really weird to talk about an "average eye colour". Similarly, gender is nominal too: male isn't better or worse than female. Neither does it make sense to try to talk about an “average gender”. In short, nominal scale variables are those for which the only thing you can say about the different possibilities is that they are different. That’s it.

Let’s take a slightly closer look at this. Suppose I was doing research on how people commute to and from work. One variable I would have to measure would be what kind of transportation people use to get to work. This “transport type” variable could have quite a few possible values, including: “train”, “bus”, “car”, “bicycle”. For now, let’s suppose that these four are the only possibilities. Then imagine that I ask 100 people how they got to work today, with this result (Table 2.1).

Table 2.1: How did 100 people get to work today

Transportation	Number of people
(1) Train	12
(2) Bus	30
(3) Car	48
(4) Bicycle	10

So, what’s the average transportation type? Obviously, the answer here is that there isn’t one. It’s a silly question to ask. You can say that travel by car is the most popular method, and travel by train is the least popular method, but that’s about all. Similarly, notice that the order in which I list the options isn’t very interesting. I could have chosen to display the data like in Table 2.2.

Table 2.2: How did 100 people get to work today, a different view

Transportation	Number of people
(3) Car	48
(1) Train	12
(4) Bicycle	10
(2) Bus	30

…and nothing really changes.

2.2.2 Ordinal scale

Ordinal scale variables have a bit more structure than nominal scale variables, but not by a lot. An ordinal scale variable is one in which there is a natural, meaningful way to order the different possibilities, but you can’t do anything else. The usual example given of an ordinal variable is “finishing position in a race”. You *can* say that the person who finished first was faster than the person who finished second, but you *do not* know how much faster. As a consequence we know that 1st > 2nd, and we know that 2nd > 3rd, but the difference between 1st and 2nd might be much larger than the difference between 2nd and 3rd.

Here’s a more psychologically interesting example. Suppose I’m interested in people’s attitudes to climate change. I then go and ask some people to pick the statement (from four listed statements) that most closely matches their beliefs:

1. Temperatures are rising because of human activity
2. Temperatures are rising but we don't know why
3. Temperatures are rising but not because of humans
4. Temperatures are not rising

Notice that these four statements actually do have a natural ordering, in terms of "the extent to which they agree with the current science". Statement 1 is a close match, statement 2 is a reasonable match, statement 3 isn't a very good match, and statement 4 is in strong opposition to current science. So, in terms of the thing I'm interested in (the extent to which people endorse the science), I can order the items as $1 > 2 > 3 > 4$. Since this ordering exists, it would be very weird to list the options like this...

1. Temperatures are rising but not because of humans
2. Temperatures are rising because of human activity
3. Temperatures are not rising
4. Temperatures are rising but we don't know why

...because it seems to violate the natural "structure" to the question.

So, let's suppose I asked 100 people these questions, and got the answers shown in Table 2.3.

Table 2.3: Attitudes to climate change

Response	**Number**
(1) Temperatures are rising because of human activity	51
(2) Temperatures are rising but we do not know why	20
(3) Temperatures are rising but not because of humans	10
(4) Temperatures are not rising	19

When analysing these data it seems quite reasonable to try to group (1), (2) and (3) together, and say that 81 out of 100 people were willing to at *least partially* endorse the science. And it's also quite reasonable to group (2), (3) and (4) together and say that 49 out of 100 people registered *at least some disagreement* with the dominant scientific view. However, it would be entirely bizarre to try to group (1), (2) and (4) together and say that 90 out of 100 people said... what? There's nothing sensible that allows you to group those responses together at all.

That said, notice that while we *can* use the natural ordering of these items to construct sensible groupings, what we can't do is average them. For instance, in my simple example here, the "average" response to the question is 1.97. If you can tell me what that means I'd love to know, because it seems like gibberish to me!

2.2.3 Interval scale

In contrast to nominal and ordinal scale variables, **interval scale** and ratio scale variables are variables for which the numerical value is genuinely meaningful. In the case of interval scale variables the *differences* between the numbers are interpretable, but the variable doesn't have a "natural" zero value. A good example of an interval scale variable is measuring temperature in degrees celsius. For instance, if it was 15° yesterday and 18° today, then the 3° difference between the two is genuinely meaningful. Moreover, that 3° difference is *exactly the same* as the 3° difference between 7° and 10°. In short, addition and subtraction are meaningful for interval scale variables.[2]

However, notice that the 0° does not mean "no temperature at all". It actually means "the temperature at which water freezes", which is pretty arbitrary. As a consequence it becomes pointless to try to multiply and divide temperatures. It is wrong to say that 20° is twice as hot as 10°, just as it is weird and meaningless to try to claim that 20° is negative two times as hot as -10°.

Again, lets look at a more psychological example. Suppose I'm interested in looking at how the attitudes of first-year university students have changed over time. Obviously, I'm going to want to record the year in which each student started. This is an interval scale variable. A student who started in 2003 did arrive 5 years before a student who started in 2008. However, it would be completely daft for me to divide 2008 by 2003 and say that the second student started "1.0024 times later" than the first one. That doesn't make any sense at all.

2.2.4 Ratio scale

The fourth and final type of variable to consider is a **ratio scale** variable, in which zero really means zero, and it's okay to multiply and divide. A good psychological example of a ratio scale variable is response time (RT). In a lot of tasks it's very common to record the amount of time somebody takes to solve a problem or answer a question, because it's an indicator of how difficult the task is. Suppose that Alan takes 2.3 seconds to respond to a question, whereas Ben takes 3.1 seconds. As with an interval scale variable, addition and subtraction are both meaningful here. Ben really did take 3.1 - 2.3 = 0.8 seconds longer than Alan did. However, notice that multiplication and division also make sense here too: Ben took 3.1/2.3 = 1.35 times as long as Alan did to answer the question. And the reason why you can do this is that, for a ratio scale variable such as RT, "zero seconds" really does mean "no time at all".

2.2.5 Continuous versus discrete variables

There's a second kind of distinction that you need to be aware of, regarding what types of variables you can run into. This is the distinction between continuous variables and discrete variables (Table 2.4). The difference between these is as follows:

[2] Actually, I've been informed by readers with greater physics knowledge than I that temperature isn't strictly an interval scale, in the sense that the amount of energy required to heat something up by 3° depends on its current temperature. So in the sense that physicists care about, temperature isn't actually an interval scale. But it still makes a cute example so I'm going to ignore this little inconvenient truth.

- A **continuous variable** is one in which, for any two values that you can think of, it's always logically possible to have another value in between.
- A **discrete variable** is, in effect, a variable that isn't continuous. For a discrete variable it's sometimes the case that there's nothing in the middle.

Table 2.4: The relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible

	continuous	**discrete**
nominal		✓
ordinal		✓
interval	✓	✓
ratio	✓	✓

These definitions probably seem a bit abstract, but they're pretty simple once you see some examples. For instance, response time is continuous. If Alan takes 3.1 seconds and Ben takes 2.3 seconds to respond to a question, then Cameron's response time will lie in between if he took 3.0 seconds. And of course it would also be possible for David to take 3.031 seconds to respond, meaning that his RT would lie in between Cameron's and Alan's. And while in practice it might be impossible to measure RT that precisely, it's certainly possible in principle. Because we can always find a new value for RT in between any two other ones we regard RT as a continuous measure.

Discrete variables occur when this rule is violated. For example, nominal scale variables are always discrete. There isn't a type of transportation that falls "in between" trains and bicycles, not in the strict mathematical way that 2.3 falls in between 2 and 3. So transportation type is discrete. Similarly, ordinal scale variables are always discrete. Although "2nd place" does fall between "1st place" and "3rd place", there's nothing that can logically fall in between "1st place" and "2nd place". Interval scale and ratio scale variables can go either way. As we saw above, response time (a ratio scale variable) is continuous. Temperature in degrees celsius (an interval scale variable) is also continuous. However, the year you went to school (an interval scale variable) is discrete. There's no year in between 2002 and 2003. The number of questions you get right on a true-or-false test (a ratio scale variable) is also discrete. Since a true-or-false question doesn't allow you to be "partially correct", there's nothing in between 5/10 and 6/10. Table 2.4 summarises the relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible. I'm trying to hammer this point home, because (a) some textbooks get this wrong, and (b) people very often say things like "discrete variable" when they mean "nominal scale variable". It's very unfortunate.

2.2.6 Some complexities

Okay, I know you're going to be shocked to hear this, but the real world is much messier than this little classification scheme suggests. Very few variables in real life actually fall into these nice neat categories, so you need to be kind of careful not to treat the scales of measurement as if they were hard and fast rules. It doesn't work like that. They're guidelines, intended to help you think about the situations in which you should treat different variables differently. Nothing more.

So let's take a classic example, maybe *the* classic example, of a psychological measurement tool: the **Likert scale**. The humble Likert scale is the bread and butter tool of all survey design. You yourself have filled out hundreds, maybe thousands, of them and odds are you've even used one yourself. Suppose we have a survey question that looks like this:

> Which of the following best describes your opinion of the statement that "all pirates are freaking awesome"?

and then the options presented to the participant are these:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

This set of items is an example of a 5-point Likert scale, in which people are asked to choose among one of several (in this case 5) clearly ordered possibilities, generally with a verbal descriptor given in each case. However, it's not necessary that all items are explicitly described. This is a perfectly good example of a 5-point Likert scale too:

1. Strongly disagree
2.
3.
4.
5. Strongly agree

Likert scales are very handy, if somewhat limited, tools. The question is what kind of variable are they? They're obviously discrete, since you can't give a response of 2.5. They're obviously not nominal scale, since the items are ordered; and they're not ratio scale either, since there's no natural zero.

But are they ordinal scale or interval scale? One argument says that we can't really prove that the difference between "strongly agree" and "agree" is of the same size as the difference between "agree" and "neither agree nor disagree". In fact, in everyday life it's pretty obvious that they're not the same at all. So this suggests that we ought to treat Likert scales as ordinal variables. On the other hand, in practice most participants do seem to take the whole "on a scale from 1 to 5" part fairly seriously, and they tend to act as if the differences between the five response options were fairly similar to one another. As a consequence, a lot of researchers treat Likert scale data as interval scale.[3] It's not interval scale, but in practice it's close enough that we usually think of it as being **quasi-interval scale**.

[3] Ah, psychology... never an easy answer to anything!

2.3 Assessing the reliability of a measurement

At this point we've thought a little bit about how to operationalise a theoretical construct and thereby create a psychological measure. And we've seen that by applying psychological measures we end up with variables, which can come in many different types. At this point, we should start discussing the obvious question: is the measurement any good? We'll do this in terms of two related ideas: *reliability* and *validity*. Put simply, the **reliability** of a measure tells you how precisely you are measuring something, whereas the **validity** of a measure tells you how accurate the measure is. In this section we'll talk about reliability; we'll talk about validity in the section on Assessing the validity of a study.

Reliability is actually a very simple concept. It refers to the repeatability or consistency of your measurement. The measurement of my weight by means of a "bathroom scale" is very reliable. If I step on and off the scales over and over again, it'll keep giving me the same answer. Measuring my intelligence by means of "asking my mum" is very unreliable. Some days she tells me I'm a bit thick, and other days she tells me I'm a complete idiot. Notice that this concept of reliability is different to the question of whether the measurements are correct (the correctness of a measurement relates to it's validity). If I'm holding a sack of potatos when I step on and off the bathroom scales the measurement will still be reliable: it will always give me the same answer. However, this highly reliable answer doesn't match up to my true weight at all, therefore it's wrong. In technical terms, this is a reliable but invalid measurement. Similarly, whilst my mum's estimate of my intelligence is a bit unreliable, she might be right. Maybe I'm just not too bright, and so while her estimate of my intelligence fluctuates pretty wildly from day to day, it's basically right. That would be an unreliable but valid measure. Of course, if my mum's estimates are too unreliable it's going to be very hard to figure out which one of her many claims about my intelligence is actually the right one. To some extent, then, a very unreliable measure tends to end up being invalid for practical purposes; so much so that many people would say that reliability is necessary (but not sufficient) to ensure validity.

Okay, now that we're clear on the distinction between reliability and validity, let's have a think about the different ways in which we might measure reliability:

- **Test-retest reliability**. This relates to consistency over time. If we repeat the measurement at a later date do we get the same answer?
- **Inter-rater reliability**. This relates to consistency across people. If someone else repeats the measurement (e.g., someone else rates my intelligence) will they produce the same answer?
- **Parallel forms reliability**. This relates to consistency across theoretically-equivalent measurements. If I use a different set of bathroom scales to measure my weight does it give the same answer?
- **Internal consistency reliability**. If a measurement is constructed from lots of different parts that perform similar functions (e.g., a personality questionnaire result is added up across several questions) do the individual parts tend to give similar answers. We'll look at this particular form of reliability later in the book, in **?@sec-Internal-consistency-reliability-analysis**.

Not all measurements need to possess all forms of reliability. For instance, educational assessment can be thought of as a form of measurement. One of the subjects that I teach, *Computational Cognitive Science*, has an assessment structure that has a research component and an exam component (plus other things). The exam component is *intended* to measure something different from the research component, so the assessment as a whole has low internal consistency. However, within the exam there are several questions that are intended to (approximately) measure the same things, and those tend to produce similar outcomes. So the exam on its own has a fairly high internal consistency. Which is as it should be. You should only demand reliability in those situations where you want to be measuring the same thing!

2.4 The "role" of variables: predictors and outcomes

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. First, 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.

When we are 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 this book 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 2.5.

Table 2.5: Variable distinctions

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

[4]Annoyingly 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 "response variable" is sometimes used where I've used "outcome". Sigh. This sort of terminological confusion is very common, I'm afraid.

2.5 Experimental and non-experimental research

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.

2.5.1 Experimental research

The key feature 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) but 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 possibility 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, but for now it's enough to say that what randomisation does is minimise (but not eliminate) the possibility 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, and not by smoking per se. In technical terms these other things (e.g. diet) are called "confounders", and we'll talk about those in just a moment.

In the meantime, let's 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 young non-smokers 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, we can feel pretty confident that (a) smoking does cause cancer and (b) we're murderers.

2.5.2 Non-experimental research

Non-experimental research is a broad term that covers "any study in which the researcher doesn't have as much control as they do in an experiment". Obviously, control is something that scientists like to have, but as the previous example illustrates there are lots of situations in which you can't or shouldn't try to obtain that control. Since it's grossly unethical (and almost certainly criminal) to force people to smoke in order to find out if they get cancer, this is a good example of a situation in which you really shouldn't try to obtain experimental control. But there are other reasons too. Even leaving aside the ethical issues, our "smoking experiment" does have a few other issues. For instance, when I suggested that we "force" half of the people to become smokers, I was talking about *starting* with a sample of non-smokers, and then forcing them to become smokers. While this sounds like the kind of solid, evil experimental design that a mad scientist would love, it might not be a very sound way of investigating the effect in the real world. For instance, suppose that smoking only causes lung cancer when people have poor diets, and suppose also that people who normally smoke do tend to have poor diets. However, since the "smokers" in our experiment aren't "natural" smokers (i.e., we forced non-smokers to become smokers, but they didn't take on all of the other normal, real-life characteristics that smokers might tend to possess) they probably have better diets. As such, in this silly example they wouldn't get lung cancer and our experiment will fail, because it violates the structure of the "natural" world (the technical name for this is an "artefactual" result).

One distinction worth making between two types of non-experimental research is the difference between **quasi-experimental research** and **case studies**. The example I discussed earlier, in which we wanted to examine incidence of lung cancer among smokers and non-smokers without trying to control who smokes and who doesn't, is a quasi-experimental design. That is, it's the same as an experiment but we don't control the predictors (IVs). We can still use statistics to analyse the results, but we have to be a lot more careful and circumspect.

The alternative approach, case studies, aims to provide a very detailed description of one or a few instances. In general, you can't use statistics to analyse the results of case studies and it's usually very hard to draw any general conclusions about "people in general" from a few isolated examples. However, case studies are very useful in some situations. Firstly, there are situations where you don't have any alternative. Neuropsychology has this issue a lot. Sometimes, you just can't find a lot of people with brain damage in a specific brain area, so the only thing you can do is describe those cases that you do have in as much detail and with as much care as you can. However, there's also some genuine advantages to case studies. Because you don't have as many people to study you have the ability to invest lots of time and effort trying to understand the specific factors at play in each case. This is a very valuable thing to do. As a consequence, case studies can complement the more statistically-oriented approaches that you see in experimental and quasi-experimental designs. We won't talk much about case studies in this book, but they are nevertheless very valuable tools!

2.6 Assessing the validity of a study

More than any other thing, a scientist wants their research to be "valid". The conceptual idea behind **validity** is very simple. Can you trust the results of your study? If not, the study is invalid. However, whilst it's easy to state, in practice it's much harder to check validity than it is to check reliability. And in all honesty, there's no precise, clearly agreed upon notion of what validity actually is. In fact, there are lots of different kinds of validity, each of which raises its own issues. And not all forms of validity are relevant to all studies. I'm going to talk about five different types of validity:

- Internal validity.
- External validity.
- Construct validity.
- Face validity.
- Ecological validity.

First, a quick guide as to what matters here. (1) Internal and external validity are the most important, since they tie directly to the fundamental question of whether your study really works. (2) Construct validity asks whether you're measuring what you think you are. (3) Face validity isn't terribly important except insofar as you care about "appearances". (4) Ecological validity is a special case of face validity that corresponds to a kind of appearance that you might care about a lot.

2.6.1 Internal validity

Internal validity refers to the extent to which you are able to draw the correct conclusions about the causal relationships between variables. It's called "internal" because it refers to the relationships between things "inside" the study. Let's illustrate the concept with a simple example. Suppose you're interested in finding out whether a university education makes you write better. To do so, you get a group of first year students, ask them to write a 1000 word essay, and count the number of spelling and grammatical errors they make. Then you find some third-year students, who obviously have had more of a university education than the first-years, and repeat the exercise. And let's suppose it turns out that the third-year students produce fewer errors. And so you conclude that a university education improves writing skills. Right? Except that the big problem with this experiment is that the third-year students are older and they've had more experience with writing things. So it's hard to know for sure what the causal relationship is. Do older people write better? Or people who have had more writing experience? Or people who have had more education? Which of the above is the true cause of the superior performance of the third-years? Age? Experience? Education? You can't tell. This is an example of a failure of internal validity, because your study doesn't properly tease apart the causal relationships between the different variables.

2.6.2 External validity

External validity relates to the **generalisability** or **applicability** of your findings. That is, to what extent do you expect to see the same pattern of results in "real life" as you saw in your study. To put it a bit more precisely, any study that you do in psychology will involve a fairly specific set of questions or tasks, will occur in a specific environment, and will involve participants that are drawn from a particular subgroup (disappointingly often it is college students!). So, if it turns out that the results don't actually generalise or apply to people and situations beyond the ones that you studied, then what you've got is a lack of external validity.

The classic example of this issue is the fact that a very large proportion of studies in psychology will use undergraduate psychology students as the participants. Obviously, however, the researchers don't care *only* about psychology students. They care about people in general. Given that, a study that uses only psychology students as participants always carries a risk of lacking external validity. That is, if there's something "special" about psychology students that makes them different to the general population in some relevant respect, then we may start worrying about a lack of external validity.

That said, it is absolutely critical to realise that a study that uses only psychology students does not necessarily have a problem with external validity. I'll talk about this again later, but it's such a common mistake that I'm going to mention it here. The external validity of a study is threatened by the choice of population if (a) the population from which you sample your participants is very narrow (e.g., psychology students), and (b) the narrow population that you sampled from is systematically different from the general population in some respect that is relevant to the *psychological phenomenon that you intend to study.* The italicised part is the bit that lots of people forget. It is true that psychology undergraduates differ from the general population in lots of ways, and so a study that uses only psychology students may have problems with external validity. However, if those differences aren't very relevant to the phenomenon that you're studying, then there's nothing to worry about. To make this a bit more concrete here are two extreme examples:

- You want to measure "attitudes of the general public towards psychotherapy", but all of your participants are psychology students. This study would almost certainly have a problem with external validity.
- You want to measure the effectiveness of a visual illusion, and your participants are all psychology students. This study is unlikely to have a problem with external validity.

Having just spent the last couple of paragraphs focusing on the choice of participants, since that's a big issue that everyone tends to worry most about, it's worth remembering that external validity is a broader concept. The following are also examples of things that might pose a threat to external validity, depending on what kind of study you're doing:

- People might answer a "psychology questionnaire" in a manner that doesn't reflect what they would do in real life.
- Your lab experiment on (say) "human learning" has a different structure to the learning problems people face in real life.

2.6.3 Construct validity

Construct validity is basically a question of whether you're measuring what you want to be measuring. A measurement has good construct validity if it is actually measuring the correct theoretical construct, and bad construct validity if it doesn't. To give a very simple (if ridiculous) example, suppose I'm trying to investigate the rates with which university students cheat on their exams. And the way I attempt to measure it is by asking the cheating students to stand up in the lecture theatre so that I can count them. When I do this with a class of 300 students 0 people claim to be cheaters. So I therefore conclude that the proportion of cheaters in my class is 0%. Clearly this is a bit ridiculous. But the point here is not that this is a very deep methodological example, but rather to explain what construct validity is. The problem with my measure is that while I'm trying to measure "the proportion of people who cheat" what I'm actually measuring is "the proportion of people stupid enough to own up to cheating, or bloody minded enough to pretend that they do". Obviously, these aren't the same thing! So my study has gone wrong, because my measurement has very poor construct validity.

2.6.4 Face validity

Face validity simply refers to whether or not a measure "looks like" it's doing what it's supposed to, nothing more. If I design a test of intelligence, and people look at it and they say "no, that test doesn't measure intelligence", then the measure lacks face validity. It's as simple as that. Obviously, face validity isn't very important from a pure scientific perspective. After all, what we care about is whether or not the measure *actually* does what it's supposed to do, not whether it *looks like* it does what it's supposed to do. As a consequence, we generally don't care very much about face validity. That said, the concept of face validity serves three useful pragmatic purposes:

- Sometimes, an experienced scientist will have a "hunch" that a particular measure won't work. While these sorts of hunches have no strict evidentiary value, it's often worth paying attention to them. Because often times people have knowledge that they can't quite verbalise, there might be something to worry about even if you can't quite say why. In other words, when someone you trust criticises the face validity of your study, it's worth taking the time to think more carefully about your design to see if you can think of reasons why it might go awry. Mind you, if you don't find any reason for concern, then you should probably not worry. After all, face validity really doesn't matter very much.
- Often (very often), completely uninformed people will also have a "hunch" that your research is crap. And they'll criticise it on the internet or something. On close inspection you may notice that these criticisms are actually focused entirely on how the study "looks", but not on anything deeper. The concept of face validity is useful for gently explaining to people that they need to substantiate their arguments further.
- Expanding on the last point, if the beliefs of untrained people are critical (e.g., this is often the case for applied research where you actually want to convince policy makers of something or other) then you have to care about face validity. Simply because, whether you like it or not, a lot of people will use face validity as a proxy for real validity. If you want the government to change a law on scientific

psychological grounds, then it won't matter how good your studies "really" are. If they lack face validity you'll find that politicians ignore you. Of course, it's somewhat unfair that policy often depends more on appearance than fact, but that's how things go.

2.6.5 Ecological validity

Ecological validity is a different notion of validity, which is similar to external validity, but less important. The idea is that, in order to be ecologically valid, the entire set up of the study should closely approximate the real-world scenario that is being investigated. In a sense, ecological validity is a kind of face validity. It relates mostly to whether the study "looks" right, but with a bit more rigour to it. To be ecologically valid the study has to look right in a fairly specific way. The idea behind it is the intuition that a study that is ecologically valid is more likely to be externally valid. It's no guarantee, of course. But the nice thing about ecological validity is that it's much easier to check whether a study is ecologically valid than it is to check whether a study is externally valid. A simple example would be eyewitness identification studies. Most of these studies tend to be done in a university setting, often with a fairly simple array of faces to look at, rather than a line up. The length of time between seeing the "criminal" and being asked to identify the suspect in the "line up" is usually shorter. The "crime" isn't real so there's no chance of the witness being scared, and there are no police officers present so there's not as much chance of feeling pressured. These things all mean that the study definitely lacks ecological validity. They might (but might not) mean that it also lacks external validity.

2.7 Confounders, artefacts and other threats to validity

If we look at the issue of validity in the most general fashion the two biggest worries that we have are *confounders* and *artefacts*. These two terms are defined in the following way:

- **Confounder:** A confounder is an additional, often unmeasured variable[5] that turns out to be related to both the predictors and the outcome. The existence of confounders threatens the internal validity of the study because you can't tell whether the predictor causes the outcome, or if the confounding variable causes it.
- **Artefact:** A result is said to be "artefactual" if it only holds in the special situation that you happened to test in your study. The possibility that your result is an artefact poses a threat to your external validity, because it raises the possibility that you can't generalise or apply your results to the actual population that you care about.

[5] The reason why I say that it's unmeasured is that if you have measured it, then you can use some fancy statistical tricks to deal with the confounder. Because of the existence of these statistical solutions to the problem of confounders, we often refer to a confounder that we have measured and dealt with as a covariate. Dealing with covariates is a more advanced topic, but I thought I'd mention it in passing since it's kind of comforting to at least know that this stuff exists.

As a general rule confounders are a bigger concern for non-experimental studies, precisely because they're not proper experiments. By definition, you're leaving lots of things uncontrolled, so there's a lot of scope for confounders being present in your study. Experimental research tends to be much less vulnerable to confounders. The more control you have over what happens during the study, the more you can prevent confounders from affecting the results. With random allocation, for example, confounders are distributed randomly, and evenly, between different groups.

However, there are always swings and roundabouts and when we start thinking about artefacts rather than confounders the shoe is very firmly on the other foot. For the most part, artefactual results tend to be more of a concern for experimental studies than for non-experimental studies. To see this, it helps to realise that the reason that a lot of studies are non-experimental is precisely because what the researcher is trying to do is examine human behaviour in a more naturalistic context. By working in a more real-world context you lose experimental control (making yourself vulnerable to confounders), but because you tend to be studying human psychology "in the wild" you reduce the chances of getting an artefactual result. Or, to put it another way, when you take psychology out of the wild and bring it into the lab (which we usually have to do to gain our experimental control), you always run the risk of accidentally studying something different to what you wanted to study.

Be warned though. The above is a rough guide only. It's absolutely possible to have confounders in an experiment, and to get artefactual results with non-experimental studies. This can happen for all sorts of reasons, not least of which is experimenter or researcher error. In practice, it's really hard to think everything through ahead of time and even very good researchers make mistakes.

Although there's a sense in which almost any threat to validity can be characterised as a confounder or an artefact, they're pretty vague concepts. So let's have a look at some of the most common examples.

2.7.1 History effects

History effects refer to the possibility that specific events may occur during the study that might influence the outcome measure. For instance, something might happen in between a pretest and a post-test. Or in-between testing participant 23 and participant 24. Alternatively, it might be that you're looking at a paper from an older study that was perfectly valid for its time, but the world has changed enough since then that the conclusions are no longer trustworthy. Examples of things that would count as history effects are:

- You're interested in how people think about risk and uncertainty. You started your data collection in December 2010. But finding participants and collecting data takes time, so you're still finding new people in February 2011. Unfortunately for you (and even more unfortunately for others), the Queensland floods occurred in January 2011 causing billions of dollars of damage and killing many people. Not surprisingly, the people tested in February 2011 express quite different beliefs about handling risk than the people tested in December 2010. Which (if any) of these reflects the "true" beliefs of participants? I think the answer is probably both. The Queensland floods genuinely changed the beliefs of the Australian public, though possibly only temporarily. The key thing here is that the "history" of the people tested in February is quite different to people tested in December.

- You're testing the psychological effects of a new anti-anxiety drug. So what you do is measure anxiety before administering the drug (e.g., by self-report, and taking physiological measures). Then you administer the drug, and afterwards you take the same measures. In the interim however, because your lab is in Los Angeles, there's an earthquake which increases the anxiety of the participants.

2.7.2 Maturation effects

As with history effects, **maturational effects** are fundamentally about change over time. However, maturation effects aren't in response to specific events. Rather, they relate to how people change on their own over time. We get older, we get tired, we get bored, etc. Some examples of maturation effects are:

- When doing developmental psychology research you need to be aware that children grow up quite rapidly. So, suppose that you want to find out whether some educational trick helps with vocabulary size among 3 year olds. One thing that you need to be aware of is that the vocabulary size of children that age is growing at an incredible rate (multiple words per day) all on its own. If you design your study without taking this maturational effect into account, then you won't be able to tell if your educational trick works.
- When running a very long experiment in the lab (say, something that lasts for 3 hours) it's very likely that people will begin to get bored and tired, and that this maturational effect will cause performance to decline regardless of anything else going on in the experiment

2.7.3 Repeated testing effects

An important type of history effect is the effect of **repeated testing**. Suppose I want to take two measurements of some psychological construct (e.g., anxiety). One thing I might be worried about is if the first measurement has an effect on the second measurement. In other words, this is a history effect in which the "event" that influences the second measurement is the first measurement itself! This is not at all uncommon. Examples of this include:

- Learning and practice: e.g., "intelligence" at time 2 might appear to go up relative to time 1 because participants learned the general rules of how to solve "intelligence-test-style" questions during the first testing session.
- Familiarity with the testing situation: e.g., if people are nervous at time 1, this might make performance go down. But after sitting through the first testing situation they might calm down a lot precisely because they've seen what the testing looks like.
- Auxiliary changes caused by testing: e.g., if a questionnaire assessing mood is boring then mood rating at measurement time 2 is more likely to be "bored" precisely because of the boring measurement made at time 1.

2.7.4 Selection bias

Selection bias is a pretty broad term. Suppose that you're running an experiment with two groups of participants where each group gets a different "treatment", and you want to see if the different treatments lead to different outcomes. However, suppose that, despite your best efforts, you've ended up with a gender imbalance across groups (say, group A has 80% females and group B has 50% females). It might sound like this could never happen but, trust me, it can. This is an example of a selection bias, in which the people "selected into" the two groups have different characteristics. If any of those characteristics turns out to be relevant (say, your treatment works better on females than males) then you're in a lot of trouble.

2.7.5 Differential attrition

When thinking about the effects of attrition, it is sometimes helpful to distinguish between two different types. The first is **homogeneous attrition**, in which the attrition effect is the same for all groups, treatments or conditions. In the example I gave above, the attrition would be homogeneous if (and only if) the easily bored participants are dropping out of all of the conditions in my experiment at about the same rate. In general, the main effect of homogeneous attrition is likely to be that it makes your sample unrepresentative. As such, the biggest worry that you'll have is that the generalisability of the results decreases. In other words, you lose external validity.

The second type of attrition is **heterogeneous attrition**, in which the attrition effect is different for different groups. More often called **differential attrition**, this is a kind of selection bias that is caused by the study itself. Suppose that, for the first time ever in the history of psychology, I manage to find the perfectly balanced and representative sample of people. I start running "Dani's incredibly long and tedious experiment" on my perfect sample but then, because my study is incredibly long and tedious, lots of people start dropping out. I can't stop this. Participants absolutely have the right to stop doing any experiment, any time, for whatever reason they feel like, and as researchers we are morally (and professionally) obliged to remind people that they do have this right. So, suppose that "Dani's incredibly long and tedious experiment" has a very high drop out rate. What do you suppose the odds are that this drop out is random? Answer: zero. Almost certainly the people who remain are more conscientious, more tolerant of boredom, etc., than those who leave. To the extent that (say) conscientiousness is relevant to the psychological phenomenon that I care about, this attrition can decrease the validity of my results.

Here's another example. Suppose I design my experiment with two conditions. In the "treatment" condition, the experimenter insults the participant and then gives them a questionnaire designed to measure obedience. In the "control" condition, the experimenter engages in a bit of pointless chitchat and then gives them the questionnaire. Leaving aside the questionable scientific merits and dubious ethics of such a study, let's have a think about what might go wrong here. As a general rule, when someone insults me to my face I tend to get much less co-operative. So, there's a pretty good chance that a lot more people are going to drop out of the treatment condition than the control condition. And this drop out isn't going to be random. The people most likely to drop out would probably be the people who don't care all that much about the importance of obediently sitting through the experiment. Since the most bloody minded and disobedient people all left the treatment group but not the control group, we've introduced a confounder: the people who actually took the questionnaire in the treatment group were already more likely to be dutiful and obedient than the people in the control group. In short, in this study insulting people doesn't make them more obedient. It makes the more disobedient people leave the experiment! The internal validity of this experiment is completely shot.

2.7.6 Non-response bias

Non-response bias is closely related to selection bias and to differential attrition. The simplest version of the problem goes like this. You mail out a survey to 1000 people but only 300 of them reply. The 300 people who replied are almost certainly not a random subsample. People who respond to surveys are systematically different to people who don't. This introduces a problem when trying to generalise from those 300 people who replied to the population at large, since you now have a very non-random sample. The issue of non-response bias is more general than this, though. Among the (say) 300 people that did respond to the survey, you might find that not everyone answers every question. If (say) 80 people chose not to answer one of your questions, does this introduce problems? As always, the answer is maybe. If the question that wasn't answered was on the last page of the questionnaire, and those 80 surveys were returned with the last page missing, there's a good chance that the missing data isn't a big deal; probably the pages just fell off. However, if the question that 80 people didn't answer was the most confrontational or invasive personal question in the questionnaire, then almost certainly you've got a problem. In essence, what you're dealing with here is what's called the problem of **missing data**. If the data that is missing was "lost" randomly, then it's not a big problem. If it's missing systematically, then it can be a big problem.

2.7.7 Regression to the mean

Regression to the mean refers to any situation where you select data based on an extreme value on some measure. Because the variable has natural variation it almost certainly means that when you take a subsequent measurement the later measurement will be less extreme than the first one, purely by chance.

Here's an example. Suppose I'm interested in whether a psychology education has an adverse effect on very smart kids. To do this, I find the 20 Psychology I students with the best high school grades and look at how well they're doing at university. It turns out that they're doing a lot better than average, but they're not topping the class at university even though they did top their classes at high school. What's going on? The natural first thought is that this must mean that the psychology classes must be having an adverse effect on those students. However, while that might very well be the explanation, it's more likely that what you're seeing is an example of "regression to the mean". To see how it works, let's take a moment to think about what is required to get the best mark in a class, regardless of whether that class be at high school or at university. When you've got a big class there are going to be lots of very smart people enrolled. To get the best mark you have to be very smart, work very hard, and be a bit lucky. The exam has to ask just the right questions for your idiosyncratic skills, and you have to avoid making any dumb mistakes (we all do that sometimes) when answering them. And that's the thing, whilst intelligence and hard work are transferable from one class to the next, luck isn't. The people who got lucky in high school won't be the same as the people who get lucky at university. That's the very definition of "luck". The consequence of this is that when you select people at the very extreme values of one measurement (the top 20 students), you're selecting for hard work, skill and luck. But because the luck doesn't transfer to the second measurement (only the skill and work), these people will all be expected to drop a little bit when you measure them a second time (at university). So their scores fall back a little bit, back towards everyone else. This is regression to the mean.

Regression to the mean is surprisingly common. For instance, if two very tall people have kids their children will tend to be taller than average but not as tall as the parents. The reverse happens with very short parents. Two very short parents will tend to have short children, but nevertheless those kids will tend to be taller than the parents. It can also be extremely subtle. For instance, there have been studies done that suggested that people learn better from negative feedback than from positive feedback. However, the way that people tried to show this was to give people positive reinforcement whenever they did good, and negative reinforcement when they did bad. And what you see is that after the positive reinforcement people tended to do worse, but after the negative reinforcement they tended to do better. But notice that there's a selection bias here! When people do very well, you're selecting for "high" values, and so you should expect, because of regression to the mean, that performance on the next trial should be worse regardless of whether reinforcement is given. Similarly, after a bad trial, people will tend to improve all on their own. The apparent superiority of negative feedback is an artefact caused by regression to the mean (see Kahneman & Tversky (1973), for discussion).

2.7.8 Experimenter bias

Experimenter bias can come in multiple forms. The basic idea is that the experimenter, despite the best of intentions, can accidentally end up influencing the results of the experiment by subtly communicating the "right answer" or the "desired behaviour" to the participants. Typically, this occurs because the experimenter has special knowledge that the participant does not, for example the right answer to the questions being asked or knowledge of the expected pattern of performance for the condition that the participant is in. The classic example of this happening is the case study of "Clever Hans", which dates back to 1907 (Pfungst, 1911). Clever Hans was a horse that apparently was able to read and count and perform other human like feats of intelligence. After Clever Hans became famous, psychologists started examining his behaviour more closely. It turned out that, not surprisingly, Hans didn't know how to do maths. Rather, Hans was responding to the human observers around him, because the humans did know how to count and the horse had learned to change its behaviour when people changed theirs.

The general solution to the problem of experimenter bias is to engage in double blind studies, where neither the experimenter nor the participant knows which condition the participant is in or knows what the desired behaviour is. This provides a very good solution to the problem, but it's important to recognise that it's not quite ideal, and hard to pull off perfectly. For instance, the obvious way that I could try to construct a double blind study is to have one of my Ph.D. students (one who doesn't know anything about the experiment) run the study. That feels like it should be enough. The only person (me) who knows all the details (e.g., correct answers to the questions, assignments of participants to conditions) has no interaction with the participants, and the person who does all the talking to people (the Ph.D. student) doesn't know anything. Except for the reality that the last part is very unlikely to be true. In order for the Ph.D. student to run the study effectively they need to have been briefed by me, the researcher. And, as it happens, the Ph.D. student also knows me and knows a bit about my general beliefs about people and psychology (e.g., I tend to think humans are much smarter than psychologists give them credit for). As a result of all this, it's almost impossible for the experimenter to avoid knowing a little bit about what expectations I have. And even a little bit of knowledge can have an effect. Suppose the experimenter accidentally conveys the fact that the participants are expected to do well in this task. Well, there's a thing called the "Pygmalion effect", where if you expect great things of people they'll tend to rise to the occasion. But if you expect them to fail then they'll do that too. In other words, the expectations become a self-fulfilling prophecy.

2.7.9 Demand effects and reactivity

When talking about experimenter bias, the worry is that the experimenter's knowledge or desires for the experiment are communicated to the participants, and that these can change people's behaviour (Rosenthal, 1966). However, even if you manage to stop this from happening, it's almost impossible to stop people from knowing that they're part of a psychological study. And the mere fact of knowing that someone is watching or studying you can have a pretty big effect on behaviour. This is generally referred to as **reactivity** or **demand effects**. The basic idea is captured by the Hawthorne effect: people alter their performance because of the attention that the study focuses on them. The effect takes its name from a study that took place in the "Hawthorne Works" factory outside of Chicago (see Adair (1984)). This study, from the 1920s, looked at the effects of factory lighting on worker productivity. But, importantly, change in worker behaviour occurred because the workers knew they were being studied, rather than any effect of factory lighting.

To get a bit more specific about some of the ways in which the mere fact of being in a study can change how people behave, it helps to think like a social psychologist and look at some of the roles that people might adopt during an experiment but might not adopt if the corresponding events were occurring in the real world:

- The *good participant* tries to be too helpful to the researcher. He or she seeks to figure out the experimenter's hypotheses and confirm them.
- The *negative participant* does the exact opposite of the good participant. He or she seeks to break or destroy the study or the hypothesis in some way.
- The *faithful participant* is unnaturally obedient. He or she seeks to follow instructions perfectly, regardless of what might have happened in a more realistic setting.
- The *apprehensive participant* gets nervous about being tested or studied, so much so that his or her behaviour becomes highly unnatural, or overly socially desirable.

2.7.10 Placebo effects

The **placebo effect** is a specific type of demand effect that we worry a lot about. It refers to the situation where the mere fact of being treated causes an improvement in outcomes. The classic example comes from clinical trials. If you give people a completely chemically inert drug and tell them that it's a cure for a disease, they will tend to get better faster than people who aren't treated at all. In other words, it is people's belief that they are being treated that causes the improved outcomes, not the drug.

However, the current consensus in medicine is that true placebo effects are quite rare and most of what was previously considered placebo effect is in fact some combination of natural healing (some people just get better on their own), regression to the mean and other quirks of study design. Of interest to psychology is that the strongest evidence for at least some placebo effect is in self-reported outcomes, most notably in treatment of pain (Hróbjartsson & Gøtzsche, 2010).

2.7.11 Situation, measurement and sub-population effects

In some respects, these terms are a catch-all term for "all other threats to external validity". They refer to the fact that the choice of sub-population from which you draw your participants, the location, timing and manner in which you run your study (including who collects the data) and the tools that you use to make your measurements might all be influencing the results. Specifically, the worry is that these things might be influencing the results in such a way that the results won't generalise to a wider array of people, places and measures.

2.7.12 Fraud, deception and self-deception

> "*It is difficult to get a man to understand something, when his salary depends on his not understanding it.*"
> - Upton Sinclair

There's one final thing I feel I should mention. While reading what the textbooks often have to say about assessing the validity of a study I couldn't help but notice that they seem to make the assumption that the researcher is honest. I find this hilarious. While the vast majority of scientists are honest, in my experience at least, some are not.[6] Not only that, as I mentioned earlier, scientists are not immune to belief bias. It's easy for a researcher to end up deceiving themselves into believing the wrong thing, and this can lead them to conduct subtly flawed research and then hide those flaws when they write it up. So you need to consider not only the (probably unlikely) possibility of outright fraud, but also the (probably quite common) possibility that the research is unintentionally "slanted". I opened a few standard textbooks and didn't find much of a discussion of this problem, so here's my own attempt to list a few ways in which these issues can arise:

[6]Some people might argue that if you're not honest then you're not a real scientist. Which does have some truth to it I guess, but that's disingenuous (look up the "No true Scotsman" fallacy). The fact is that there are lots of people who are employed ostensibly as scientists, and whose work has all of the trappings of science, but who are outright fraudulent. Pretending that they don't exist by saying that they're not scientists is just muddled thinking.

- **Data fabrication**. Sometimes, people just make up the data. This is occasionally done with "good" intentions. For instance, the researcher believes that the fabricated data do reflect the truth, and may actually reflect "slightly cleaned up" versions of actual data. On other occasions, the fraud is deliberate and malicious. Some high-profile examples where data fabrication has been alleged or shown include Cyril Burt (a psychologist who is thought to have fabricated some of his data), Andrew Wakefield (who has been accused of fabricating his data connecting the MMR vaccine to autism) and Hwang Woo-suk (who falsified a lot of his data on stem cell research).
- **Hoaxes**. Hoaxes share a lot of similarities with data fabrication, but they differ in the intended purpose. A hoax is often a joke, and many of them are intended to be (eventually) discovered. Often, the point of a hoax is to discredit someone or some field. There's quite a few well known scientific hoaxes that have occurred over the years (e.g., Piltdown man) and some were deliberate attempts to discredit particular fields of research (e.g., the Sokal affair).
- **Data misrepresentation**. While fraud gets most of the headlines, it's much more common in my experience to see data being misrepresented. When I say this I'm not referring to newspapers getting it wrong (which they do, almost always). I'm referring to the fact that often the data don't actually say what the researchers think they say. My guess is that, almost always, this isn't the result of deliberate dishonesty but instead is due to a lack of sophistication in the data analyses. For instance, think back to the example of Simpson's paradox that I discussed in the beginning of this book. It's very common to see people present "aggregated" data of some kind and sometimes, when you dig deeper and find the raw data yourself you find that the aggregated data tell a different story to the disaggregated data. Alternatively, you might find that some aspect of the data is being hidden, because it tells an inconvenient story (e.g., the researcher might choose not to refer to a particular variable). There's a lot of variants on this, many of which are very hard to detect.
- **Study "misdesign"**. Okay, this one is subtle. Basically, the issue here is that a researcher designs a study that has built-in flaws and those flaws are never reported in the paper. The data that are reported are completely real and are correctly analysed, but they are produced by a study that is actually quite wrongly put together. The researcher really wants to find a particular effect and so the study is set up in such a way as to make it "easy" to (artefactually) observe that effect. One sneaky way to do this, in case you're feeling like dabbling in a bit of fraud yourself, is to design an experiment in which it's obvious to the participants what they're "supposed" to be doing, and then let reactivity work its magic for you. If you want you can add all the trappings of double blind experimentation but it won't make a difference since the study materials themselves are subtly telling people what you want them to do. When you write up the results the fraud won't be obvious to the reader. What's obvious to the participant when they're in the experimental context isn't always obvious to the person reading the paper. Of course, the way I've described this makes it sound like it's always fraud. Probably there are cases where this is done deliberately, but in my experience the bigger concern is with unintentional misdesign. The researcher believes and so the study just happens to end up with a built-in flaw, and that flaw then magically erases itself when the study is written up for publication.
- **Data mining and post hoc hypothesising**. Another way in which the authors

of a study can more or less misrepresent the data is by engaging in what's referred to as "data mining" (see Gelman and Loken 2014, for a broader discussion of this as part of the "garden of forking paths" in statistical analysis). As we'll discuss later, if you keep trying to analyse your data in lots of different ways, you'll eventually find something that "looks" like a real effect but isn't. This is referred to as "data mining". It used to be quite rare because data analysis used to take weeks, but now that everyone has very powerful statistical software on their computers it's becoming very common. Data mining per se isn't "wrong", but the more that you do it the bigger the risk you're taking. The thing that is wrong, and I suspect is very common, is unacknowledged data mining. That is, the researcher runs every possible analysis known to humanity, finds the one that works, and then pretends that this was the only analysis that they ever conducted. Worse yet, they often "invent" a hypothesis after looking at the data to cover up the data mining. To be clear. It's not wrong to change your beliefs after looking at the data, and to reanalyse your data using your new "post hoc" hypotheses. What is wrong (and I suspect common) is failing to acknowledge what you've done. If you acknowledge that you did it then other researchers are able to take your behaviour into account. If you don't, then they can't. And that makes your behaviour deceptive. Bad!

- **Publication bias and self-censoring**. Finally, a pervasive bias is "non-reporting" of negative results. This is almost impossible to prevent. Journals don't publish every article that is submitted to them. They prefer to publish articles that find "something". So, if 20 people run an experiment looking at whether reading Finnegans Wake causes insanity in humans, and 19 of them find that it doesn't, which one do you think is going to get published? Obviously, it's the one study that did find that Finnegans Wake causes insanity.[7] This is an example of a publication bias. Since no-one ever published the 19 studies that didn't find an effect, a naive reader would never know that they existed. Worse yet, most researchers "internalise" this bias and end up self-censoring their research. Knowing that negative results aren't going to be accepted for publication, they never even try to report them. As a friend of mine says "for every experiment that you get published, you also have 10 failures". And she's right. The catch is, while some (maybe most) of those studies are failures for boring reasons (e.g., you stuffed something up) others might be genuine "null" results that you ought to acknowledge when you write up the "good" experiment. And telling which is which is often hard to do. A good place to start is a paper by Ioannidis (2005) with the depressing title "Why most published research findings are false". I'd also suggest taking a look at work by Kühberger et al. (2014) presenting statistical evidence that this actually happens in psychology.

There's probably a lot more issues like this to think about, but that'll do to start with. What I really want to point out is the blindingly obvious truth that real-world science is conducted by actual humans, and only the most gullible of people automatically assume that everyone else is honest and impartial. Actual scientists aren't usually that naive, but for some reason the world likes to pretend that we are, and the textbooks we usually write seem to reinforce that stereotype.

[7]Clearly, the real effect is that only insane people would even try to read *Finnegans Wake*.

2.8 Summary

This chapter isn't really meant to provide a comprehensive discussion of psychological research methods. It would require another volume just as long as this one to do justice to the topic. However, in real life statistics and study design are so tightly intertwined that it's very handy to discuss some of the key topics. In this chapter, I've briefly discussed the following topics:

- Introduction to psychological measurement. What does it mean to operationalise a theoretical construct? What does it mean to have variables and take measurements?
- Scales of measurement and types of variables. Remember that there are two different distinctions here. There's the difference between discrete and continuous data, and there's the difference between the four different scale types (nominal, ordinal, interval and ratio).
- Assessing the reliability of a measurement. If I measure the "same" thing twice, should I expect to see the same result? Only if my measure is reliable. But what does it mean to talk about doing the "same" thing? Well, that's why we have different types of reliability. Make sure you remember what they are.
- The "role" of variables: predictors and outcomes. What roles do variables play in an analysis? Can you remember the difference between predictors and outcomes? Dependent and independent variables? Etc.
- Experimental and non-experimental research designs. What makes an experiment an experiment? Is it a nice white lab coat, or does it have something to do with researcher control over variables?
- Assessing the validity of a study. Does your study measure what you want it to? How might things go wrong? And is it my imagination, or was that a very long list of possible ways in which things can go wrong?

All this should make clear to you that study design is a critical part of research methodology. I built this chapter from the classic little book by Campbell & Stanley (1963), but there are of course a large number of textbooks out there on research design. Spend a few minutes with your favourite search engine and you'll find dozens.

Part II

An introduction to jamovi

Chapter 3

Getting started with jamovi

> "*Robots are nice to work with.*"
> – Roger Zelazny[1]

In this chapter I'll discuss how to get started in jamovi. I'll briefly talk about how to download and install jamovi, but most of the chapter will be focused on getting you started with finding your way around the jamovi graphical user interface (GUI). Our goal in this chapter is not to learn any statistical concepts: we're just trying to learn the basics of how jamovi works and get comfortable interacting with the system. To do this we'll spend a bit of time looking at datasets and variables. In doing so, you'll get something of a feel for what it's like to work in jamovi.

However, before going into any of the specifics, it's worth talking a little about why you might want to use jamovi 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 jamovi for the class. Of course, I don't really know why *other* people choose jamovi 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 conventional spreadsheet (e.g., Microsoft Excel) is generally a bad idea in the long run. Although many people 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.

[1]Source: "Dismal Light" (1968).

- 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" (crippled 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 jamovi 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 jamovi is basically a sophisticated front end for the free R statistical programming language. 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 look at recent advanced data analysis textbooks then a lot of them use R.

Those are the main reasons I use jamovi. It's not without its flaws, though. It's relatively new[2] so there is not a huge set of textbooks and other resources to support it, and it has a few annoying quirks 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.

3.1 Installing jamovi

Okay, enough with the sales pitch. Let's get started. Just as with any piece of software, jamovi 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, jamovi is freely distributed online and you can download it from the jamovi homepage, which is: https://www.jamovi.org/

At the top of the page under the heading "Download", 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. At the time of writing, the current version of jamovi is 2.3, but they usually issue updates every few months, so you'll probably have a newer version.[3]

[2] At the time of first writing this in August 2018. Later versions of this book will use later versions of jamovi.

[3] Although jamovi 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 it didn't make much difference at all to what is in this book.

3.1.1 Starting up jamovi

One way or another, regardless of what operating system you're using, it's time to open jamovi and get started. When first starting jamovi you will be presented with a user interface which looks something like Figure 3.1.

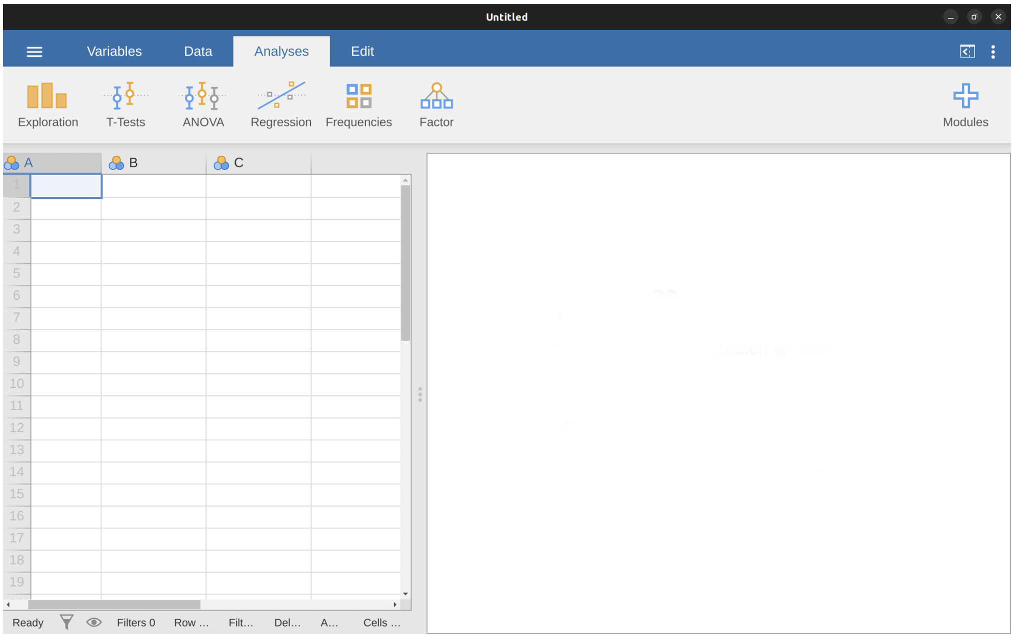


Figure 3.1: jamovi starts up!

To the left is the spreadsheet view, and to the right is where the results of statistical tests appear. Down the middle is a bar separating these two regions and this can be dragged to the left or the right to change their sizes.

It is possible to simply begin typing values into the jamovi spreadsheet as you would in any other spreadsheet software. Alternatively, existing data sets in the csv (.csv) file format can be opened in jamovi. Additionally, you can easily import SPSS, SAS, STATA and JASP files directly into jamovi. To open a file select the 'File'[4] tab (three horizontal lines signify this tab) at the top left hand corner, select 'Open' and then choose from the files listed on 'Browse' depending on whether you want to open an example or a file stored on your computer.

3.2 Analyses

Analyses can be selected from the analysis ribbon or menu along the top. Selecting an analysis will present an 'Options panel' for that particular analysis, allowing you to assign different variables to different parts of the analysis, and select different options.

[4] From now on, we'll use single quote marks to signify a label, command, option, or outputs in the jamovi interface.

At the same time, the results for the analysis will appear in the right 'Results panel' and will update in real time as you make changes to the options.

When you have the analysis set up correctly you can dismiss the analysis options by clicking the arrow to the top right of the optional panel. If you wish to return to these options, you can click on the results that were produced. In this way, you can return to any analysis that you (or say, a colleague) created earlier.

If you decide you no longer need a particular analysis, you can remove it with the results context menu. Right-clicking on the analysis results will bring up a menu and by selecting 'Analysis' and then 'Remove' the analysis can be removed. But more on this later. First, let's take a more detailed look at the spreadsheet view.

3.3 The spreadsheet

In jamovi data is represented in a spreadsheet with each column representing a 'variable' and each row representing a 'case' or 'participant'.

3.3.1 Variables

The most commonly used variables in jamovi are 'Data variables', these variables simply contain data either loaded from a data file, or 'typed in' by the user. Data variables can be one of several measurement levels (Figure 3.2).

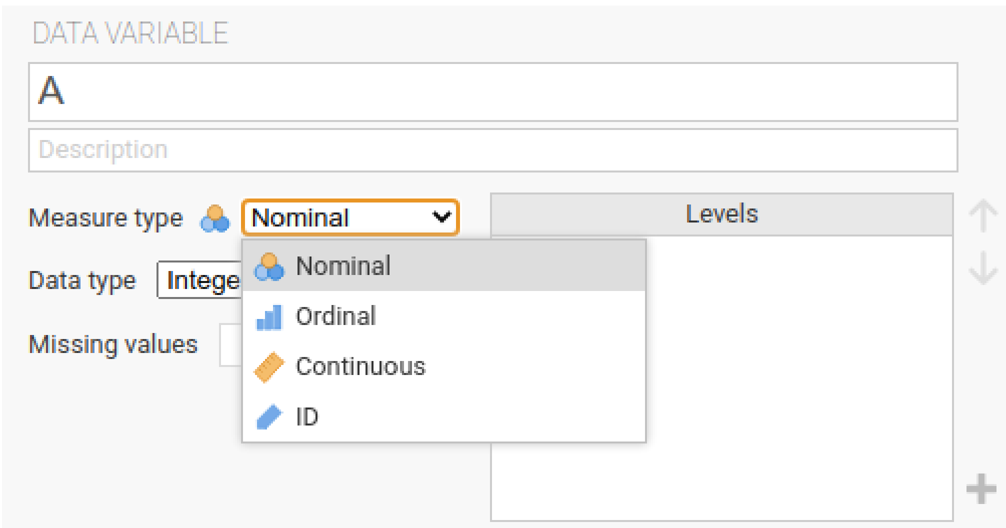


Figure 3.2: measurement levels

These levels are designated by the symbol in the header of the variable's column. The ID variable type is unique to jamovi. It's intended for variables that contain identifiers that you would almost never want to analyse. For example, a person's name, or a participant

ID. Specifying an ID variable type can improve performance when interacting with very large data sets.

Nominal variables are for categorical variables which are text labels, for example a column called 'gender' with the values 'male' and 'female' would be nominal. So would a person's name. Nominal variable values can also have a numeric value. These variables are used most often when importing data which codes values with numbers rather than text. For example, a column in a dataset may contain the values 1 for males, and 2 for females. It is possible to add nice 'human-readable' labels to these values with the variable editor (more on this later).

Ordinal variables are like nominal variables, except the values have a specific order. An example is a Likert scale with 3 being 'strongly agree' and -3 being 'strongly disagree'.

Continuous variables are variables which exist on a continuous scale. Examples might be height or weight. This is also referred to as 'Interval' or 'Ratio scale'.

In addition, you can also specify different data types: variables have a data type of either 'Text', 'Integer' or 'Decimal'.

When starting with a blank spreadsheet and typing values in the variable type will change automatically depending on the data you enter. This is a good way to get a feel for which variable types go with which sorts of data. Similarly, when opening a data file jamovi will try and guess the variable type from the data in each column. In both cases this automatic approach may not be correct, and it may be necessary to manually specify the variable type with the variable editor.

The variable editor can be opened by selecting 'Setup' from the data tab or by double-clicking on the variable column header. The variable editor allows you to change the name of the variable and, for data variables, the variable type, the order of the levels, and the label displayed for each level. Changes can be applied by clicking the 'tick' to the top right. The variable editor can be dismissed by clicking the 'Hide' arrow.

New variables can be inserted or appended to the data set using the 'Add' button from the data ribbon. The 'Add' button also allows the addition of computed variables.

3.3.2 Computed variables

Computed Variables are those which take their value by performing a computation on other variables. Computed Variables can be used for a range of purposes, including log transforms, z-scores, sum-scores, negative scoring and means.

Computed variables can be added to the data set with the 'Add' button available on the data tab. This will produce a formula box where you can specify the formula. The usual arithmetic operators are available. Some examples of formulas are:

A + B LOG10(len) MEAN(A, B) (len - VMEAN(len)) / VSTDEV(len)

In order, these are the sum of A and B, a log (base 10) transform of len, the mean of A and B, and the z-score of the variable len.[5] Figure 3.3 shows the jamovi screen for the new variable computed as the z-score of len (from the 'Tooth Growth' example data set).

[5]In later versions of jamovi there is a pre-defined function 'Z' to compute z-scores, which is much easier!

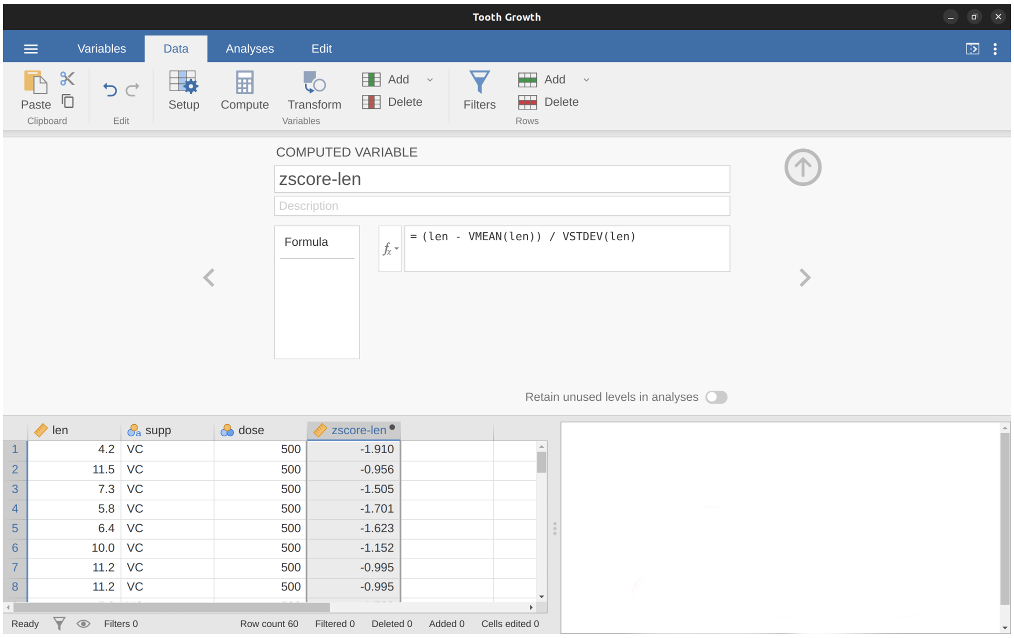


Figure 3.3: A newly computed variable, the *z*-score of 'dose'

3.3.2.1 V-functions

Several functions are already available in jamovi and available from the drop down box labelled fx. A number of functions appear in pairs, one prefixed with a V and the other not. V functions perform their calculation on a variable as a whole, where as non-V functions perform their calculation row by row. For example, MEAN(A, B) will produce the mean of A and B for each row. Where as VMEAN(A) gives the mean of all the values in A.

3.3.3 Copy and paste

jamovi produces nice American Psychological Association (APA) formatted tables and attractive plots. It is often useful to be able to copy and paste these, perhaps into a Word document, or into an email to a colleague. To copy results right click on the object of interest and from the menu select exactly what you want to copy. The menu allows you to choose to copy only the image or the entire analysis. Selecting 'Copy' copies the content to the clipboard and this can be pasted into other programs in the usual way. You can practice this later on when we do some analyses.

3.3.4 Syntax mode

jamovi also provides an 'R Syntax mode'. In this mode jamovi produces equivalent R code for each analysis. To change to syntax mode, select the 'Application' menu to the top right of jamovi (a button with three vertical dots) and click the 'Syntax mode' checkbox there. You can turn off syntax mode by clicking this a second time.

In syntax mode analyses continue to operate as before but now they produce R syntax, and "ascii output" like an R session. Like all results objects in jamovi, you can right click on these items (including the R syntax) and copy and paste them, for example into an R session. At present, the provided R syntax does not include the data import step and so this must be performed manually in R. There are many resources explaining how to import data into R and if you are interested we recommend you take a look at these; just search on the interweb.

3.4 Loading data in jamovi

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

- *jamovi files* are those with a .omv file extension. This is the standard kind of file that jamovi uses to store data, and variables and analyses.
- *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 many different software programs. It's quite typical for people to store data in csv files, precisely because they're so simple.

There are also several other kinds of data file that you might want to import into jamovi. For instance, you might want to open Microsoft Excel spreadsheets (.xls files), or data files that have been saved in the native file formats for other statistics software, such as SPSS or SAS. Whichever file formats you are using, it's a good idea to create a folder or folders especially for your jamovi data sets and analyses and to make sure you keep these backed up regularly.

3.4.1 Importing data from csv files

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 3.4, which shows a file called *booksales.csv* that I've created. As you can see, 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.

It's easy to open csv files in jamovi. From the top left menu (the button with three parallel lines) choose 'Open' and browse to where you have stored the csv file on your computer. 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 3.5. 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.

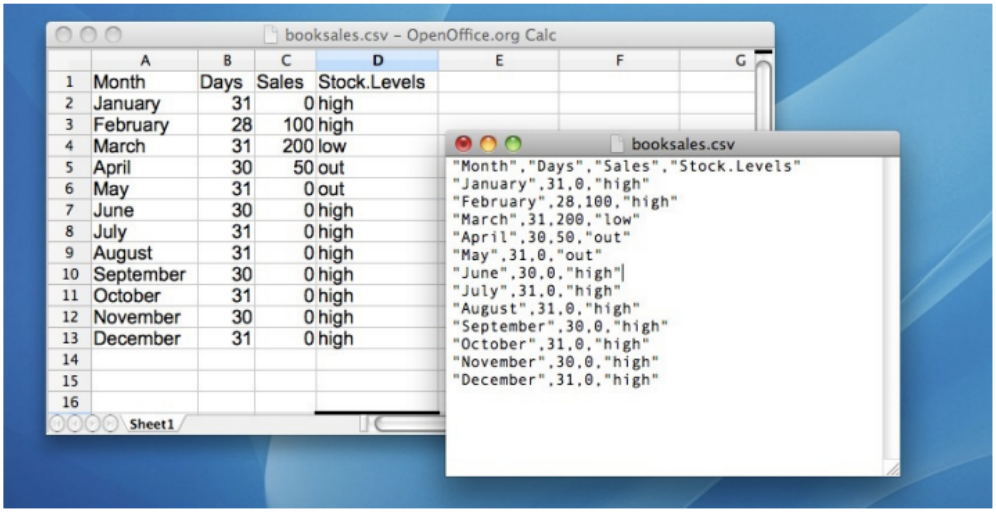


Figure 3.4: The *booksales.csv* data file. On the left I have opened the file using a spreadsheet program (OpenOffice), which shows that the file is basically a table. On the right the same file is open in a standard text editor (the TextEdit program on a Mac), which shows how the file is formatted. The entries in the table are wrapped in quote marks and separated by commas

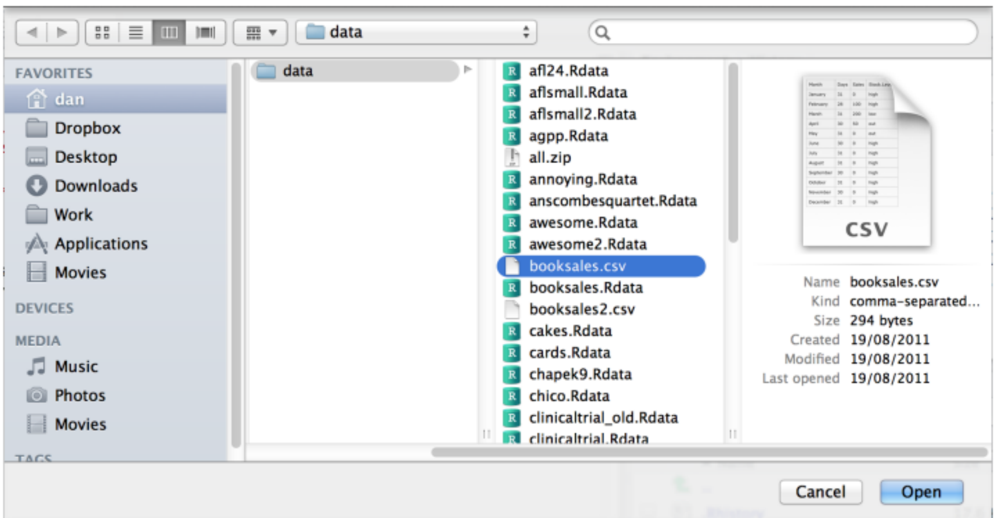


Figure 3.5: A dialog box on a Mac asking you to select the csv file jamovi should try to import. Mac users will recognise this immediately, as it is the usual way in which a Mac asks you to find a file. Windows users will not see this, instead they will see the usual explorer window that Windows always gives you when it wants you to select a file

There are a few things that you can check to make sure that the data gets imported correctly:

- Heading. Does the first row of the file contain the names for each variable – a 'header' row? The *booksales.csv* file has a header, so that's a yes.
- 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 though, many European countries use a comma.
- 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.

3.5 Importing unusual data files

Throughout this book I've assumed that your data are stored as a jamovi .omv file or as a "properly" formatted csv file. However, in real life that's not a terribly plausible assumption to make so I'd better talk about some of the other possibilities that you might run into.

3.5.1 Loading data from text files

The first thing I should point out is that if your data are saved as a text file but aren't quite in the proper csv format then there's still a pretty good chance that jamovi will be able to open it. You just need to try it and see if it works. Sometimes though you will need to change some of the formatting. The ones that I've often found myself needing to change are:

- header. A lot of the time when you're storing data as a csv file the first row actually contains the column names and not data. If that's not true then it's a good idea to open up the csv file in a spreadsheet programme such as Open Office and add the header row manually.
- sep. As the name "comma separated value" indicates, the values in a row of a csv file are usually separated by commas. This isn't universal, however. In Europe the decimal point is typically written as , instead of . and as a consequence it would be somewhat awkward to use , as the separator. Therefore it is not unusual to use ; instead of , as the separator. At other times, I've seen a TAB character used.
- quote. It's conventional in csv files to include a quoting character for textual data. As you can see by looking at the *booksales.csv* file, this is usually a double quote character, ". But sometimes there is no quoting character at all, or you might see a single quote mark ' used instead.
- skip. It's actually very common to receive csv files in which the first few rows have nothing to do with the actual data. Instead, they provide a human readable summary of where the data came from, or maybe they include some technical info that doesn't relate to the data.
- missing values. Often you'll get given data with missing values. For one reason or another, some entries in the table are missing. The data file needs to include a "special" value to indicate that the entry is missing. By default jamovi assumes

that this value is 99,[6] for both numeric and text data, so you should make sure that, where necessary, all missing values in the csv file are replaced with 99 (or -9999; whichever you choose) before opening / importing the file into jamovi. Once you have opened / imported the file into jamovi all the missing values are converted to blank or greyed out cells in the jamovi spreadsheet view. You can also change the missing value for each variable as an option in the Data - Setup view.

3.5.2 Loading data from SPSS (and other statistics packages)

The commands listed above are the main ones we'll need for data files in this book. But in real life we have many more possibilities. For example, you might want to read data files in from other statistics programs. Since SPSS is probably the most widely used statistics package in psychology, it's worth mentioning that jamovi can also import SPSS data files (file extension .sav). Just follow the instructions above for how to open a csv file, but this time navigate to the .sav file you want to import. For SPSS files, jamovi will regard all values as missing if they are regarded as "system missing" files in SPSS. The 'Default missings' value does not seem to work as expected when importing SPSS files, so be aware of this - you might need another step: import the SPSS file into jamovi, then export as a csv file before re-opening in jamovi.[7]

And that's pretty much it, at least as far as SPSS goes. As far as other statistical software goes, jamovi can also directly open / import SAS and STATA files.

3.5.3 Loading Excel files

A different problem is posed by Excel files. Despite years of yelling at people for sending data to me encoded in a proprietary data format, I get sent a lot of Excel files. The way to handle Excel files is to open them up first in Excel or another spreadsheet programme that can handle Excel files, and then export the data as a csv file before opening / importing the csv file into jamovi.

3.6 Changing data from one level to another

Sometimes you want to change the variable level. This can happen for all sorts of reasons. Sometimes when you import data from files, it can come to you in the wrong format. Numbers sometimes get imported as nominal, text values. Dates may get imported as text. ParticipantID values can sometimes be read as continuous: nominal values can sometimes be read as ordinal or even continuous. There's a good chance that sometimes you'll want to convert a variable from one measurement level into another

[6] You can change the default value for missing values in jamovi from the top right menu (three vertical dots), but this only works at the time of importing data files into jamovi. The default missing value in the dataset should not be a valid number associated with any of the variables, e.g., you could use -9999 as this is unlikely to be a valid value.

[7] I know this is a bit of a fudge, but it does work and hopefully this will be fixed in a later version of jamovi.

one. Or, to use the correct term, you want to **coerce** the variable from one class into another.

Earlier we saw how to specify different variable levels, and if you want to change a variable's measurement level then you can do this in the jamovi data view for that variable. Just click the check box for the measurement level you want – continuous, ordinal, or nominal.

3.7 Installing add-on modules into jamovi

A really great feature of jamovi is the ability to install add-on modules from the jamovi library. These add-on modules have been developed by the jamovi community, i.e., jamovi users and developers who have created special software add-ons that do other, usually more advanced, analyses that go beyond the capabilities of the base jamovi program.

To install add-on modules, just click on the large + in the top right of the jamovi window, select "jamovi-library" and then browse through the various add-on modules that are available. Choose the one(s) you want, and then install them, as in Figure 3.6. It's that easy. The newly installed modules can then be accessed from the "Analyses" button bar. Try it...useful add-on modules to install include "scatr" (added under "Descriptives") and R_j.

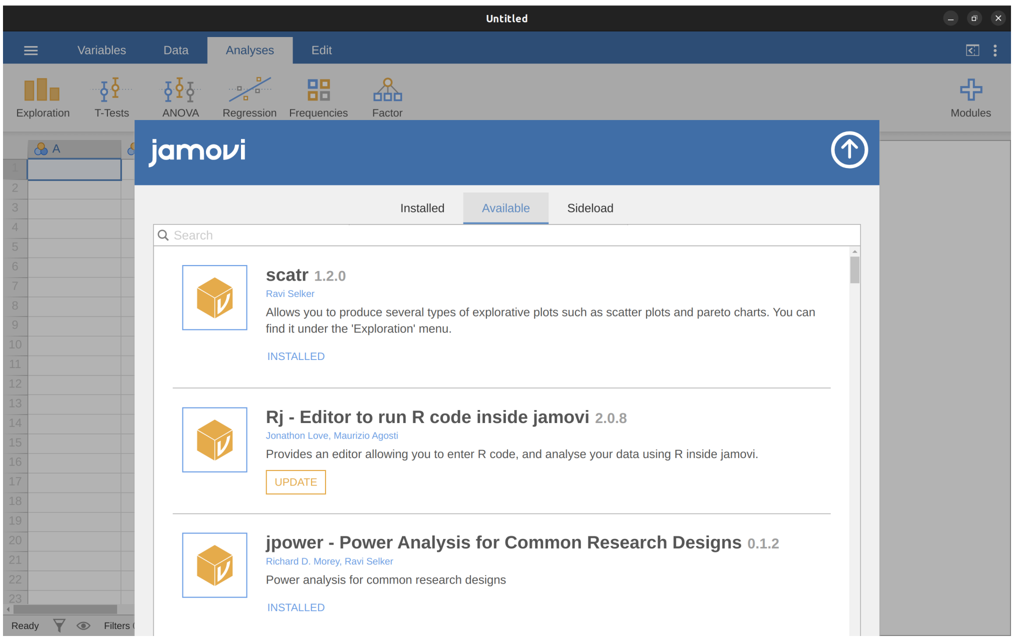


Figure 3.6: Installing add-on modules in jamovi

3.8 Quitting jamovi

There's one last thing I should cover in this chapter: how to quit jamovi. It's not hard, just close the program the same way you would any other program. However, what you might want to do before you quit is save your work! There are two parts to this: saving any changes to the data set, and saving the analyses that you ran.

It is good practice to save any changes to the data set as a *new* data set. That way you can always go back to the original data. To save any changes in jamovi, select 'Export'…'Data' from the main jamovi menu (button with three horizontal bars in the top left) and create a new file name for the changed data set.

Alternatively, you can save *both* the changed data and any analyses you have undertaken by saving as a jamovi file. To do this, from the main jamovi menu select 'Save as' and type in a file name for this 'jamovi file (.omv)'. Remember to save the file in a location where you can find it again later. I usually create a new folder for specific data sets and analyses.

3.9 Summary

Every book that tries to teach a new statistical software program to novices has to cover roughly the same topics, and in roughly the same order. Ours 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:

- Installing jamovi. We downloaded and installed jamovi, and started it up.
- Analyses. We very briefly oriented to the part of jamovi where analyses are done and results appear, but then deferred this until later in the book.
- The spreadsheet. We spent more time looking at the spreadsheet part of jamovi, and considered different variable types, and how to compute new variables.
- Loading data in jamovi. We also saw how to load data files in jamovi.
- Importing unusual data files. Then we figured out how to open other data files, from different file types.
- Changing data from one level to another. And saw that sometimes we need to coerce data from one type to another.
- Installing add-on modules into jamovi. Installing add-on modules from the jamovi community really extends jamovi capabilities.
- Quitting jamovi. Finally, we looked at good practice in terms of saving your data set and analyses when you have finished and are about to quit jamovi.

We still haven't arrived at anything that resembles data analysis. Maybe the next chapter will get us a bit closer!

Part III

Working with data

Chapter 4

Descriptive statistics

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 open the *aflsmall_margins* file and see what variables are stored in the file, see Figure 4.1.

In fact, there is just one variable here, *afl.margins*. We'll focus a bit on this variable in this chapter, so I'd better tell you what it is. Unlike most of the data sets in this book, this is 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.

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 the data are actually saying we need to calculate some descriptive statistics (this chapter) and draw some nice pictures (Chapter 5). 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, see Figure 4.2. We'll talk a lot more about how to draw histograms in Section 5.1 in the next chapter. For now, it's enough to look at the histogram and note that it provides a fairly interpretable representation of the *afl.margins* data.

4.1 Measures of central tendency

Drawing pictures of the data, as I did in Figure 4.2, 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

[1] Note 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.

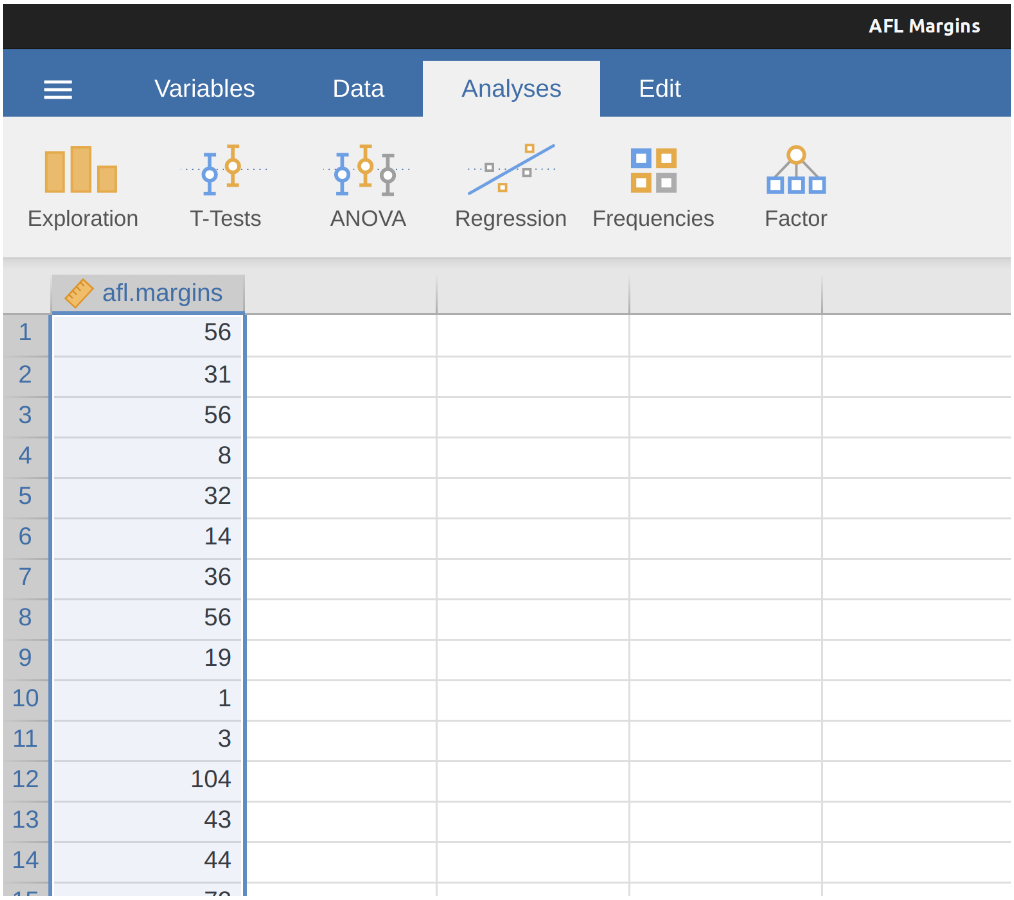


Figure 4.1: A screenshot of jamovi showing the variables stored in the aflsmall_margins.csv file

Plots

afl.margins

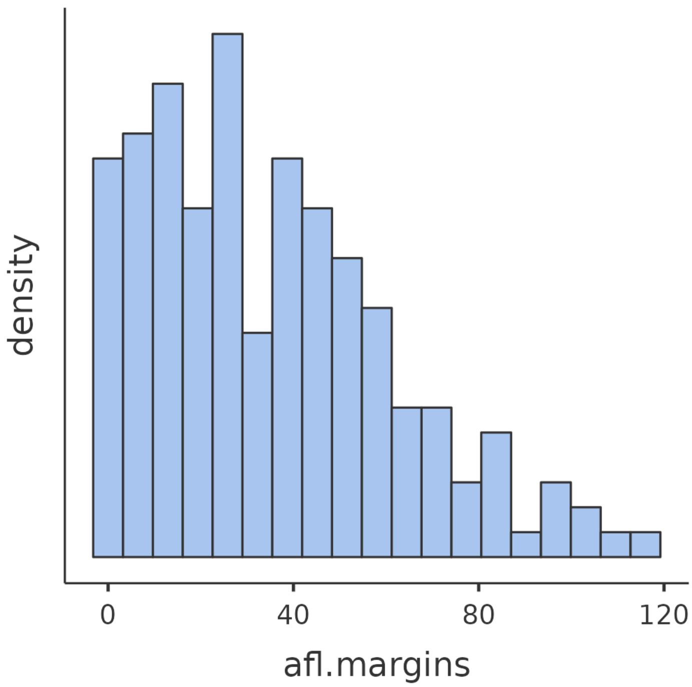


Figure 4.2: A histogram of the AFL 2010 winning margin data (the *afl.margins* variable). As you might expect, the larger the winning margin the less frequently you tend to see it

thing that you'll want to calculate is a measure of **central tendency**. That is, you'd like to know something about where the "average" or "middle" of your data lies. The three most commonly used measures are the mean, median and mode. I'll explain each of these in turn, and then discuss when each of them is useful.

4.1.1 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 winning 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 jamovi.

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, Table 4.1 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.

Table 4.1: Observations in the *afl.margins* variable

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

[Additional technical detail[2]]

[2]Okay, now let's try to write a formula for the mean. By tradition, we use $\bar{X}$ as the notation for the

4.1.2 Calculating the mean in jamovi

Okay, that's the maths. So how do we get the magic computing box to do the work for us? When the number of observations starts to become large it's much easier to do these sorts of calculations using a computer. To calculate the mean using all the data we can use jamovi. The first step is to click on the 'Exploration' button and then click 'Descriptives'. Then you can highlight the *afl.margins* variable and click the 'Right arrow' to move it across into the 'Variables box'. As soon as you do that a Table appears on the right-hand side of the screen containing default 'Descriptives' information; see Figure 4.3.

As you can see in Figure 4.3, the mean value for the *afl.margins* variable is 35.30. Other information presented includes the total number of observations (N=176), the number of missing values (none), and the median, minimum and maximum values for the variable.

4.1.3 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 are interested in the first 6 games rather than

---

mean. So the calculation for the mean could be expressed using the following formula:

$$\bar{X} = \frac{X_1 + X_2 ... + X_{N-1} + X_N}{N}$$

This formula is entirely correct but it's terribly long, so we make use of the **summation symbol** $\sum$ to shorten it. 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, $\sum$ for the idea of summation, X_i for the ith 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". *The choice to use* $\sum$ 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 $\prod$ symbol for this (the Greek upper case pi, which is the analogue of the letter P.

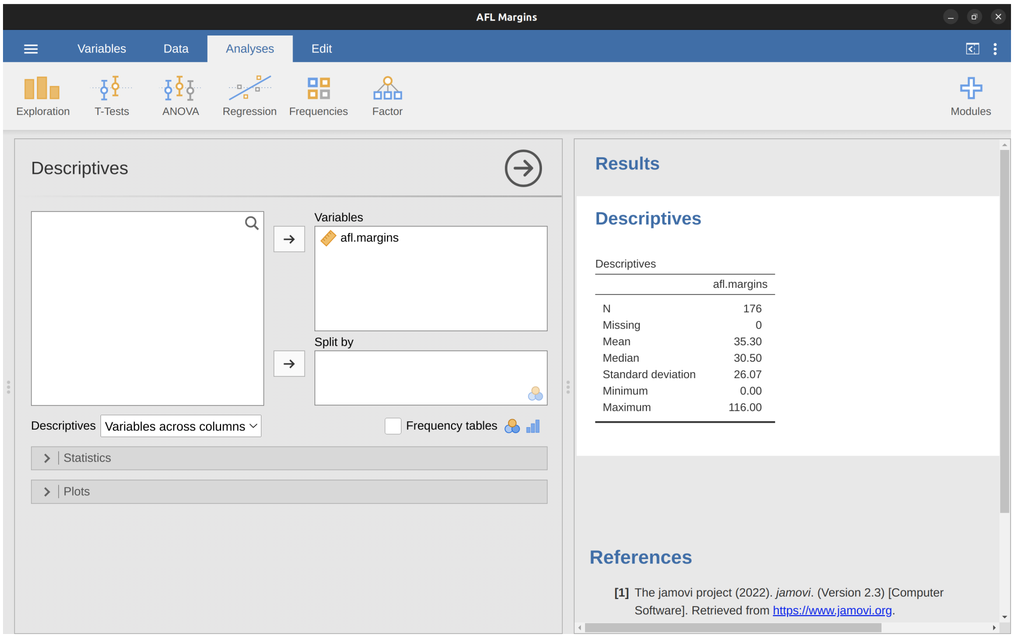


Figure 4.3: Default descriptives for the AFL 2010 winning margin data (the *afl.margins* variable)

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. 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 a computer to do the heavy lifting for us, and jamovi has provided us with a median value of 30.50 for the *afl.margins* variable (Figure 4.3).

4.1.4 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 4.4. 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 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.
- 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, outlying values.

Let's expand on that last part a little. One consequence is that there are systematic differences between the mean and the median when the histogram is asymmetric (Skew and kurtosis). This is illustrated in Figure 4.4. 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.

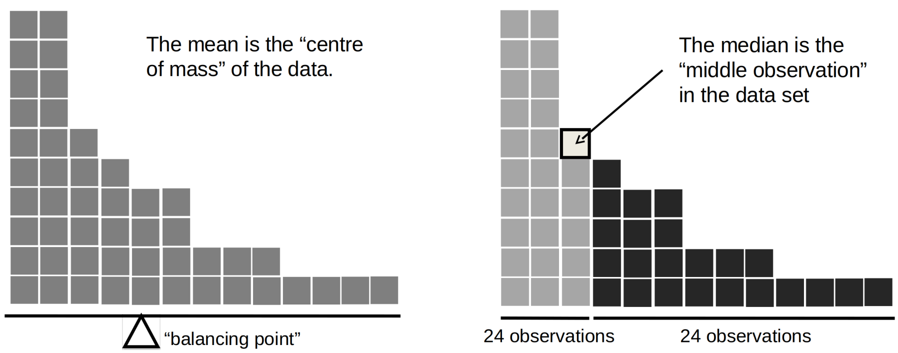


Figure 4.4: An illustration of the difference between how the mean and the median should be interpreted. The mean is basically the 'centre of gravity' of the data set. If you imagine that the histogram of the data is a solid object, then the point on which you could balance it (as if on a see-saw) is the mean. In contrast, the median is the middle observation, with half of the observations smaller and half of the observations larger

4.1.5 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 an excellent article by Michael Janda from the ABC news website[3] from 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 is correct, and the Commonwealth Bank is wrong. As the article points out:

> [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

[3] www.abc.net.au/news/2010-09-24/housing-bubble-debate-boils-over/2273406

> 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 correct. 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.

4.1.6 Mode

The mode of a sample is very simple. It is the value that occurs most frequently. We can illustrate the mode using a different AFL variable: who has played in the most finals? Open the aflsmall finalists file and take a look at the *afl.finalists* variable, see Figure 4.5. This variable contains the names of all 400 teams that played in all 200 finals matches played during the period 1987 to 2010.

What we could do is read through all 400 entries 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 jamovi to do this for us. Under 'Exploration' - 'Descriptives' click the small check box labelled 'Frequency tables' and you should get something like Figure 4.6.

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 *afl.finalists* data is "Geelong". We can see that Geelong (39 finals) played in more finals than any other team during the 1987-2010 period. It's also worth noting that in the 'Descriptives' Table no results are calculated for mean, median, minimum or maximum. This is because the *afl.finalists* variable is a nominal text variable so it makes no sense to calculate these values.

One last point to make regarding the mode. Whilst the mode is most often calculated when you have nominal 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 our *afl.margins* variable. This variable is clearly ratio scale (if it's not clear to you, it may help to re-read Section 2.2), 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 and they pick a football game at random. Without knowing who is playing you

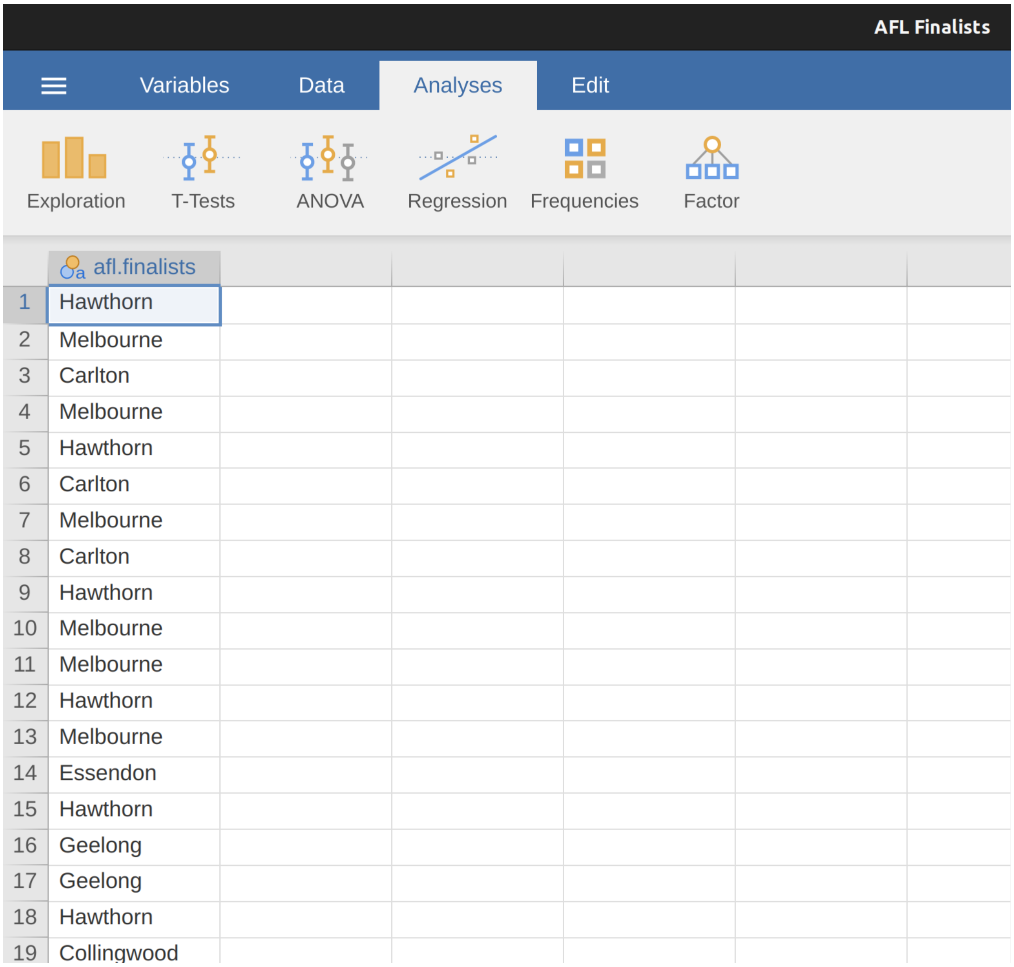


Figure 4.5: A screenshot of jamovi showing the variables stored in the aflsmall finalists.csv file

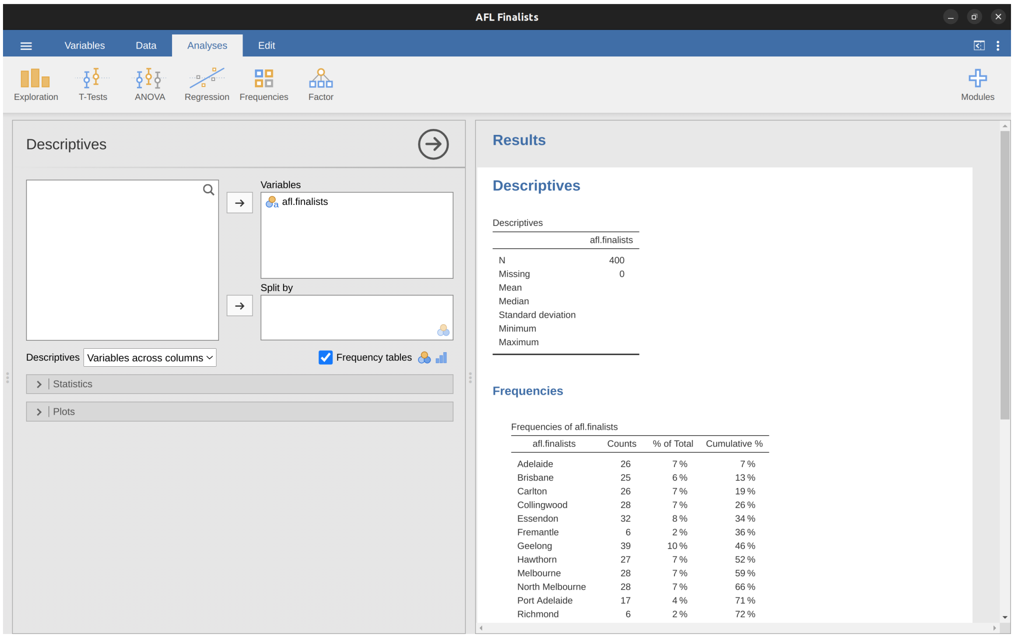


Figure 4.6: A screenshot of jamovi showing the frequency table for the *afl.finalists* variable

have to guess the exact winning 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 margin. For this bet, the mean and the median are completely useless to you. It is the mode that you should bet on. To calculate the mode for the *afl.margins* variable in jamovi, go back to that data set and on the 'Exploration' - 'Descriptives' screen you will see you can expand the section marked 'Statistics'. Click on the checkbox marked 'Mode' and you will see the modal value presented in the 'Descriptives' Table, as in Figure 4.7. So the 2010 data suggest you should bet on a 3 point margin.

4.2 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, and 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.

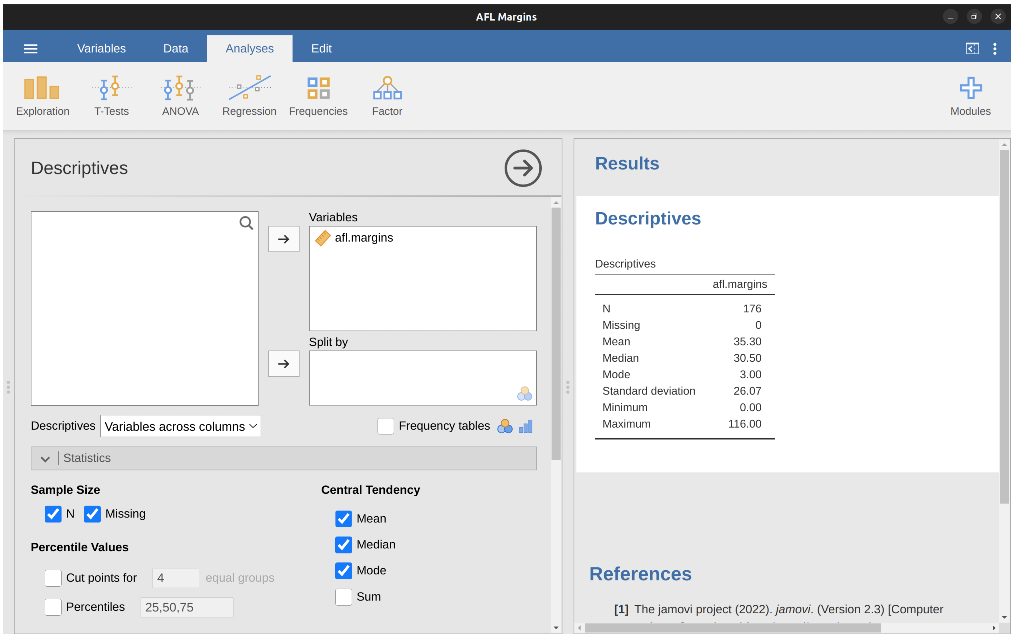


Figure 4.7: A screenshot of jamovi showing the modal value for the *afl.margins* variable

4.2.1 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. 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 to not be unduly influenced by these cases. For example, in a variable containing very extreme outliers

-100, 2, 3, 4, 5, 6, 7, 8, 9, 10

it is clear that the range is not robust. This variable has a range of 110 but if the outlier were removed we would have a range of only 8.

4.2.2 Interquartile range

The **interquartile range** (IQR) is like the range, but instead of the difference between the biggest and smallest value the difference between the 25th percentile and the 75th percentile is taken. If you don't already know what a **percentile** is, 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 percentile! In jamovi you can easily specify the 25th, 50th and 75th percentiles by clicking the checkbox 'Quartiles' in the 'Exploration' - 'Descriptives' - 'Statistics' screen.

And not surprisingly, in Figure 4.8 the 50th percentile is the same as the median value.

And, by noting that $50.50 - 12.75 = 37.75$, we can see that the interquartile range for the 2010 AFL winning margins data is 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 and 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.

4.2.3 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 percentiles 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, this is 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 are 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, and 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 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, Table 4.2 shows these calculations for all five observations.

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.

Descriptives

Descriptives

	afl.margins
N	176
Missing	0
Mean	35.30
Median	30.50
Mode	3.00
Standard deviation	26.07
Minimum	0.00
Maximum	116.00
25th percentile	12.75
50th percentile	30.50
75th percentile	50.50

Figure 4.8: A screenshot of jamovi showing the Quartiles for the *afl.margins* variable

Table 4.2: Measures of variability

English	notation	value	deviation from mean	absolute deviation
notation:	i	X_i	$X_i - \bar{X}$	$\mid X_i - \bar{X} \mid$
	1	56	19.4	19.4
	2	31	-5.6	5.6
	3	56	19.4	19.4
	4	8	-28.6	28.6
	5	32	-4.6	4.6

[Additional technical detail[4]]

4.2.4 Variance

Although the average 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 called the **variance**, which has a lot of really nice statistical properties that I'm going to ignore,[5] 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 Var(X), but it's more commonly denoted s^2 (the reason for this will become clearer shortly).

[Additional technical detail[6]]

[4] However, whilst our calculations for this little example are at an end, we do have a couple of things left to talk about. First, 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 so 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:

$$AAD(X) = \frac{1}{N} \sum_{i=1}^{N} \mid X_i - \bar{X} \mid = 15.52$$

[5] Well, 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 $Var(X)$ and $Var(Y)$ 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 $Var(X)+Var(Y)$. This is a very useful property, but it's not true of the other measures that I talk about in this section.

[6] The formula that we use to calculate the variance of a set of observations is as follows:

$$VAR(X) = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

As you can see, it's basically the same formula that we used to calculate the average 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 information shown in Table 4.3.

Table 4.3: Measures of variability for the first five AFL games

English	**maths:**	**value**	**deviation from mean**	**absolute deviation**
notation:	i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
	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

That last column contains all of our squared deviations, so all we have to do is average them. If we do that by hand, i.e. using a calculator, 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 jamovi, because this will reveal something very weird. Start a new jamovi session by clicking on the main menu button (three horizontal lines in the top left corner and selecting 'New'. Now type in the first five values from the *afl.margins* data set in column A (56, 31, 56, 8, 32. Change the variable type to 'Continuous' and under 'Descriptives' click the 'Variance' check box, and you get the same values for variance as the one we calculated by hand (324.64). No, wait, you get a completely different answer (405.80) - see Figure 4.9. That's just weird. Is jamovi broken? Is this a typo? Am I an idiot?

As it happens, the answer is no.[7] It's not a typo, and jamovi is not making a mistake. In fact, it's very simple to explain what jamovi is doing here, but slightly trickier to explain why jamovi is doing it. So let's start with the "what". What jamovi is doing is evaluating a 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, jamovi has chosen to divide by $N - 1$.

[Additional technical detail[8]]

So that's the *what*. The real question is why jamovi 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 8, 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

[7]With the possible exception of the third question.

[8]In other words, the formula that jamovi is using is this one:

$$\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

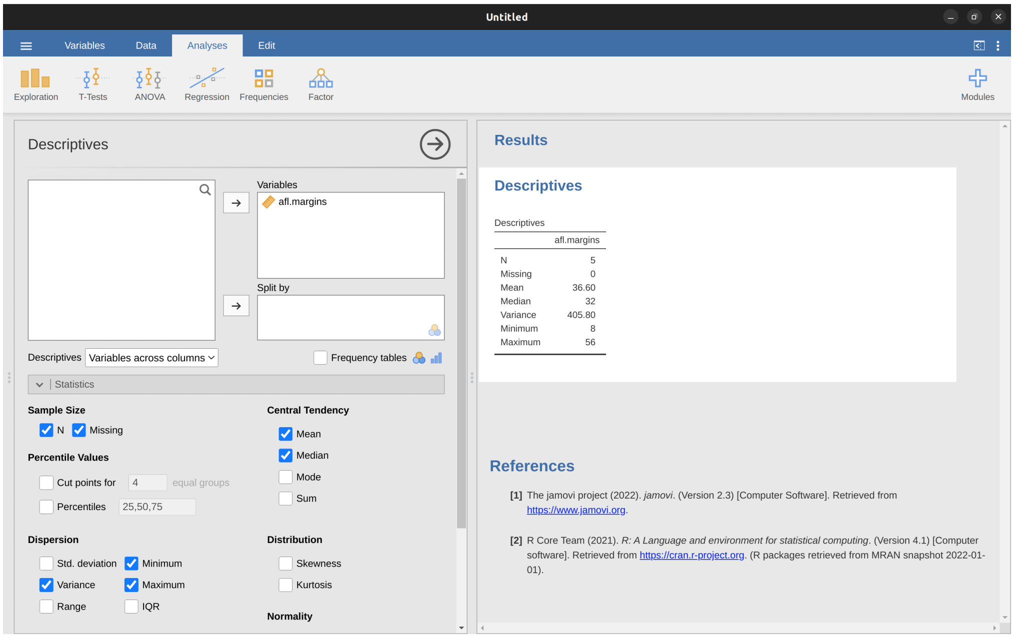


Figure 4.9: A screenshot of jamovi showing the Variance for the first 5 values of the afl.margins variable

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 jamovi knows what it's doing, and we'll revisit the question later on when we talk about estimation in Chapter 8.

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 jamovi 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 Table 4.3, 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 at no time has anyone ever 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.

4.2.5 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 our problem fairly neatly. Whilst 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.

[Additional technical detail[9]]

However, as you might have guessed from our discussion of the variance, what jamovi actually calculates is slightly different to the formula given above. Just like we saw with the variance, what jamovi calculates is a version that divides by $N-1$ rather than N.

[Additional technical detail[10]]

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". As you can tell from looking at the AFL winning margins histogram in Figure 4.2, 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 4.10.

[9] 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 in jamovi there is a check box for 'Std. deviation' right above the check box for 'Variance'. Selecting this gives a value of 26.07 for the standard deviation.

[10] For reasons that will make sense when we return to this topic in Chapter 8 I'll refer to this new quantity as $\hat{\sigma}$ (read as: "sigma hat"), and the formula for this is:

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

4.2.6 Which measure to use?

We've discussed quite a few measures of spread: range, IQR, mean absolute deviation, 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.
- *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.

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.

4.3 Skew and kurtosis

There are two more descriptive statistics that you will sometimes see reported in the psychological literature: skew and kurtosis. In practice, neither one is used anywhere near as frequently as the measures of central tendency and variability that we've been talking about. Skew is pretty important, so you do see it mentioned a fair bit, but I've actually never seen kurtosis reported in a scientific article to date.

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 4.11 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. If there are relatively more values that are far greater than the mean, the distribution is positively skewed or right skewed, with a tail stretching to the right. Negative or left skew is the opposite. A symmetric distribution has a skewness of 0.

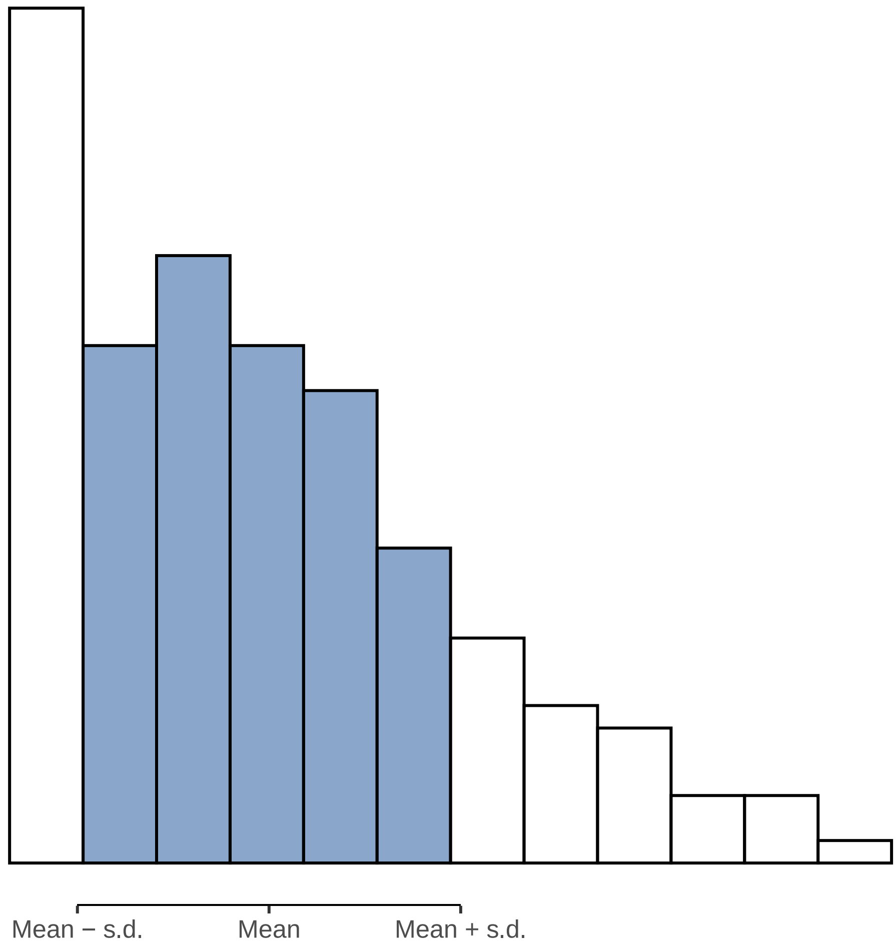


Figure 4.10: An illustration of the standard deviation from the AFL winning margins data. The shaded bars in the histogram show how much of the data fall within one standard deviation of the mean. In this case, 65.3% of the data set lies within this range, which is pretty consistent with the "approximately 68% rule" discussed in the main text

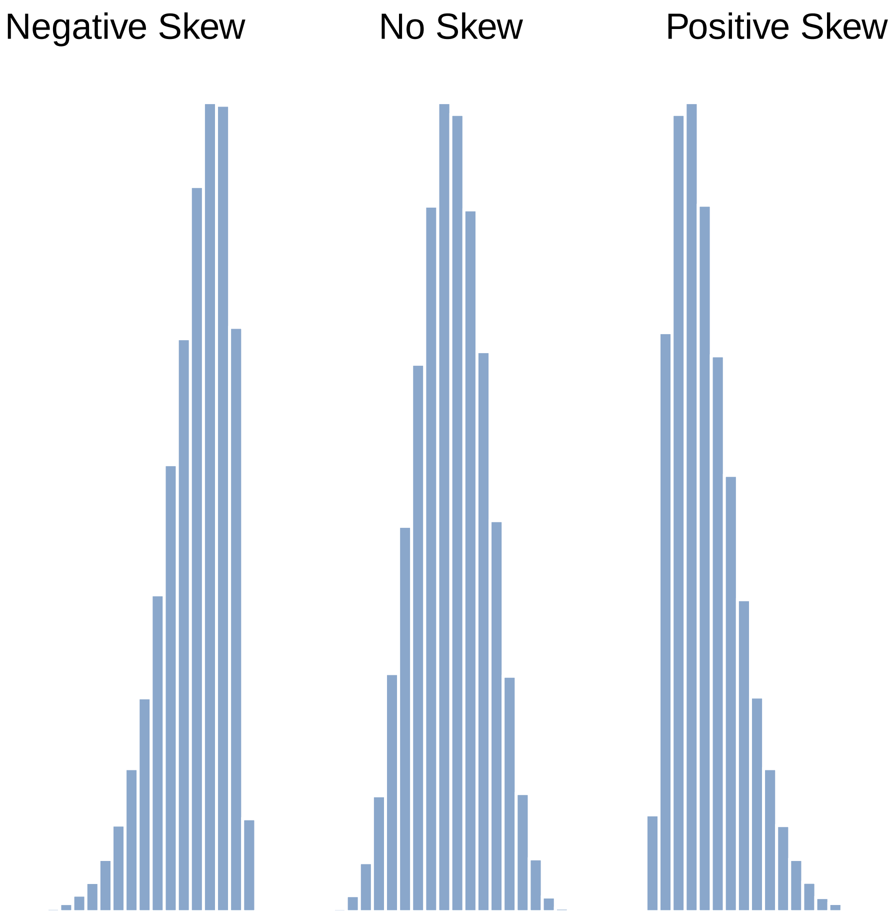


Figure 4.11: An illustration of skewness. On the left we have a negatively skewed data set, in the middle we have a data set with no skew, and on the right we have a positively skewed data set

The skewness value for a positively skewed distribution is positive, and a negative value for a negatively skewed distribution.

[Additional technical detail[11]]

Perhaps more helpfully, you can use jamovi to calculate skewness: it's a check box in the 'Statistics' options under 'Exploration' - 'Descriptives'. For the *afl.margins* variable, the skewness figure is 0.780. If you divide the skewness estimate by the Std. error for skewness you have an indication of how skewed the data is. Especially in small samples (N $<$ 50), one rule of thumb suggests that a value of 2 or less can mean that the data is not very skewed, and a value of over 2 that there is sufficient skew in the data to possibly limit its use in some statistical analyses. Though there is no clear agreement on this interpretation. That said, this does indicate that the AFL winning margins data is somewhat skewed ($\frac{0.780}{0.183} = 4.262$).

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 how thin or fat the tails of a distribution are, as illustrated in Figure 4.12. By convention, we say that the "normal curve" (black lines) has zero kurtosis, so the degree of kurtosis is assessed relative to this curve.

In this Figure, the data on the left have a pretty flat distribution, with thin tails, so the kurtosis is negative and we call the data platykurtic. The data on the right have a distribution with fat tails, so the kurtosis is positive and we say that the data is leptokurtic. But the data in the middle have neither thin or fat tails, so we say that it is mesokurtic and has kurtosis zero. This is summarised in Table 4.4:

Table 4.4: Thin to fat tails to illustrate kurtosis

English	informal term	kurtosis value
"tails too thin"	platykurtic	negative
"tails neither thin or fat"	mesokurtic	zero
"tails too fat"	leptokurtic	positive

[Additional technical detail[12]]

[11] One formula for the skewness of a data set is as follows

$$skewness(X) = \frac{1}{N\hat{\sigma}^3}\sum_{i=1}^{N}(X_i - \bar{X})^3$$

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).

[12] 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:[b]

$$kurtosis(X) = \frac{1}{N\hat{\sigma}^4}\sum_{i=1}^{N}(X_i - \bar{X})^4 - 3$$

I know, it's not terribly interesting to me either. — [b] The "-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.

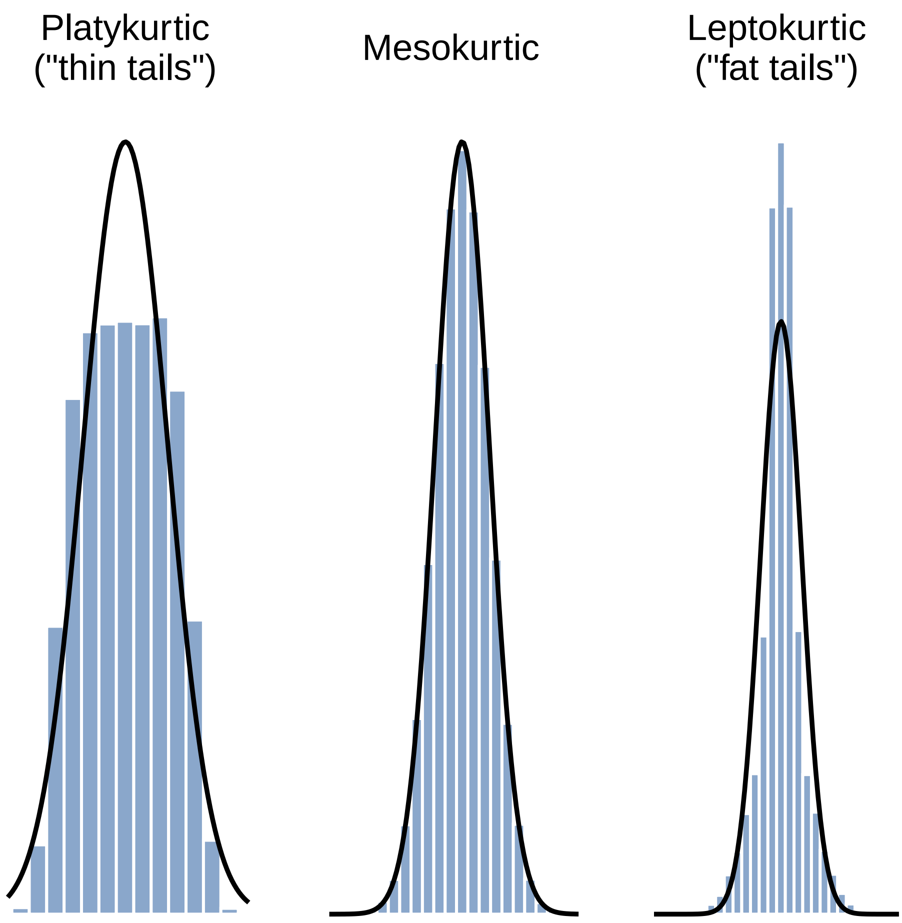


Figure 4.12: An illustration of kurtosis. On the left, we have a "platykurtic" distribution (kurtosis = -.95) meaning that the distribution has "thin" or flat tails. In the middle we have a "mesokurtic" distribution (kurtosis is almost exactly 0) which means that the tails are neither thin or fat. Finally, on the right, we have a "leptokurtic" distribution (kurtosis = 2.12) indicating that the distribution has "fat" tails. Note that kurtosis is measured with respect to a normal curve (black line)

More to the point, jamovi has a check box for kurtosis just below the check box for skewness, and this gives a value for kurtosis of 0.101 with a standard error of 0.364. This means that the AFL winning margins data has only a small kurtosis, which is ok.

4.4 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 jamovi. For instance, let's say I want to look at the descriptive statistics for some clinical trial data, broken down separately by therapy type. This is a new data set, one that you've never seen before. The data is stored in the *clinicaltrial.csv* file and we'll use it a lot later on in **?@sec-Comparing-several-means-one-way-ANOVA** (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 (Figure 4.13):

Evidently there were three drugs: a placebo, something called "anxifree" 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 'Descriptives' of the mood.gain variable that most people did show a mood gain ($mean = 0.88$), 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.

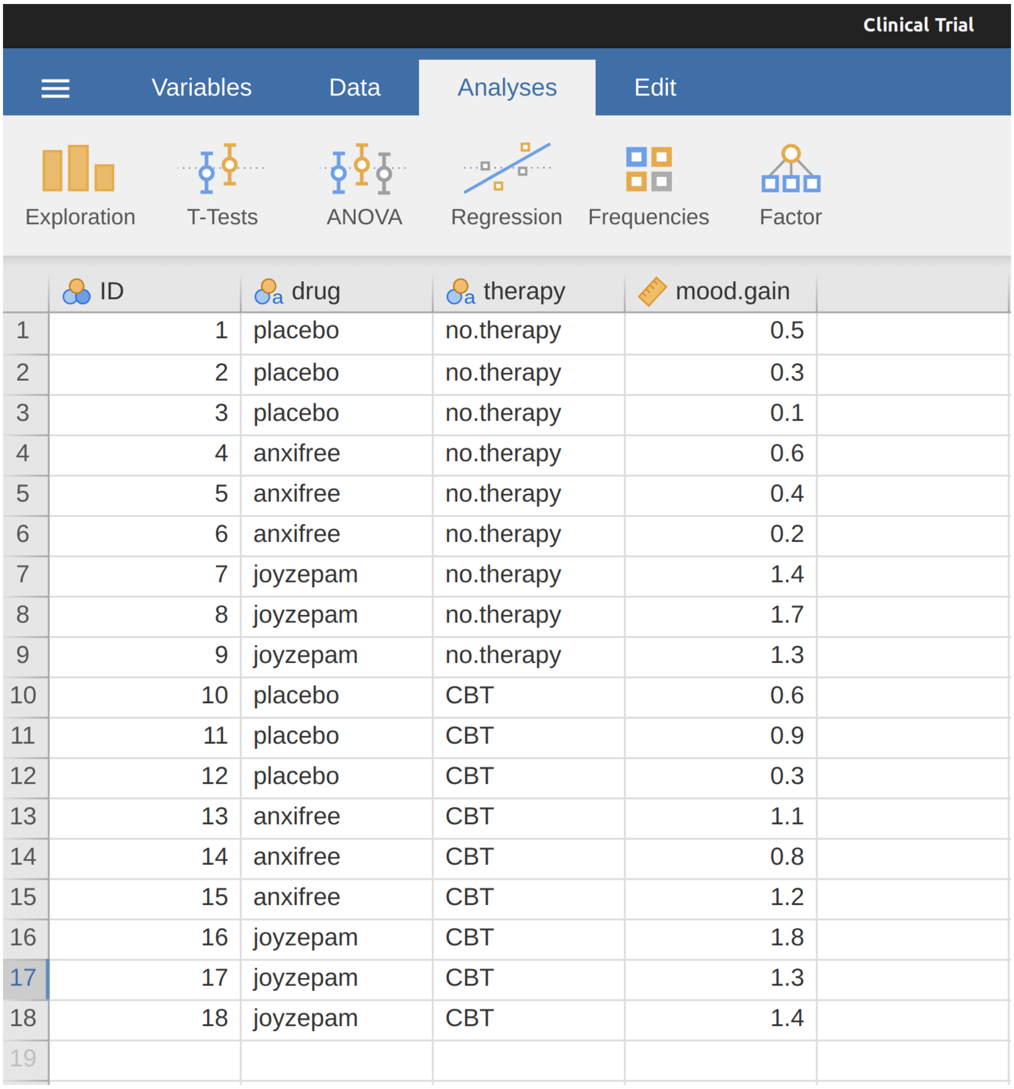
We can also go ahead and look at some other descriptive statistics, and this time separately for each type of therapy. In jamovi, check Std. deviation, Skewness and Kurtosis in the 'Statistics' options. At the same time, transfer the therapy variable into the 'Split by' box, and you should get something like Figure 4.14.

What if you have multiple grouping variables? Suppose you want to look at the average mood gain separately for all possible combinations of drug and therapy. It is possible to do this by adding another variable, drug, into the 'Split by' box. Easy peasy, though sometimes if you split too much there isn't enough data in each breakdown combination to make meaningful calculations. In this case jamovi tells you this by stating something like NaN or Inf.[13]

4.5 Standard scores

Suppose my friend is putting together a new questionnaire intended to measure "grumpiness". The survey has 50 questions which you can answer in a grumpy way or not. Across a big sample (hypothetically, let's imagine a million people or so!) the data are fairly normally distributed, with the mean grumpiness score being 17 out of 50 questions answered in a grumpy way, and the standard deviation is 5. In contrast, when I take the

[13]Sometimes jamovi will also present numbers in an unusual way. If a number is very small, or very large, then jamovi switches to an exponential form for numbers. For example 6.51e-4 is the same as saying that the decimal point is moved 4 places to the left, so the actual number is 0.000651. If there is a plus sign (i.e. 6.51e+4 then the decimal point is moved to the right, i.e. 65,100.00. Usually only very small or very large numbers are expressed in this way, for example 6.51e-16, which would be quite unwieldy to write out in the normal way.



	ID	drug	therapy	mood.gain
1	1	placebo	no.therapy	0.5
2	2	placebo	no.therapy	0.3
3	3	placebo	no.therapy	0.1
4	4	anxifree	no.therapy	0.6
5	5	anxifree	no.therapy	0.4
6	6	anxifree	no.therapy	0.2
7	7	joyzepam	no.therapy	1.4
8	8	joyzepam	no.therapy	1.7
9	9	joyzepam	no.therapy	1.3
10	10	placebo	CBT	0.6
11	11	placebo	CBT	0.9
12	12	placebo	CBT	0.3
13	13	anxifree	CBT	1.1
14	14	anxifree	CBT	0.8
15	15	anxifree	CBT	1.2
16	16	joyzepam	CBT	1.8
17	17	joyzepam	CBT	1.3
18	18	joyzepam	CBT	1.4
19				

Figure 4.13: A screenshot of jamovi showing the variables stored in the *clinicaltrial.csv* file

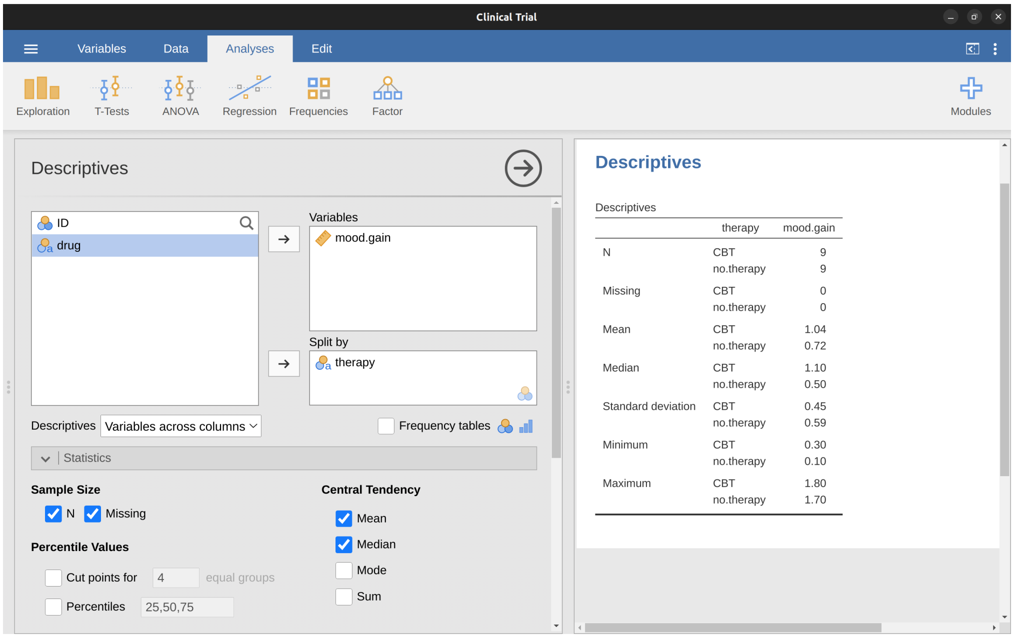


Figure 4.14: A screenshot of jamovi showing Descriptives split by therapy type

questionnaire I answer 35 out of 50 questions in a grumpy way. So, how grumpy am I? One way to think about it would be to say that I have grumpiness of $\frac{35}{50}$, so you might say that I'm 70% grumpy. But that's a bit weird, when you think about it. If my friend had phrased her questions a bit differently people might have answered them in a different way, so the overall distribution of answers could easily move up or down depending on the precise way in which the questions were asked. So, I'm only 70% grumpy *with respect to this set of survey questions*. Even if it's a very good questionnaire this isn't a very informative statement.

A simpler way around this is to describe my grumpiness by comparing me to other people. Shockingly, out of my friend's sample of 1,000,000 people, only 159 people were as grumpy as me (that's not at all unrealistic, frankly) suggesting that I'm in the top 0.016% of people for grumpiness. This makes much more sense than trying to interpret the raw data. This idea, that we should describe my grumpiness in terms of the overall distribution of the grumpiness of humans, is the qualitative idea that standardisation attempts to get at. One way to do this is to do exactly what I just did and describe everything in terms of percentiles. However, the problem with doing this is that "it's lonely at the top". Suppose that my friend had only collected a sample of 1000 people (still a pretty big sample for the purposes of testing a new questionnaire, I'd like to add), and this time had gotten, let's say, a mean of 16 out of 50 with a standard deviation of 5. The problem is that almost certainly not a single person in that sample would be as grumpy as me.

However, all is not lost. A different approach is to convert my grumpiness score into a **standard score**, also referred to as a z-score. The standard score is defined as the number of standard deviations above the mean that my grumpiness score lies. To phrase it in "pseudomaths" the standard score is calculated like this:

$$\text{standard score} = \frac{\text{raw score} - mean}{\text{standard deviation}}$$

[Additional technical detail[14]]

So, going back to the grumpiness data, we can now transform Dani's raw grumpiness into a standardised grumpiness score.

$$z = \frac{35 - 17}{5} = 3.6$$

To interpret this value, recall the rough heuristic that I provided in Section 4.2.5 in which I noted that 99.7% of values are expected to lie within 3 standard deviations of the mean. So the fact that my grumpiness corresponds to a z-score of 3.6 indicates that I'm very grumpy indeed. In fact this suggests that I'm grumpier than 99.98% of people. Sounds about right.

In addition to allowing you to interpret a raw score in relation to a larger population (and thereby allowing you to make sense of variables that lie on arbitrary scales), standard scores serve a second useful function. Standard scores can be compared to one another in situations where the raw scores can't. Suppose, for instance, my friend also had another questionnaire that measured extraversion using a 24 item questionnaire. The overall mean for this measure turns out to be 13 with standard deviation 4, and I scored a 2. As you can imagine, it doesn't make a lot of sense to try to compare my raw score of 2 on the extraversion questionnaire to my raw score of 35 on the grumpiness questionnaire. The raw scores for the two variables are "about" fundamentally different things, so this would be like comparing apples to oranges.

What about the standard scores? Well, this is a little different. If we calculate the standard scores we get ($z = \frac{(35-17)}{5} = 3.6$) for grumpiness and ($z = \frac{(2-13)}{4} = -2.75$) for extraversion. These two numbers can be compared to each other.[15] I'm much less extraverted than most people ($z = -2.75$) and much grumpier than most people ($z = 3.6$). But the extent of my unusualness is much more extreme for grumpiness, since 3.6 is a bigger number than 2.75. Because each standardised score is a statement about where an observation falls relative to its own population, it is possible to compare standardised scores across completely different variables.

4.6 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

[14] In actual maths, the equation for the z-score is

$$z_i = \frac{X_i - \bar{X}}{\hat{\sigma}}$$

[15] Though some caution is usually warranted. It's not always the case that one standard deviation on variable A corresponds to the same "kind" of thing as one standard deviation on variable B. Use common sense when trying to determine whether or not the z-scores of two variables can be meaningfully compared.

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.
- Measures of variability. In contrast, measures of variability tell you about how "spread out" the data are. The key measures are: range, standard deviation, and interquartile range.
- Skew and kurtosis. We also looked at assymetry in a variable's distribution (skew) and thin or fat tailed distributions (kurtosis).
- Descriptive statistics separately for each group. Since this book focuses on doing data analysis in jamovi, we spent a bit of time talking about how descriptive statistics are computed for different subgroups.
- Standard scores. The z-score is a slightly unusual beast. It's not quite a descriptive statistic, and not quite an inference. Make sure you understand this section. It'll come up again later.

In the next chapter we'll move on to a discussion of how to draw pictures! Everyone loves a pretty picture, right? But before we do, I want to end on an important point. 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. With that in mind…

Chapter 5

Drawing graphs

> "*Above all else show the data.*"
> – Edward Tufte[1]

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 5.1 shows a redrawing of one of the most famous data visualisations of all time. This is 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 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 and ended the outbreak that had killed over 500 people. Such is the power of a good data visualisation.

The goals in this chapter are twofold. First, to discuss several fairly standard graphs that we use a lot when analysing and presenting data, and second to show you how to create these graphs in jamovi. The graphs themselves tend to be pretty straightforward, so in one respect this chapter is pretty simple. Where people usually struggle is learning how to produce graphs, and especially learning how to produce good graphs. Fortunately, learning how to draw graphs in jamovi 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 jamovi has a lot of very good default graphs, or plots, that most of the time produce a clean, high-quality graphic. However, on those occasions when you do want to do something

[1] The origin of this quote is Tufte's lovely book *The Visual Display of Quantitative Information.*

non-standard, or if you need to make highly specific changes to the figure, then the graphics functionality in jamovi is not yet capable of supporting advanced work or detail editing.

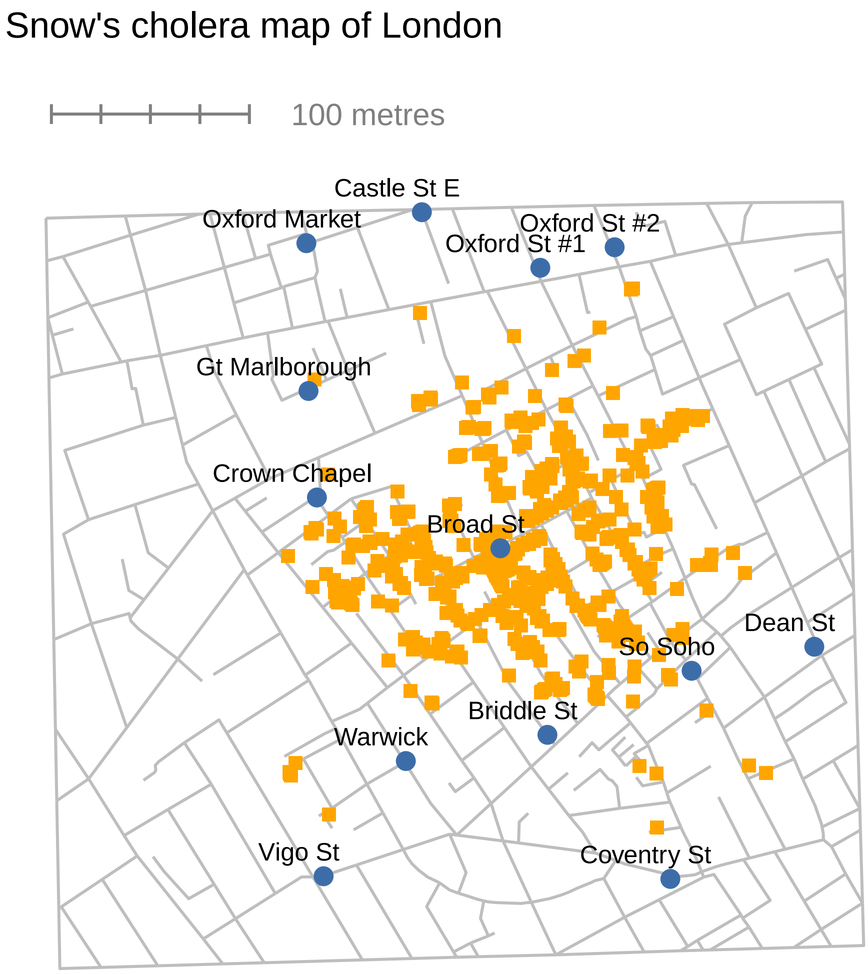


Figure 5.1: A stylised redrawing of John Snow's original cholera map. Each small orange square represents the location of a cholera death and each blue circle shows the location of a water pump. As the plot makes clear, the cholera outbreak is centred very closely on the Broad St pump

5.1 Histograms

Let's 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 variable (e.g., the *afl.margins* data from Chapter 4) and you want to get an overall impression of the variable. 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 or density of the bin and is displayed as a vertical 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 that we showed earlier in Chapter 4, Figure 4.2. With these earlier graphs we used an advanced plotting package in R which, for now, is beyond the capability of jamovi. But jamovi gets us close, and drawing this histogram in jamovi is pretty straightforward. Open up the 'plots' options under 'Exploration' - 'Descriptives' and click the 'histogram' check box, as in Figure 5.1. jamovi defaults to labelling the y-axis as 'density' and the x-axis with the variable name. The **bins** are selected automatically, and there is no scale, or count, information on the y-axis unlike the previous Figure 4.2. But this does not matter too much because after all what we are really interested in is our impression of the shape of the distribution: is it normally distributed or is there a skew or kurtosis? Our first impressions of these characteristics come from drawing a **histogram**.

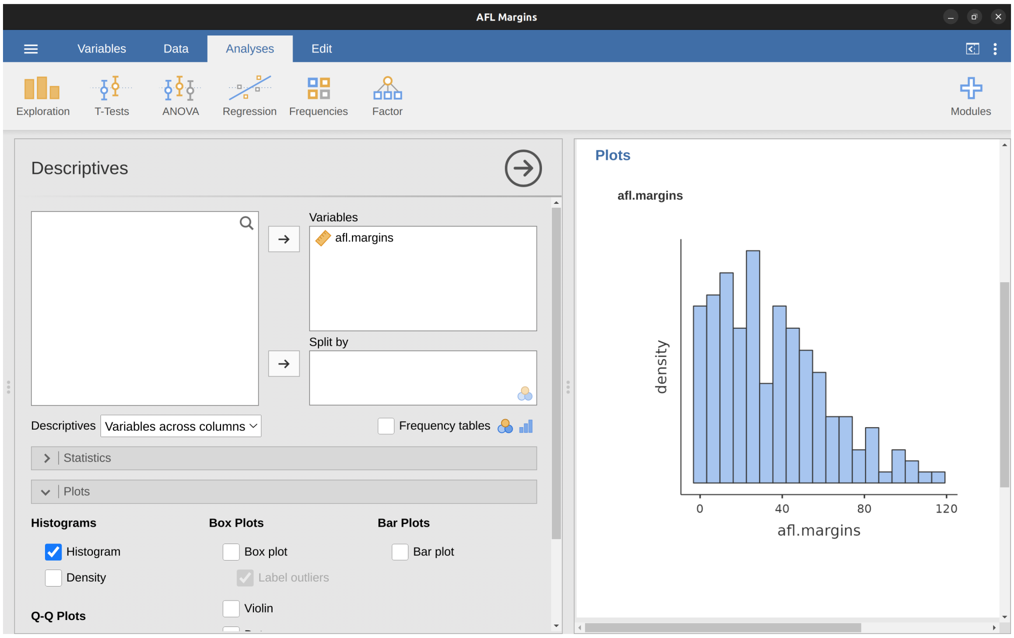


Figure 5.2: jamovi screen showing the histogram check box

One additional feature that jamovi provides is the ability to plot a 'Density' curve. You can do this by clicking the 'Density' check box under the 'Plots' options (and unchecking 'Histogram'), and this gives us the plot shown in Figure 5.3. A density plot visualises the distribution of data over a continuous interval or time period. This chart is a variation of a histogram that uses **kernel smoothing** to plot values, allowing for smoother distributions by smoothing out the noise. The peaks of a density plot help display where values are concentrated over the interval. An advantage density plots have over histograms is that they are better at determining the distribution shape because they're not affected by the number of bins used (each bar used in a typical histogram). A histogram comprising of only 4 bins wouldn't produce a distinguishable enough shape of distribution as a 20-bin histogram would. However, with density plots, this isn't an issue.

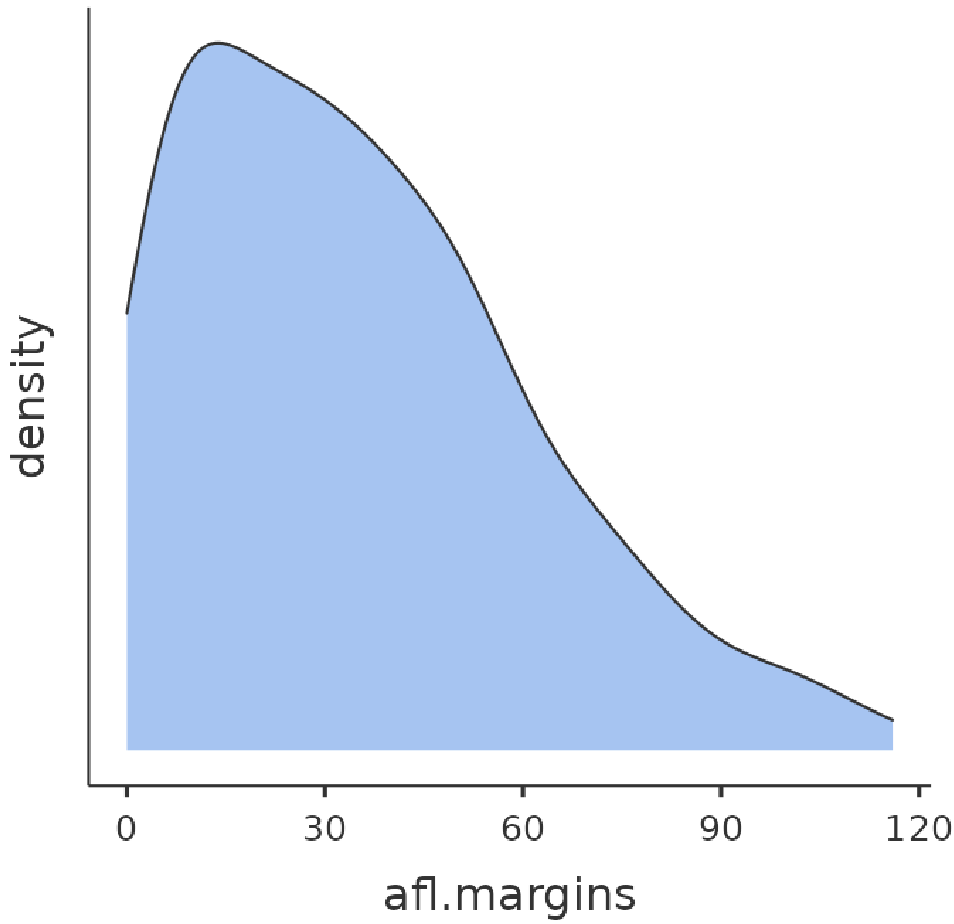


Figure 5.3: A density plot of the *afl.margins* variable plotted in jamovi

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 or density plot 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 then histograms are useless.

5.2 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.

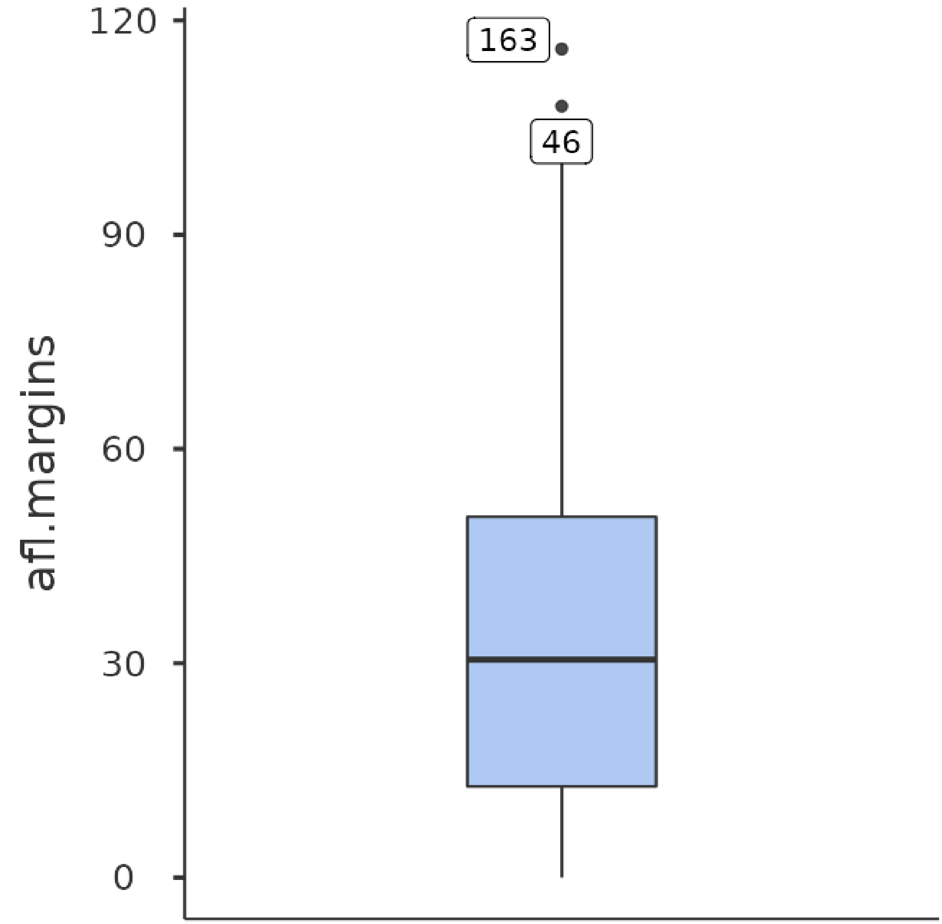


Figure 5.4: A box plot of the *afl.margins* variable plotted in jamovi

The easiest way to describe what a boxplot looks like is just to draw one. Click on the 'Box plot' check box and you will get the plot shown on the lower right of Figure 5.4. jamovi has drawn 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" 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 (IQR), calculated as 25th percentile - (1.5*IQR) for the lower boundary, and 75th percentile + (1.5*IQR) for the upper boundary. Any observation whose value falls outside this range is plotted as a circle or dot instead of being covered by the whiskers, and is commonly referred to as an **outlier**. For our *afl.margins* data there are two observations that fall outside this range, and these observations are plotted as dots (the upper boundary is 107, and looking over the data column in the spreadsheet there are two observations with values higher than this, 108 and 116, so these are the dots).

5.2.1 Violin plots

A variation to the traditional box plot is the violin plot. Violin plots are similar to box plots except that they also show the kernel probability density of the data at different values. Typically, violin plots will include a marker for the median of the data and a box indicating the interquartile range, as in standard box plots. In jamovi you can achieve this sort of functionality by checking both the 'Violin' and the 'Box plot' check boxes. See Figure 5.5, which also has the 'Data' check box turned on to show the actual data points on the plot. This does tend to make the graph a bit too busy though, in my opinion. Clarity is simplicity, so in practice it might be better to just use a simple box plot.

5.2.2 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 **aflmarginbyyear.csv** file. So let's load it into jamovi and see what is in it. You will see that it is a pretty big data set. It contains 4296 games and the variables that we're interested in. What we want to do is have jamovi draw boxplots for the margin variable, but plotted separately for each year. The way to do this is to move the year variable across into the 'Split by' box, as in Figure 5.6.

The result is shown in Figure 5.7. This version of the box plot, split by year, gives a sense of why it's sometimes useful to choose box plots instead of histograms. 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.

5.2.3 Using box plots to detect outliers

Because the boxplot automatically separates out those observations that lie outside a certain range, depicting them with a dot in jamovi, 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*

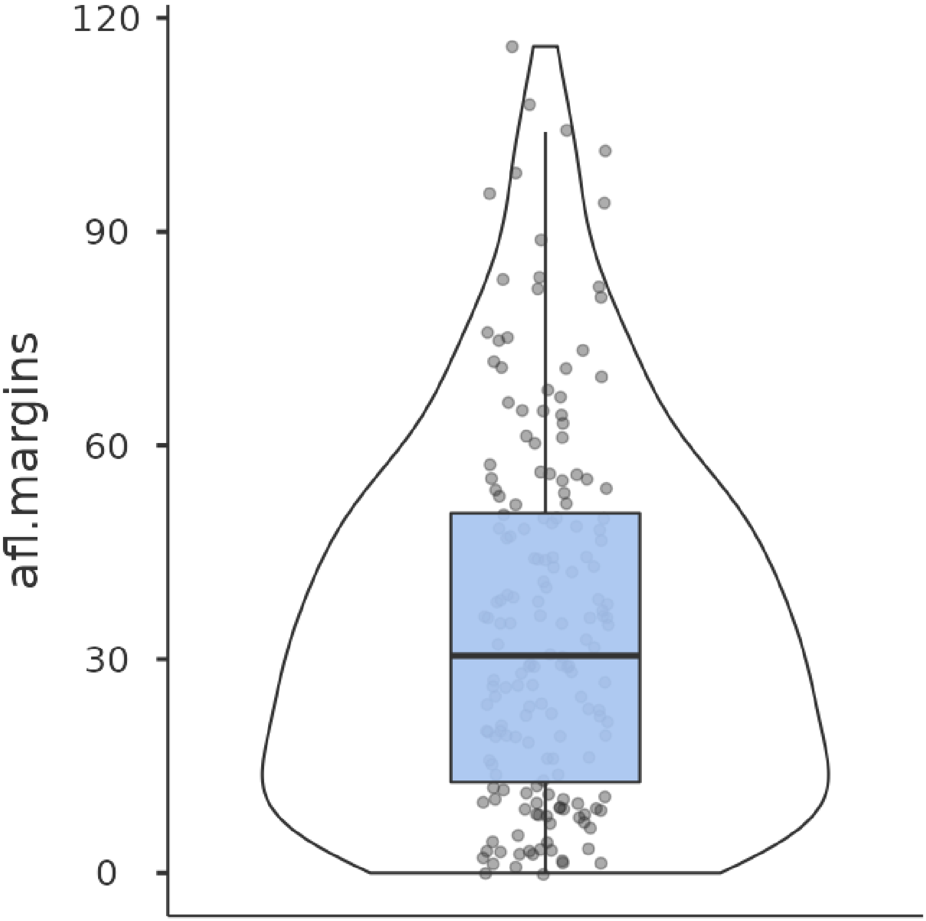


Figure 5.5: A violin plot of the *afl.margins* variable plotted in jamovi, also showing a box plot and data points

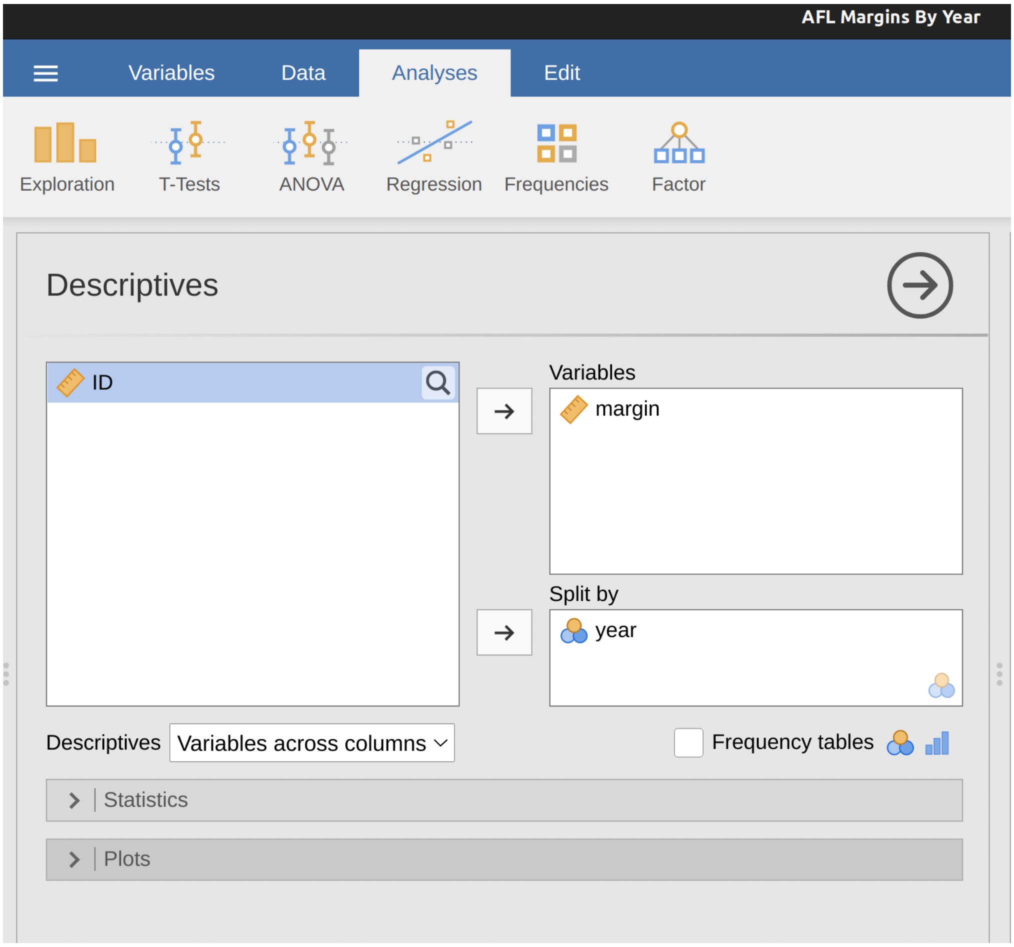


Figure 5.6: jamovi screen shot showing the ‘Split by’ window

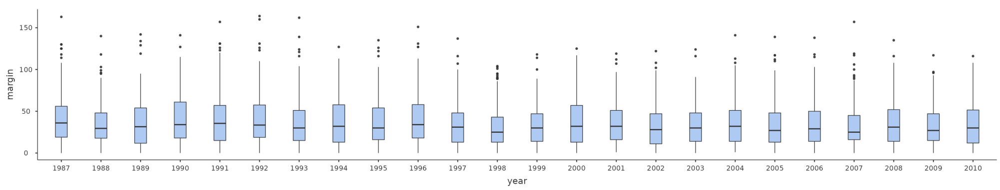


Figure 5.7: Multiple boxplots plotted in jamovi, for the margin by year variables

data and it came up looking like Figure 5.8. It's pretty clear that something funny is going on with two of the observations. Apparently, there were two games in which the margin was over 300 points![2] 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. In jamovi you can quickly find out which of these observations are suspicious and then you can go back to the raw data to see if there has been a mistake in data entry. One way to do this is to tell jamovi to label the outliers, by checking the box next to the 'Box plot' check box. This adds a row number label next to the outlier in the boxplot, so you can go look at that row and find the extreme value. Another, more flexible way, is to set up a filter so that only those observations with values over a certain threshold are included. In our example, the threshold is over 300, so that is the filter we will create. First, click on the 'Filters' button at the top of the jamovi window, and then type 'margin > 300' into the filter field, as in Figure 5.9.

This filter creates a new column in the spreadsheet view where only those observations that pass the filter are included. One neat way to quickly identify which observations these are is to tell jamovi to produce a 'Frequency table' (in the 'Exploration' - 'Descriptives' window) for the ID variable (which must be a nominal variable otherwise the Frequency table is not produced). In Figure 5.10 you can see that the ID values for the observations where the margin was over 300 are 176 and 202. These are suspicious cases, or observations, where you should go back to the original data source to find out what is going on.

Usually you find that someone has just typed in the wrong number. Whilst 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 and in practice it can take up a huge chunk of our time: data cleaning. It involves searching for typing mistakes ("typos"), missing data and all sorts of other obnoxious errors in raw data files.

For less extreme values, even if they are flagged in a a boxplot as outliers, the decision about whether to include outliers or exclude them in any analysis depends heavily on why you think the data look they way they do and what you want to use the data for. You really need to exercise good judgement here. If the outlier looks legitimate to you, then keep it. In any case, I'll return to the topic again in **?@sec-Model-checking** in **?@sec-Correlation-and-linear-regression**.

5.3 Bar graphs

Another form of graph that you often want to plot is the **bar graph**. Let's use the *afl.finalists* data set with the *afl.finalists* variable that I introduced in Section 4.1.6. 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.finalists* data set. There are lots of teams, but I am particularly interested in just four: Brisbane, Carlton, Fremantle and Richmond. So the first step is to set up a filter so just those four teams are included in

[2] This altered version of the AFL Margins By Year dataset isn't available to open / load into jamovi. I simply changed a couple of the values of margin in the dataset so that they were over 300. You can do this yourself if you want.

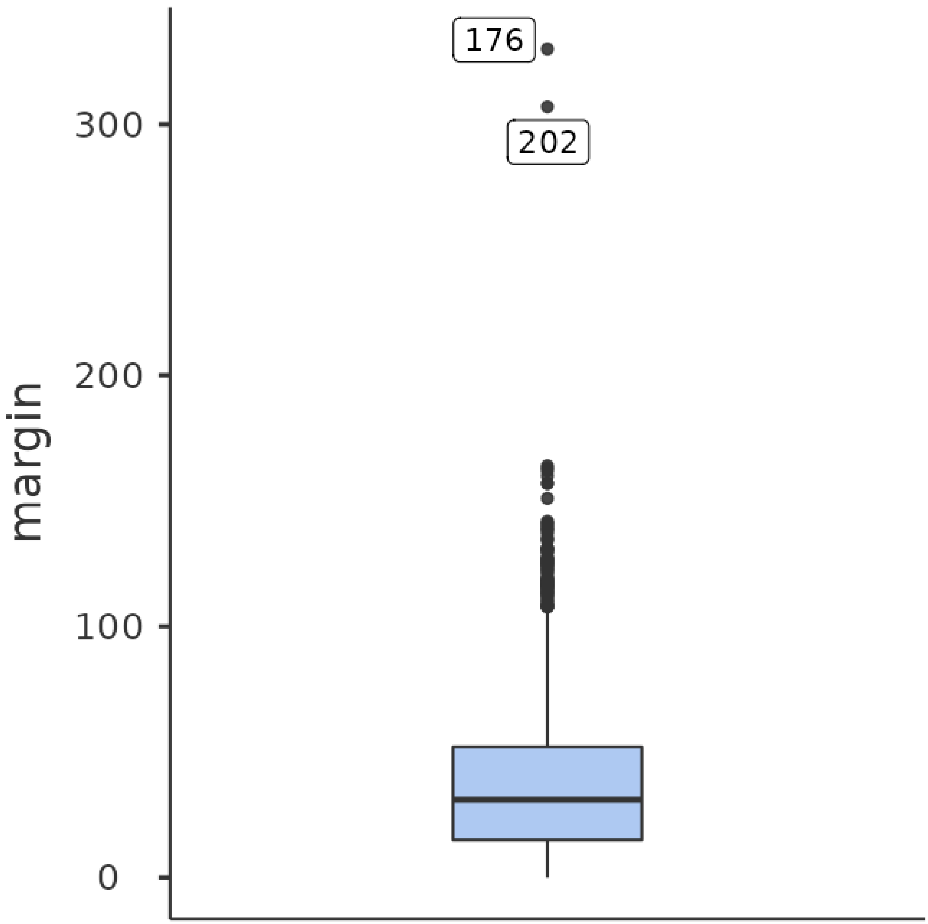


Figure 5.8: A boxplot showing two very suspicious outliers!

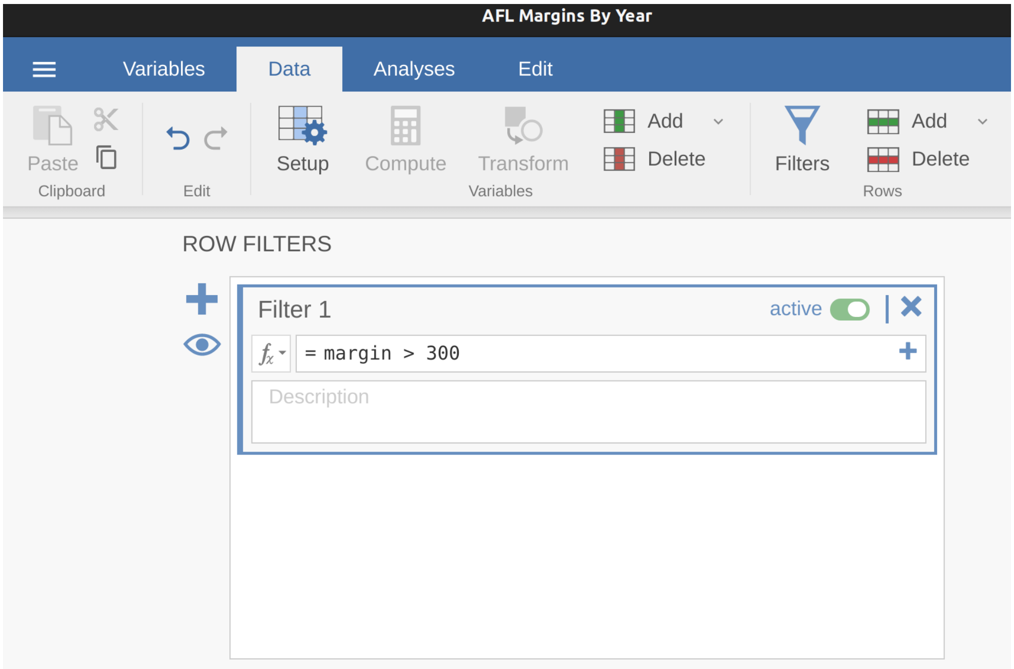


Figure 5.9: The jamovi filter screen

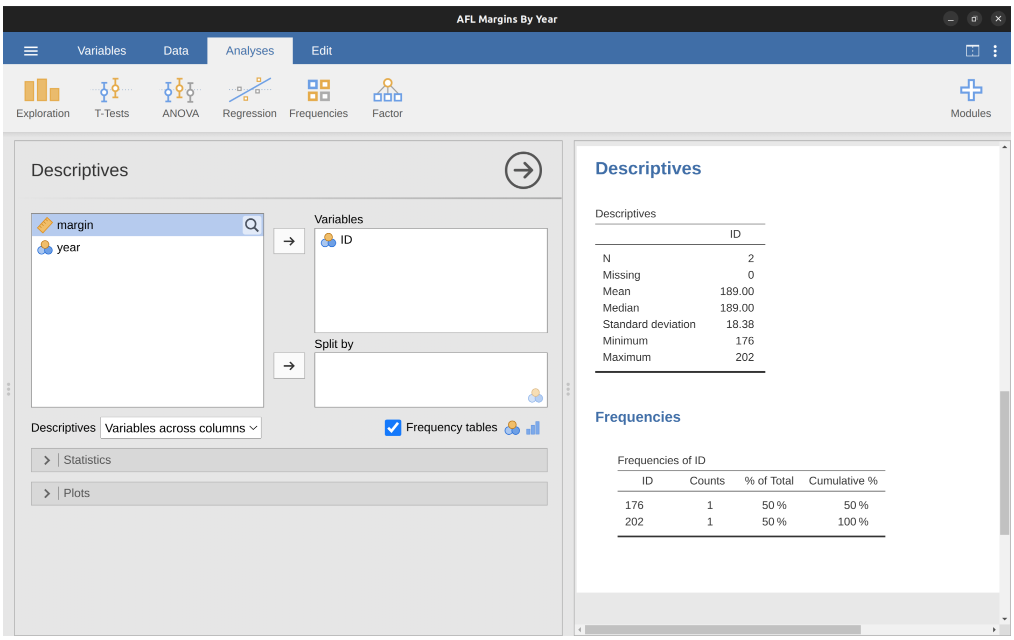


Figure 5.10: Frequency table for ID showing the ID numbers for the two suspicious outliers, 176 and 202

the bar graph. This is straightforward in jamovi and you can do it by using the ‘Filters’ function that we used previously. Open up the ‘Filters’ screen and type in the following exactly as written – including the single quote marks:

> afl.finalists == ‘Brisbane’ or afl.finalists == ‘Carlton’ or afl.finalists == ‘Fremantle’ or afl.finalists == ‘Richmond’ [3]

When you have done this you will see, in the ‘Data’ view, that jamovi has filtered out all values apart from those we have specified. Next, open up the ‘Exploration’ - ‘Descriptives’ window and click on the ‘Bar plot’ check box (remember to move the ‘afl.finalists’ variable across into the ‘Variables’ box so that jamovi knows which variable to use). You should then get a bar graph, something like that shown in Figure 5.11.

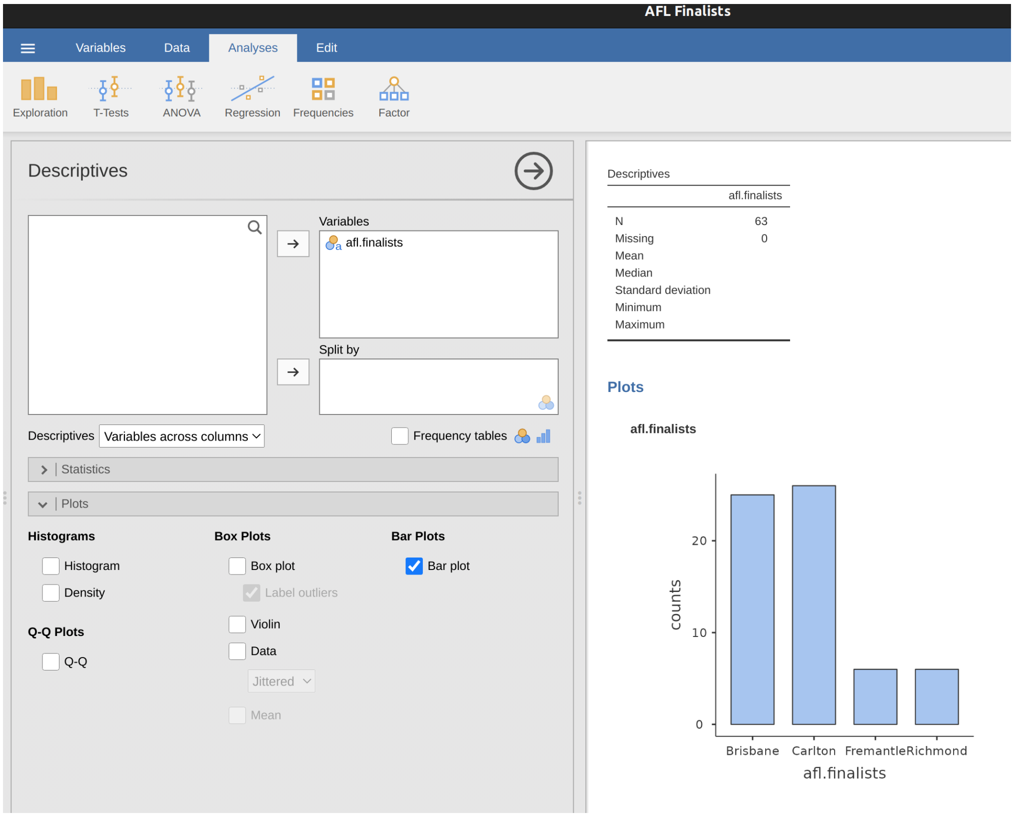


Figure 5.11: Filtering to include just four AFL teams, and drawing a bar plot in jamovi

5.4 Saving image files using jamovi

Hold on, you might be thinking. What’s the good of being able to draw pretty pictures in jamovi if I can’t save them and send them to friends to brag about how awesome my

[3]jamovi uses the symbol “==” here to mean “matches”.

data is? How do I save the picture? Simples. Just right click on the plot image and export it to a file, either as 'png', 'eps', 'svg' or 'pdf'. These formats all produce nice images that you can then send to your friends, or include in your assignments or papers.

5.5 Summary

Perhaps I'm a simple-minded person, but I love pictures. Every time I write a new scientific paper one of the first things I do is sit down and think about what the pictures will be. In my head an article is really just a sequence of pictures linked together by a story. All the rest of it is just window dressing. What I'm really trying to say here is that the human visual system is a very powerful data analysis tool. Give it the right kind of information and it will supply a human reader with a massive amount of knowledge very quickly. Not for nothing do we have the saying "a picture is worth a thousand words". With that in mind, I think that this is one of the most important chapters in the book. The topics covered were:

- *Common plots.* Much of the chapter was focused on standard graphs that statisticians like to produce: Histograms, Boxplots and Bar graphs.
- Saving image files using jamovi. Importantly, we also covered how to export your pictures.

One final thing to point out. Whilst jamovi produces some really neat default graphics, editing the plots is currently not possible. For more advanced graphics and plotting capability the packages available in R are much more powerful. One of the most popular graphics systems is provided by the ggplot2 package (see https://ggplot2.tidyverse.org/), which is loosely based on "The grammar of graphics" (Wilkinson et al., 2006). It's not for novices. You need to have a pretty good grasp of R before you can start using it, and even then it takes a while to really get the hang of it. But when you're ready it's worth taking the time to teach yourself, because it's a much more powerful and cleaner system.

Chapter 6

Pragmatic matters

"*The garden of life never seems to confine itself to the plots philosophers have laid out for its convenience. Maybe a few more tractors would do the trick.*"
– Roger Zelazny[1]

This is a somewhat strange chapter, even by my standards. My goal in this chapter is to talk a bit more honestly about the realities of working with data than you'll see anywhere else in the book. The problem with real world data sets is that they are *messy*. Very often the data file that you start out with doesn't have the variables stored in the right format for the analysis you want to do. Sometimes there might be a lot of missing values in your data set. Sometimes you only want to analyse a subset of the data. Et cetera. In other words, there's a lot of **data manipulation** that you need to do just to get the variables in your data set into the format that you need it. The purpose of this chapter is to provide a basic introduction to these pragmatic topics. Although the chapter is motivated by the kinds of practical issues that arise when manipulating real data, I'll stick with the practice that I've adopted throughout most of the book and rely on very small, toy data sets that illustrate the underlying issue. Because this chapter is essentially a collection of techniques and doesn't tell a single coherent story, it may be useful to start with a list of topics:

- Tabulating and cross-tabulating data.
- Logical expressions in jamovi.
- Transforming and recoding a variable.
- A few more mathematical functions and operations.
- Extracting a subset of the data.

As you can see, the list of topics that the chapter covers is pretty broad, and there's a lot of content there. Even though this is one of the longest and hardest chapters in the book, I'm really only scratching the surface of several fairly different and important topics. My advice, as usual, is to read through the chapter once and try to follow as much of it as you can. Don't worry too much if you can't grasp it all at once, especially

[1]The quote comes from *Home is the Hangman*, published in 1975.

the later sections. The rest of the book is only lightly reliant on this chapter so you can get away with just understanding the basics. However, what you'll probably find is that later on you'll need to flick back to this chapter in order to understand some of the concepts that I refer to here.

6.1 Tabulating and cross-tabulating data

A very common task when analysing data is the construction of frequency tables, or crosstabulation of one variable against another. These tasks can be achieved in jamovi and I'll show you how in this section.

6.1.1 Creating tables for single variables

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.csv* file, I've transcribed a short section of the dialogue. The file contains two variables of interest, speaker and utterance. Open up this data set in jamovi and take a look at the data in the 'spreadsheet' view. You will see that the data looks something like this:

'speaker' variable: upsy-daisy upsy-daisy upsy-daisy upsy-daisy tombliboo tombliboo makka-pakka makka-pakka makka-pakka makka-pakka 'utterance' variable: pip pip onk onk ee oo pip pip onk onk

Looking at this it becomes very clear what happened to my sanity! 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 jamovi 'Descriptives' screen has a check box called 'Frequency tables' which does just this, see Table 6.1.

Table 6.1: Frequency table for the speaker variable

levels	**Counts**	**% of Total**	**Cumulative %**
makka-pakka	4	40%	40%
tombliboo	2	20%	60%
upsy-daisy	4	40%	100%

The output here tells us on the first line that what we're looking at is a tabulation of the speaker variable. In the 'Levels' column it lists all the different speakers that exist in the data, and in the 'Counts' column it tells you how many times that speaker appears in the data. In other words, it's a frequency table.

In jamovi, the 'Frequency tables' check box will only produce a table for single variables. For a table of two variables, for example combining speaker and utterance so that we can see how many times each speaker said a particular utterance, we need a cross-tabulation or contingency table. In jamovi you can do this by selecting the 'Frequencies' - 'Contingency Tables' - 'Independent Samples' analysis, and moving the speaker variable into the 'Rows' box, and the utterances variable into the 'Columns' box. You then should have a contingency table like the one shown in Figure 6.1.

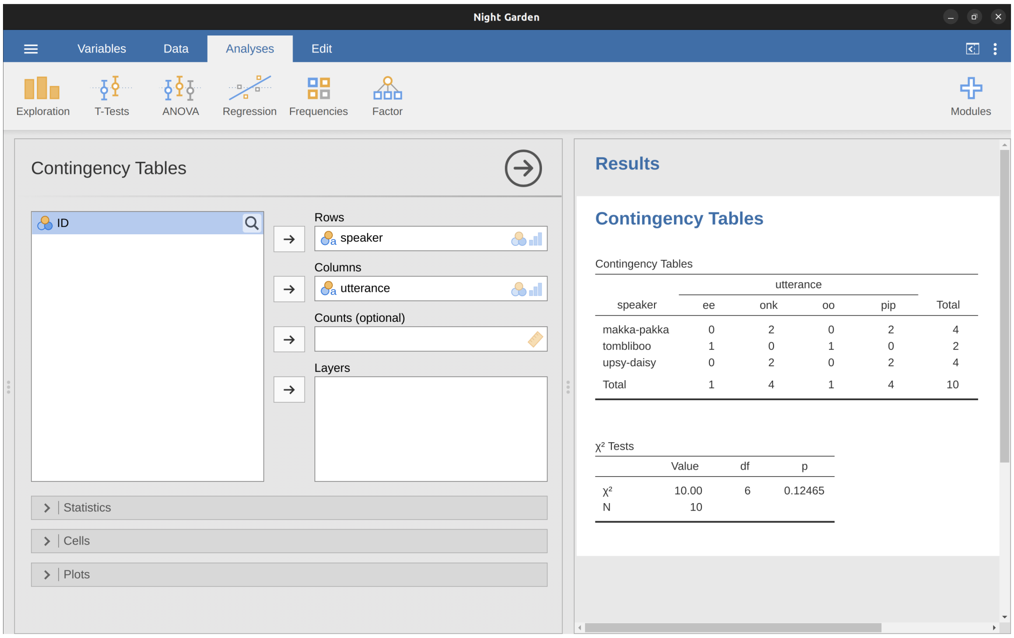


Figure 6.1: Contingency table for the speaker and utterances variables

Don't worry about the "χ^2 Tests" table that is produced. We are going to cover this later on in **?@sec-Categorical-data-analysis**. When interpreting the contingency table remember that these are counts, so the fact that the first row and second column of numbers corresponds to a value of 2 indicates that Makka-Pakka (row 1) says "onk" (column 2) twice in this data set.

6.1.2 Adding percentages to a contingency table

The contingency table shown in Figure 6.1 shows a table of raw frequencies. That is, a count of the total number of cases for different combinations of levels of the specified variables. However, often you want your data to be organised in terms of percentages as well as counts. You can find the check boxes for different percentages under the 'Cells' option in the 'Contingency Tables' window. First, click on the 'Row' check box and the Contingency Table in the output window will change to the one in Figure 6.2.

What we're looking at here is the percentage 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 table we get when we calculate column percentages (uncheck 'Row' and check 'Column' in the Cells options window), see Figure 6.3. In this version, what we're seeing is the percentage 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.

Contingency Tables

Contingency Tables

speaker		utterance				
		ee	onk	oo	pip	Total
makka-pakka	Observed	0	2	0	2	4
	% within row	0 %	50 %	0 %	50 %	100 %
tombliboo	Observed	1	0	1	0	2
	% within row	50 %	0 %	50 %	0 %	100 %
upsy-daisy	Observed	0	2	0	2	4
	% within row	0 %	50 %	0 %	50 %	100 %
Total	Observed	1	4	1	4	10
	% within row	10 %	40 %	10 %	40 %	100 %

Figure 6.2: Contingency table for the speaker and utterances variables, with row percentages

Contingency Tables

Contingency Tables

speaker		utterance				
		ee	onk	oo	pip	Total
makka-pakka	Observed	0	2	0	2	4
	% within column	0 %	50 %	0 %	50 %	40 %
tombliboo	Observed	1	0	1	0	2
	% within column	100 %	0 %	100 %	0 %	20 %
upsy-daisy	Observed	0	2	0	2	4
	% within column	0 %	50 %	0 %	50 %	40 %
Total	Observed	1	4	1	4	10
	% within column	100 %	100 %	100 %	100 %	100 %

Figure 6.3: Contingency table for the speaker and utterances variables, with column percentages

6.2 Logical expressions in jamovi

A key concept that a lot of data transformations in jamovi rely 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 jamovi in a pretty straightforward way. There are two logical values, namely TRUE and FALSE. Despite the simplicity, logical values are very useful things. Let's see how they work.

6.2.1 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 humans[2] and it's definitely not true of jamovi. jamovi 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 $2 + 2$,[3] it always gives the same answer, and it's not bloody 5!

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

What I've done here is use the **equality operator**, ==, to force jamovi to make a "true or false" judgement.[4] Okay, let's see what jamovi thinks of the Party slogan, so type this into the compute new variable 'formula' box:

$$2 + 2 == 5$$

And what do you get? It should be a whole set of 'false' values in the spreadsheet column for your newly computed variable. It was worth having a look at what happens if I try to force jamovi to believe that two plus two is five by making a statement like $2 + 2 = 5$. I know that if I do this in another program, say R, then it throws up an error message. But wait, if you do this in jamovi you get a whole set of 'false' values. So what is going on? Well, it seems that jamovi is being pretty smart and realises that you are testing whether it is TRUE or FALSE that $2 + 2 = 5$, regardless of whether you use the correct **equality operator** ==, or the equals sign =.

[2] I offer up my teenage attempts to be "cool" as evidence that some things just can't be done.

[3] You can do this in the 'Compute new variable' screen, though just calculating 2 + 2 for every cell of a new variable is not very useful!

[4] Note that this is a very different operator to the equals operator =. A common typo that people make when trying to write logical commands in jamovi (or other languages, since the "= versus ==" distinction is important in many computer and statistical programs) 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.

6.2.2 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: $3 \times 3 + 4 \times 4 == 5 \times 5$ or this $SQRT(25) == 5$

Not only that, but as Table 6.2 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 jamovi returns an answer of TRUE, but if the two numbers are equal, or if the one on the right is larger, then jamovi returns an answer of 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 on the left-hand side is less than or equal to the number on the right-hand side. 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 would get 'true' as the value for our newly computed variable. Try it and see:

$$2 + 2 \; != 5$$

We're not quite done yet. There are three more logical operations that are worth knowing about, listed in Table 6.3. These are the **not** operator !, the **and** operator *and*, and the **or** operator *or*. 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 or operator does:[5]

$$(2 + 2 == 4) \text{ or } (2 + 2 == 5)$$

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 *and* operator does:

$$(2 + 2 == 4) \text{ and } (2 + 2 == 5)$$

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

[5] Now, here's a quirk in jamovi. When you have simple logical expressions like the ones we have already met, e.g. $2+2 == 5$ then jamovi neatly states 'false' (or 'true') in the corresponding spreadsheet column. Underneath the hood, jamovi stores 'false' as 0 and 'true' as 1. When we have more complex logical expressions, such as $(2 + 2 == 4)$ or $(2 + 2 == 5)$, then jamovi just displays either 0 or 1, depending whether the logical expression is evaluated as false, or true.

Table 6.2: Some logical operators

operation	**operator**	**example input**	**answer**
less than	$<$	$2 < 3$	TRUE
less than or equal to	$<=$	$2 <= 2$	TRUE
greater than	$>$	$2 > 3$	FALSE
greater than or equal to	$>=$	$2 >= 2$	TRUE
equal to	$==$	$2 == 3$	FALSE
not equal to	!=	2 != 3	TRUE

Table 6.3: Some more logical operators

operation	**operator**	**example input**	**answer**
not	NOT	NOT(1==1)	FALSE
or	or	(1==1) or (2==3)	TRUE
and	and	(1==1) and (2==3)	FALSE

that my claim is true, because actually my claim is that "$2 + 2 = 5$ is false". And I'm right. If we write this in jamovi we use this:

$$NOT(2 + 2 == 5)$$

In other words, since $2 + 2 == 5$ is a FALSE statement, it must be the case that $NOT(2 + 2 == 5)$ 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 jamovi lives in a much more black or white world. For jamovi everything is either true or false. No shades of grey are allowed.

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

$$2 + 2 \mathrel{!}= 5$$

6.2.3 Applying logical operation to text

I also want to briefly point out that you can apply these logical operators to text as well as to logical data. It's just that we need to be a bit more careful in understanding how jamovi 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 ! =.

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

"cat" == "dog" That's pretty obvious, and it's good to know that even jamovi can figure that out. Similarly, jamovi does recognise that a "cat" is a "cat": "cat" == "cat" Again, that's exactly what we'd expect. However, what you need to keep in mind is that jamovi is not at all tolerant when it comes to grammar and spacing. If two strings differ in any way whatsoever, jamovi will say that they're not equal to each other, as with the following: " cat" == "cat" "cat" == "CAT" "cat" == "c a t"

You can also use other logical operators too. For instance jamovi also allows you to use the > and < operators to determine which of two text 'strings' comes first, alphabetically speaking. Sort of. Actually, it's a bit more complicated than that, but let's start with a simple example:

"cat" < "dog"

In jamovi, this example evaluates to 'true'. This is because "cat" does does come before "dog" alphabetically, so jamovi judges the statement to be true. However, if we ask jamovi to tell us if "cat" comes before "anteater" then it will evaluate the expression as false. So far, so good. But text data is a bit more complicated than the dictionary suggests. What about "cat" and "CAT"? Which of these comes first? Try it and find out:

"CAT" < "cat"

This in fact evaluates to 'true'. In other words, jamovi assumes that uppercase letters come before lowercase ones. Fair enough. No-one is likely to be surprised by that. What you might find surprising is that jamovi assumes that all uppercase letters come before all lowercase ones. That is, while "anteater" < "zebra" is a true statement, and the uppercase equivalent "ANTEATER" < "ZEBRA" is also true, it is not true to say that "anteater" < "ZEBRA", as the following extract illustrates. Try this:

"anteater" < "ZEBRA"

This evaluates to 'false', and this may seem slightly counter-intuitive. With that in mind, it may help to have a quick look at Table 6.4 which lists various text characters in the order that jamovi processes them.

6.3 Transforming and recoding a variable

It's not uncommon in real-world data analysis to find that one of your variables isn't quite equivalent to the variable that you really want. For instance, it's often convenient to take a continuous-valued variable (e.g., age) and break it up into a smallish number of categories (e.g., younger, middle, older). At other times, you may need to convert a numeric variable into a different numeric variable (e.g., you may want to analyse at the absolute value of the original variable). In this section I'll describe a few key ways you can do these things in jamovi.

Table 6.4: Text characters in the order that jamovi processes them

!	”	#	$	%	&	’	(
)	*	+	,	-	.	/	0
1	2	3	4	5	6	7	8
9	:	;	<	=	>	?	@
A	B	C	D	E	F	G	H
I	J	K	L	M	N	O	P
Q	R	S	T	U	V	W	X
Y	Z	[	\	]	ˆ	_	‘
a	b	c	d	e	g	h	i
j	k	l	m	n	o	p	q
r	s	t	u	v	w	x	y
z	{	|	}				

6.3.1 Creating a transformed variable

The first trick to discuss is the idea of **transforming** a variable. Taken literally, anything you do to a variable is a transformation, but in practice what it usually means is that you apply a relatively simple mathematical function to the original variable in order to create a new variable that either (a) provides a better way of describing the thing you're actually interested in, or (b) is more closely in agreement with the assumptions of the statistical tests you want to do. Since, at this stage, I haven't talked about statistical tests or their assumptions, I'll show you an example based on the first case.

Suppose I've run a short study in which I ask 10 people a single question: On a scale of 1 (strongly disagree) to 7 (strongly agree), to what extent do you agree with the proposition that “Dinosaurs are awesome”?

Now let's load and look at the data. The data file *likert.omv* contains a single variable that contains raw Likert-scale responses for these 10 people. However, if you think about it, this isn't the best way to represent these responses. Because of the fairly symmetric way that we set up the response scale, there's a sense in which the midpoint of the scale should have been coded as 0 (no opinion), and the two endpoints should be “3 (strongly agree)” and “3 (strongly disagree)”. By recoding the data in this way it's a bit more reflective of how we really think about the responses. The recoding here is pretty straightforward, we just subtract 4 from the raw scores. In jamovi you can do this by computing a new variable: click on the ‘Data’ - ‘Compute’ button and you will see that a new variable has been added to the spreadsheet. Let's call this new variable ‘likert.centred’ (go ahead and type that in) and then add the following in the formula box, like in Figure 6.4: ‘likert.raw - 4’

One reason why it might be useful to have the data in this format is that there are a lot of situations where you might prefer to analyse the strength of the opinion separately from the direction of the opinion. We can do two different transformations on this likert.centred variable in order to distinguish between these two different concepts. First, to compute an opinion.strength variable, we want to take the absolute value of the

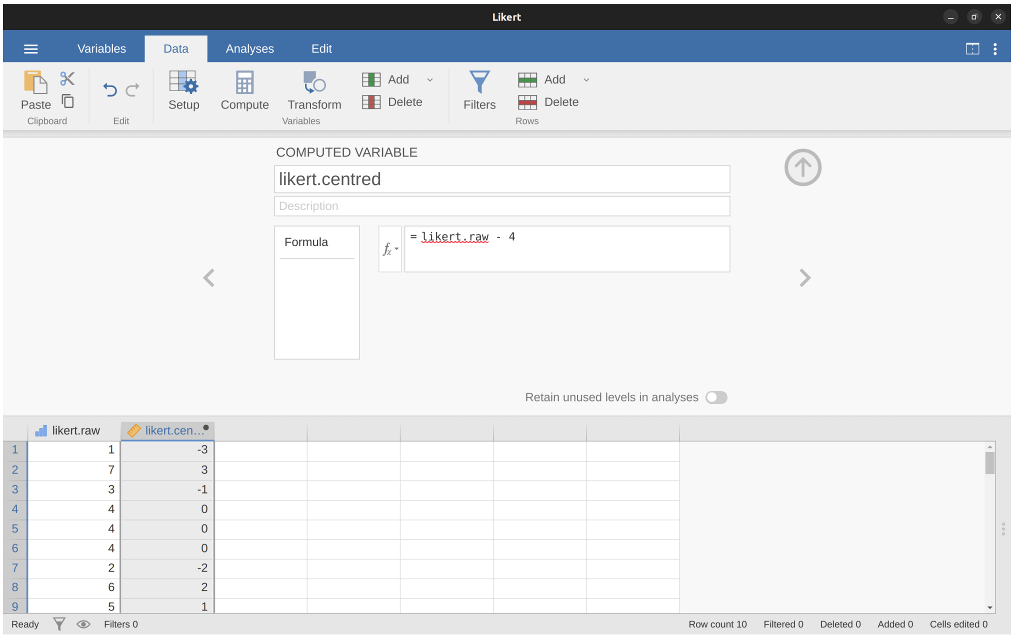


Figure 6.4: Creating a new computed variable in jamovi

centred data (using the 'ABS' function).[6] In jamovi, create another new variable using the 'Compute' button. Name the variable opinion.strength and this time click on the fx button next to the 'Formula' box. This shows the different 'Functions' and 'Variables' that you can add to the 'Formula' box, so double click on 'ABS' and then double click on "likert.centred' and you will see that the 'Formula' box is populated with 'ABS(likert.centred)' and a new variable has been created in the spreadsheet view, as in Figure 6.5.

Second, to compute a variable that contains only the direction of the opinion and ignores the strength, we want to calculate the 'sign' of the variable. In jamovi we can use the IF function to do this. Create another new variable using the 'Compute' button, name this one opinion.sign, and then type the following into the function box:

IF(likert.centred == 0, 0, likert.centred / opinion.strength) When done, you'll see that all negative numbers from the likert.centred variable are converted to -1, all positive numbers are converted to 1 and zero stays as 0, like so:

-1 1 -1 0 0 0 -1 1 1 1

Let's break down what this 'IF' command is doing. In jamovi there are three parts to an 'IF' statement, written as 'IF(expression, value, else)'. The first part, 'expression' can be a logical or mathematical statement. In our example, we have specified 'likert.centred == 0', which is TRUE for values where likert.centred is zero. The next part, 'value', is the new value where the expression in part one is TRUE. In our example, we have

[6] The absolute value of a number is its distance from zero, regardless of whether its sign is negative or positive.

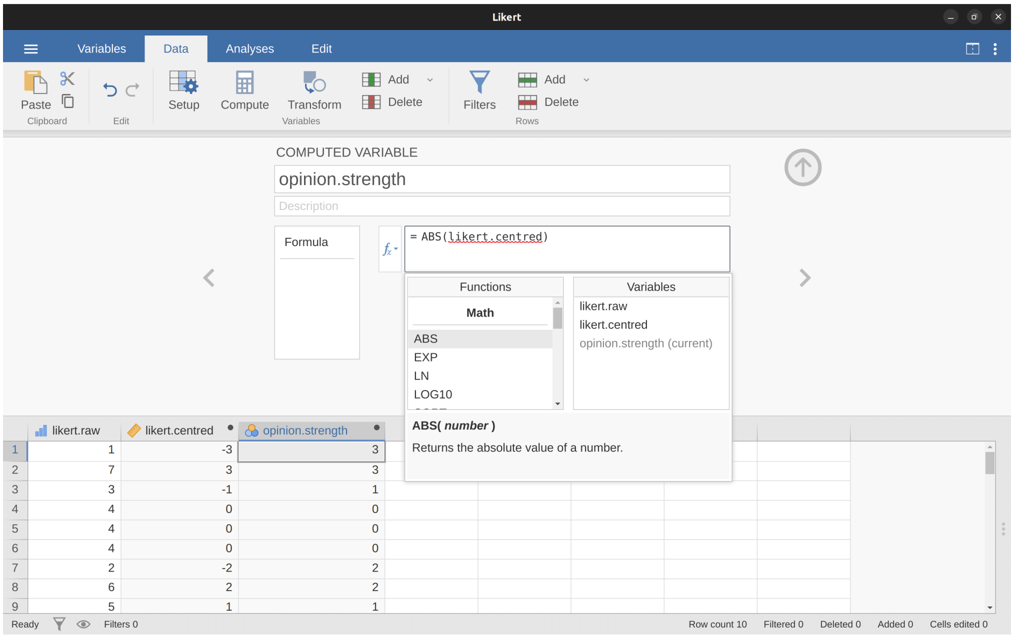


Figure 6.5: Using the f_x button to select functions and variables

said that for all those values where likert.centred is zero, keep them zero. In the next part, 'else', we can enter another logical or mathematical statement to be used if part one evaluates to FALSE, i.e. where likert.centred is not zero. In our example we have divided likert.centred by opinion.strength to give '-1' or '+1' depending of the sign of the original value in likert.centred.[7]

And we're done. We now have three shiny new variables, all of which are useful transformations of the original likert.raw data.

6.3.2 Collapsing a variable into a smaller number of discrete levels or categories

One pragmatic task that comes up quite often is the problem of collapsing a variable into a smaller number of discrete levels or categories. For instance, suppose I'm interested in looking at the age distribution of people at a social gathering:

60,58,24,26,34,42,31,30,33,2,9

In some situations it can be quite helpful to group these into a smallish number of categories. For example, we could group the data into three broad categories: young (0-20), adult (21-40) and older (41-60). This is a quite coarse-grained classification,

[7]The reason we have to use the 'IF' command and keep zero as zero is that you cannot just use likert.centred / opinion.strength to calculate the sign of likert.centred, because mathematically dividing zero by zero does not work. Try it and see.

and the labels that I've attached only make sense in the context of this data set (e.g., viewed more generally, a 42 year old wouldn't consider themselves as "older"). We can slice this variable up quite easily using the jamovi 'IF' function that we have already used. This time we have to specify nested 'IF' statements, meaning simply that IF the first logical expression is TRUE, insert a first value, but IF a second logical expression is TRUE, insert a second value, but IF a third logical expression is TRUE, then insert a third value. This can be written as:

> IF(Age >= 0 and Age <= 20, 1, IF(Age >= 21 and Age <= 40, 2, IF(Age >= 41 and Age <= 60, 3)))

Note that there are three left parentheses used during the nesting, so the whole statement has to end with three right parentheses otherwise you will get an error message. The jamovi screenshot for this data manipulation, along with an accompanying frequency table, is shown in Figure 6.6.

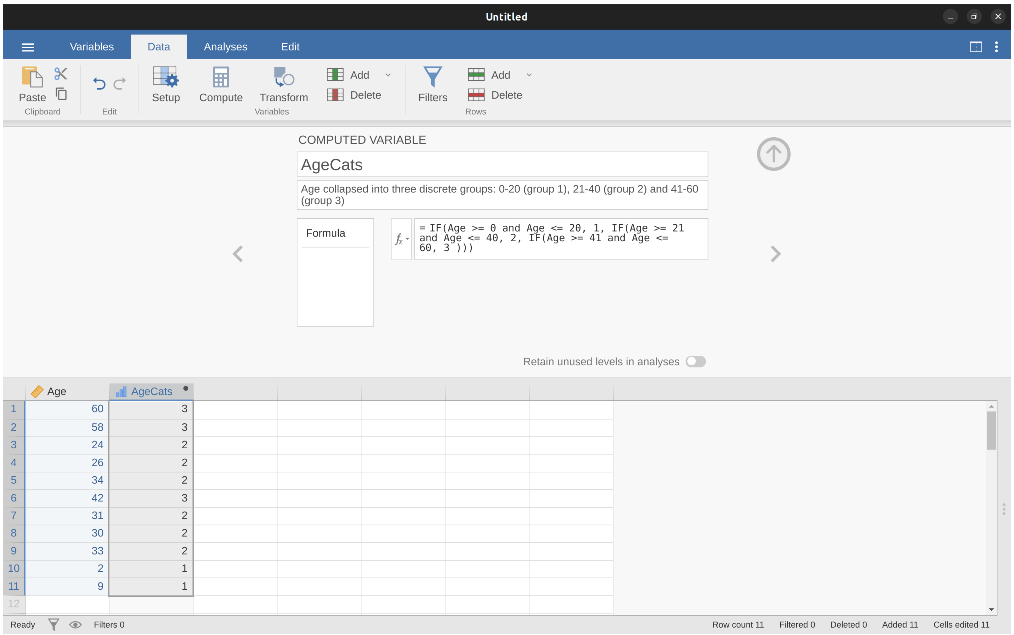


Figure 6.6: Collapsing a variable into a smaller number of discrete levels using the jamovi 'IF' function

It's important to take the time to figure out whether or not the resulting categories make any sense at all in terms of your research project. If they don't make any sense to you as meaningful categories, then any data analysis that uses those categories is likely to be just as meaningless. More generally, in practice I've noticed that people have a very strong desire to carve their (continuous and messy) data into a few (discrete and simple) categories, and then run analyses using the categorised data instead of the original data.[8] I wouldn't go so far as to say that this is an inherently bad idea, but it

[8]If you've read further into the book, and are re-reading this section, then a good example of this

does have some fairly serious drawbacks at times, so I would advise some caution if you are thinking about doing it.

6.3.3 Creating a transformation that can be applied to multiple variables

Sometimes you want to apply the same transformation to more than one variable, for example when you have multiple questionnaire items that all need to be recalculated or recoded in the same way. And one of the neat features in jamovi is that you can create a transformation, using the 'Data' - 'Transform' button, that can then be saved and applied to multiple variables. Let's go back to the first example above, using the data file *likert.omv* that contains a single variable with raw Likert-scale responses for 10 people. To create a transformation that you can save and then apply across multiple variables (assuming you had more variables like this in your data file), first in the spreadsheet editor select (i.e., click) the variable you want to use to initially create the transformation. In our example this is likert.raw. Next click the 'Transform' button in the jamovi 'Data' ribbon, and you'll see something like Figure 6.7.

Give your new variable a name, let's call it 'opinion.strength', and then click on the 'using transform' selection box and select 'Create New Transform...'. This is where you will create, and name, the transformation that can be re-applied to as many variables as you like. The transformation is automatically named for us as 'Transform 1' (imaginative, huh. You can change this if you like). Then type the expression "ABS($source - 4)" into the function text box, as in Figure 6.8, press Enter or Return on your keyboard and, hey presto, you have created a new transformation and applied it to the likert.raw variable! Good, eh. Note that instead of using the variable label in the expression, we have instead used '$source'. This is so that we can then use the same transformation with as many different variables as we like - jamovi requires you to use '$source' to refer to the source variable you are transforming. Your transformation has also been saved and can be re-used any time you like (providing you save the dataset as an '.omv' file, otherwise you'll lose it!).

You can also create a transformation with the second example we looked at, the age distribution of people at a social gathering. Go on, you know you want to! Remember that we collapsed this variable into three groups: younger, adult and older. This time we will achieve the same thing, but using the jamovi 'Transform' - 'Add condition' button. With this data set (go back to it or create it again if you didn't save it) set up a new variable transformation. Call the transformed variable 'AgeCats' and the transformation you will create 'Agegroupings'. Then click on the big "+" sign next to the function box. This is the 'Add condition' button and I've stuck a big red arrow onto Figure 6.9 so you can see exactly where this is. Re-create the transformation shown in Figure 6.9 and when you are done, you will see the new values appear in the spreadsheet window. What's more, the age groupings transformation has been saved and can be re-applied any time you like. Ok, so I know that it's unlikely you will have more than

would be someone choosing to do an ANOVA using AgeCats as the grouping variable, instead of running a regression using Age as a predictor. There are sometimes good reasons for doing this. For instance, if the relationship between Age and your outcome variable is highly non-linear and you aren't comfortable with trying to run non-linear regression! However, unless you really do have a good rationale for doing this, it's best not to. It tends to introduce all sorts of other problems (e.g., the data will probably violate the normality assumption) and you can lose a lot of statistical power.

one ‘Age’ variable, but you get the idea now of how to set up transformations in jamovi, so you can follow this idea with other sorts of variables. A typical scenario for this is when you have a questionnaire scale with, say, 20 items (variables) and each item was originally scored from 1 to 6 but, for some reason or quirk of the data you decide to recode all the items as 1 to 3. You can easily do this in jamovi by creating and then re-applying your transformation for each variable that you want to recode.

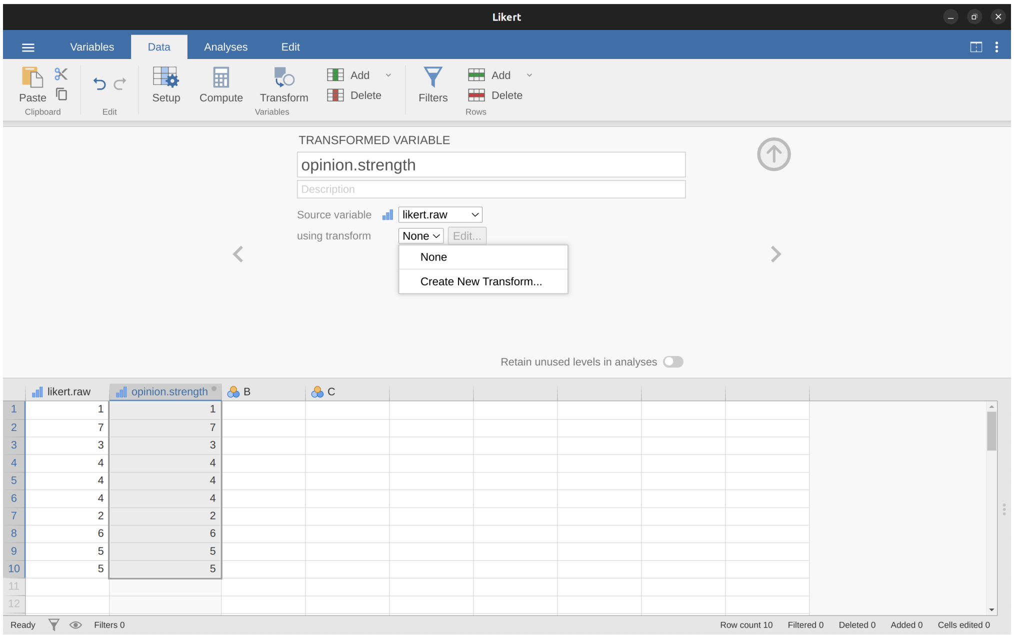


Figure 6.7: Creating a new variable transformation using the jamovi ‘Transform’ command

6.4 A few more mathematical functions and operations

In the section on Transforming and recoding a variable I discussed the ideas behind variable transformations and showed that a lot of the transformations that you might want to apply to your data are based on fairly simple mathematical functions and operations. In this section I want to return to that discussion and mention several other mathematical functions and arithmetic operations that are actually quite useful for a lot of real world data analysis. Table 6.5 gives a brief overview of the various mathematical functions I want to talk about here, or later.[9] Obviously this doesn't even come close to cataloguing the range of possibilities available, but it does cover a range of functions that are used regularly in data analysis and that are available in jamovi.

[9] We'll leave the box-cox function until later on.

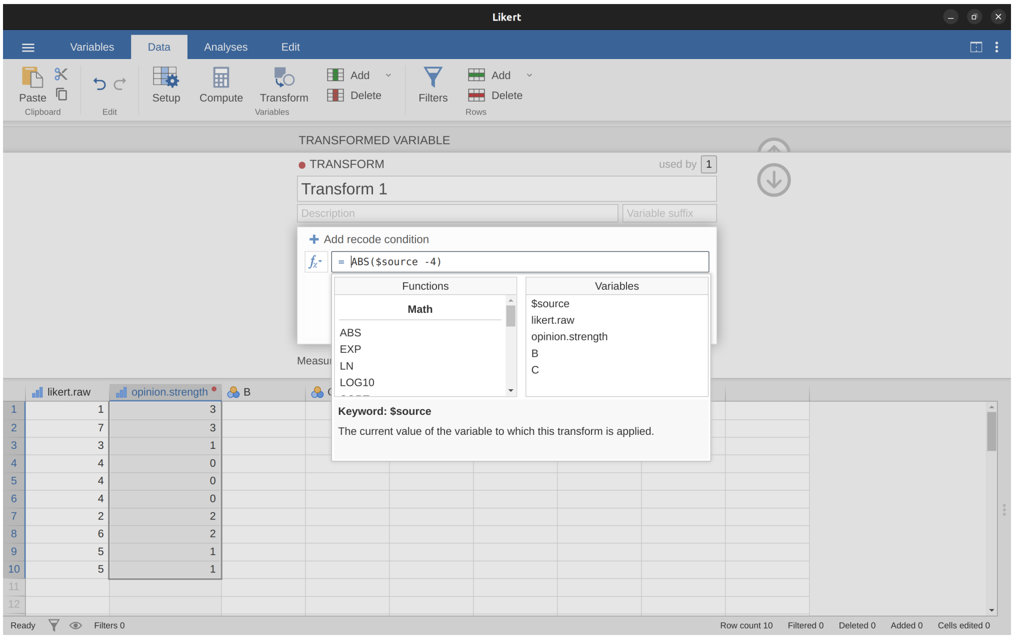


Figure 6.8: Specifying a transformation in jamovi, to be saved as the imaginatively named 'Transform 1'

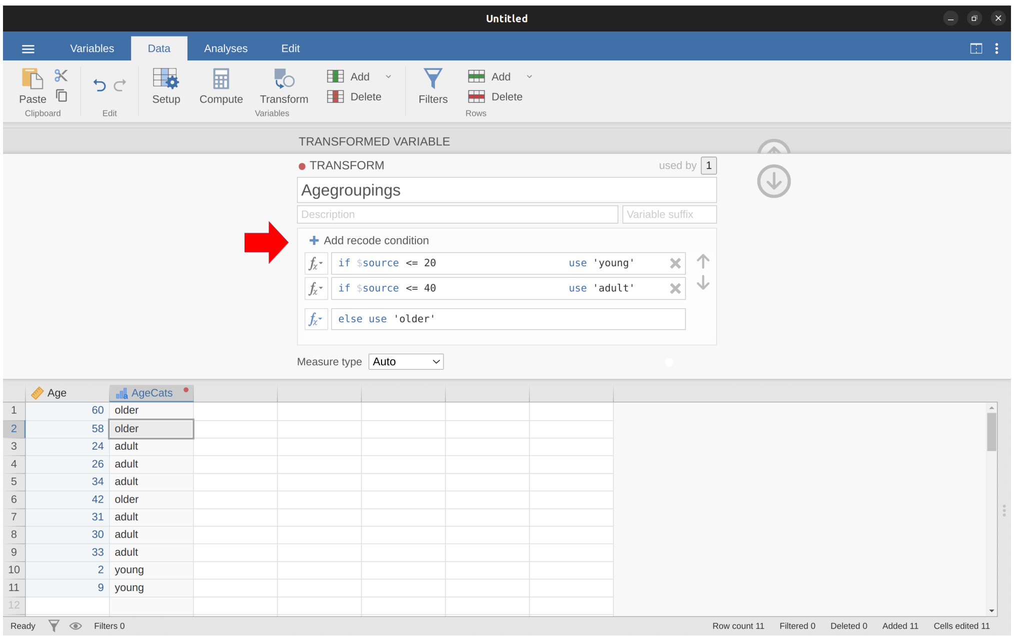


Figure 6.9: jamovi transformation into three age categories, using the 'Add condition' button

Table 6.5: Some mathematical operators

	function	example input	(answer)
square root	SQRT(x)	SQRT(25)	5
absolute value	ABS(x)	ABS(-23)	23
logarithm (base 10)	LOG10(x)	LOG10(1000)	3
logarithm (base e)	LN(x)	LN(1000)	6.91
exponentiation	EXP(x)	EXP(6.908)	1e+03
box-cox	BOXCOX(x, lamda)	BOXCOX(6.908, 3)	110

6.4.1 Logarithms and exponentials

As I've mentioned earlier, jamovi has a useful range of mathematical functions built into it and there really wouldn't be much point in trying to describe or even list all of them. For the most part, I've focused only on those functions that are strictly necessary for this book. However I do want to make an exception for logarithms and exponentials. Although they aren't needed anywhere else in this book, they are *everywhere* in statistics more broadly. And not only that, there are a *lot of* situations in which it is convenient to analyse the logarithm of a variable (i.e., to take a "log-transform" of the variable). I suspect that many (maybe most) readers of this book will have encountered logarithms and exponentials before, but from past experience I know that there's a substantial proportion of students who take a social science statistics class who haven't touched logarithms since high school, and would appreciate a bit of a refresher.

In order to understand logarithms and exponentials, the easiest thing to do is to actually calculate them and see how they relate to other simple calculations. There are three jamovi functions in particular that I want to talk about, namely LN(), LOG10() and EXP(). To start with, let's consider LOG10(), which is known as the "logarithm in base 10". The trick to understanding a **logarithm** is to understand that it's basically the "opposite" of taking a power. Specifically, the logarithm in base 10 is closely related to the powers of 10. So let's start by noting that 10-cubed is 1000. Mathematically, we would write this:

$$10^3 = 1000$$

The trick to understanding a logarithm is to recognise that the statement that "10 to the power of 3 is equal to 1000" is equivalent to the statement that "the logarithm (in base 10) of 1000 is equal to 3". Mathematically, we write this as follows,

$$\log_{10}(1000) = 3$$

Okay, since the LOG10() function is related to the powers of 10, you might expect that there are other logarithms (in bases other than 10) that are related to other powers too.

And of course that's true: there's not really anything mathematically special about the number 10. You and I happen to find it useful because decimal numbers are built around the number 10, but the big bad world of mathematics scoffs at our decimal numbers. Sadly, the universe doesn't actually care how we write down numbers. Anyway, the consequence of this cosmic indifference is that there's nothing particularly special about calculating logarithms in base 10. You could, for instance, calculate your logarithms in base 2. Alternatively, a third type of logarithm, and one we see a lot more of in statistics than either base 10 or base 2, is called the **natural logarithm**, and corresponds to the logarithm in base e. Since you might one day run into it, I'd better explain what e is. The number e, known as **Euler's number**, is one of those annoying "irrational" numbers whose decimal expansion is infinitely long, and is considered one of the most important numbers in mathematics. The first few digits of e are:

$$e \approx 2.718282$$

There are quite a few situations in statistics that require us to calculate powers of e, though none of them appear in this book. Raising e to the power x is called the **exponential** of x, and so it's very common to see e^x written as $\exp(x)$. And so it's no surprise that jamovi has a function that calculates exponentials, called EXP(). Because the number e crops up so often in statistics, the natural logarithm (i.e., logarithm in base e) also tends to turn up. Mathematicians often write it as $\log_e(x)$ or $\ln(x)$. In fact, jamovi works the same way: the LN() function corresponds to the natural logarithm.

And with that, I think we've had quite enough exponentials and logarithms for this book!

6.5 Extracting a subset of the data

One very important kind of data handling is being able to extract a particular subset of the data. For instance, you might be interested only in analysing the data from one experimental condition, or you may want to look closely at the data from people over 50 years in age. To do this, the first step is getting jamovi to filter the subset of the data corresponding to the observations that you're interested in.

This section returns to the *nightgarden.csv* data set. If you're reading this whole chapter in one sitting, then you should already have this data set loaded into a jamovi window. For this section, let's focus on the two variables speaker and utterance (see Tabulating and cross-tabulating data if you've forgotten what those variables look like). Suppose that what I want to do is pull out only those utterances that were made by Makka-Pakka. To that end, we need to specify a filter in jamovi. First open up a filter window by clicking on 'Filters' on the main jamovi 'Data' toolbar. Then, in the 'Filter 1' text box, next to the '=' sign, type the following, including the single quote marks:

speaker == 'makka-pakka'

When you have done this, you will see that a new column has been added to the spreadsheet window (see Figure 6.10), labelled 'Filter 1', with the cases where speaker is not 'makka-pakka' greyed-out (i.e., filtered out) and, conversely, the cases where

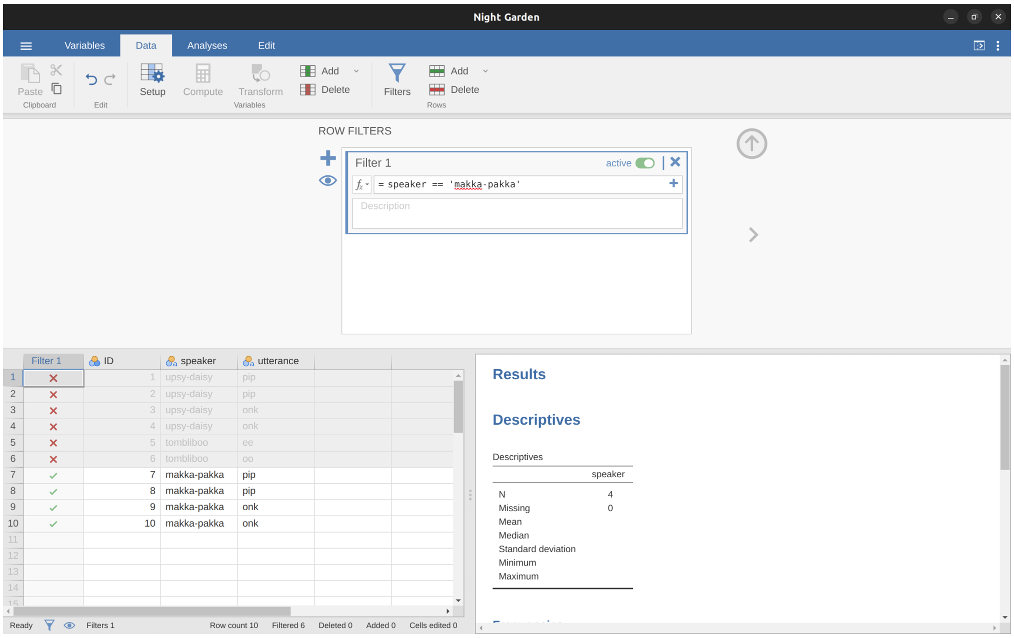


Figure 6.10: Creating a subset of the nightgarden data using the jamovi 'Filters' option

speaker is 'makka-pakka' have a green check mark indicating they are filtered in. You can test this by running 'Exploration' - 'Descriptives' - 'Frequency tables' for the speaker variable and seeing what that shows. Go on, try it!

Following on from this simple example, you can also build up more complex filters using logical expressions in jamovi. For instance, suppose I wanted to keep only those cases when the utterance is either "pip" or "oo". In this case in the 'Filter 1' text box, next to the '=' sign, you would type the following:

utterance == 'pip' or utterance == 'oo'

6.6 Summary

Obviously, there's no real coherence to this chapter. It's just a grab bag of topics and tricks that can be handy to know about, so the best wrap up I can give here is just to repeat this list:

- Tabulating and cross-tabulating data.
- Logical expressions in jamovi.
- Transforming and recoding a variable.
- A few more mathematical functions and operations.
- Extracting a subset of the data.

Part IV

Statistical theory

Prelude

Part IV of the book is by far the most theoretical, focusing as it does on the theory of statistical inference. Over the next three chapters my goal is to give you an introduction to probability theory, sampling and estimation in Chapter 8 and statistical hypothesis testing in **?@sec-Hypothesis-testing**. Before we get started though, I want to say something about the big picture. Statistical inference is primarily about learning from data. The goal is no longer merely to describe our data but to use the data to draw conclusions about the world. To motivate the discussion I want to spend a bit of time talking about a philosophical puzzle known as the riddle of induction, because it speaks to an issue that will pop up over and over again throughout the book: statistical inference relies on assumptions. This sounds like a bad thing. In everyday life people say things like "you should never make assumptions", and psychology classes often talk about assumptions and biases as bad things that we should try to avoid. From bitter personal experience I have learned never to say such things around philosophers!

On the limits of logical reasoning

> *"The whole art of war consists in getting at what is on the other side of the hill, or, in other words, in learning what we do not know from what we do."*
> - Arthur Wellesley, 1st Duke of Wellington

I am told that quote above came about as a consequence of a carriage ride across the countryside.[10] He and his companion, J. W. Croker, were playing a guessing game, each trying to predict what would be on the other side of each hill. In every case it turned out that Wellesley was right and Croker was wrong. Many years later when Wellesley was asked about the game he explained that "the whole art of war consists in getting at what is on the other side of the hill". Indeed, war is not special in this respect. All of life is a guessing game of one form or another, and getting by on a day-to-day basis requires us to make good guesses. So let's play a guessing game of our own.

Suppose you and I are observing the Wellesley-Croker competition and after every three hills you and I have to predict who will win the next one, Wellesley or Croker. Let's say that W refers to a Wellesley victory and C refers to a Croker victory. After three hills, our data set looks like this:

WWW

[10] http://www.bartleby.com/344/400.html

Our conversation goes like this:

> you: Three in a row doesn't mean much. I suppose Wellesley might be better at this than Croker, but it might just be luck. Still, I'm a bit of a gambler. I'll bet on Wellesley.
>
> me: I agree that three in a row isn't informative and I see no reason to prefer Wellesley's guesses over Croker's. I can't justify betting at this stage. Sorry. No bet for me.

Your gamble paid off: three more hills go by and Wellesley wins all three. Going into the next round of our game the score is 1-0 in favour of you and our data set looks like this: WWW WWW. I've organised the data into blocks of three so that you can see which batch corresponds to the observations that we had available at each step in our little side game. After seeing this new batch, our conversation continues:

> you: Six wins in a row for Duke Wellesley. This is starting to feel a bit suspicious. I'm still not certain, but I reckon that he's going to win the next one too.
>
> me: I guess I don't see that. Sure, I agree that Wellesley has won six in a row, but I don't see any logical reason why that means he'll win the seventh one. No bet. you: Do you really think so? Fair enough, but my bet worked out last time and I'm okay with my choice.

For a second time you were right, and for a second time I was wrong. Wellesley wins the next three hills, extending his winning record against Croker to 9-0. The data set available to us is now this: WWW WWW WWW. And our conversation goes like this:

> you: Okay, this is pretty obvious. Wellesley is way better at this game. We both agree he's going to win the next hill, right?
>
> me: Is there really any logical evidence for that? Before we started this game, there were lots of possibilities for the first 10 outcomes, and I had no idea which one to expect. WWW WWW WWW W was one possibility, but so was WCC CWC WWC C and WWW WWW WWW C or even CCC CCC CCC C. Because I had no idea what would happen so I'd have said they were all equally likely. I assume you would have too, right? I mean, that's what it means to say you have "no idea", isn't it?
>
> you: I suppose so.
>
> me: Well then, the observations we've made logically rule out all possibilities except two: WWW WWW WWW C or WWW WWW WWW W. Both of these are perfectly consistent with the evidence we've encountered so far, aren't they?

you: Yes, of course they are. Where are you going with this?

me: So what's changed then? At the start of our game, you'd have agreed with me that these are equally plausible and none of the evidence that we've encountered has discriminated between these two possibilities. Therefore, both of these possibilities remain equally plausible and I see no logical reason to prefer one over the other. So yes, while I agree with you that Wellesley's run of 9 wins in a row is remarkable, I can't think of a good reason to think he'll win the 10th hill. No bet.

you: I see your point, but I'm still willing to chance it. I'm betting on Wellesley.

Wellesley's winning streak continues for the next three hills. The score in the Wellesley-Croker game is now 12-0, and the score in our game is now 3-0. As we approach the fourth round of our game, our data set is this: WWW WWW WWW WWW. And the conversation continues:

you: Oh yeah! Three more wins for Wellesley and another victory for me. Admit it, I was right about him! I guess we're both betting on Wellesley this time around, right?

me: I don't know what to think. I feel like we're in the same situation we were in last round, and nothing much has changed. There are only two legitimate possibilities for a sequence of 13 hills that haven't already been ruled out, WWW WWW WWW WWW C and WWW WWW WWW WWW W. It's just like I said last time. If all possible outcomes were equally sensible before the game started, shouldn't these two be equally sensible now given that our observations don't rule out either one? I agree that it feels like Wellesley is on an amazing winning streak, but where's the logical evidence that the streak will continue?

you: I think you're being unreasonable. Why not take a look at our scorecard, if you need evidence? You're the expert on statistics and you've been using this fancy logical analysis, but the fact is you're losing. I'm just relying on common sense and I'm winning. Maybe you should switch strategies.

me: Hmm, that is a good point and I don't want to lose the game, but I'm afraid I don't see any logical evidence that your strategy is better than mine. It seems to me that if there were someone else watching our game, what they'd have observed is a run of three wins to you. Their data would look like this: YYY. Logically, I don't see that this is any different to our first round of watching Wellesley and Croker. Three wins to you doesn't seem like a lot of evidence, and I see no reason to think that your strategy is working out any better than mine. If I didn't think that WWW was good evidence then for Wellesley being better than Croker at their game, surely I have no reason now to think that YYY is good evidence that you're better at ours?

you: Okay, now I think you're being a jerk.

me: I don't see the logical evidence for that.

Learning without making assumptions is a myth

There are lots of different ways in which we could dissect this dialogue, but since this is a statistics book pitched at psychologists and not an introduction to the philosophy and psychology of reasoning, I'll keep it brief. What I've described above is sometimes referred to as the riddle of induction. It seems entirely reasonable to think that a 12-0 winning record by Wellesley is pretty strong evidence that he will win the 13th game, but it is not easy to provide a proper logical justification for this belief. On the contrary, despite the obviousness of the answer, it's not actually possible to justify betting on Wellesley without relying on some assumption that you don't have any logical justification for.

The riddle of induction is most associated with the philosophical work of David Hume and more recently Nelson Goodman, but you can find examples of the problem popping up in fields as diverse as literature (Lewis Carroll) and machine learning (the "no free lunch" theorem). There really is something weird about trying to "learn what we do not know from what we do know". The critical point is that assumptions and biases are unavoidable if you want to learn anything about the world. There is no escape from this, and it is just as true for statistical inference as it is for human reasoning. In the dialogue I was taking aim at your perfectly sensible inferences as a human being, but the common sense reasoning that you relied on is no different to what a statistician would have done. Your "common sense" half of the dialog relied on an implicit assumption that there exists some difference in skill between Wellesley and Croker, and what you were doing was trying to work out what that difference in skill level would be. My "logical analysis" rejects that assumption entirely. All I was willing to accept is that there are sequences of wins and losses and that I did not know which sequences would be observed. Throughout the dialogue I kept insisting that all logically possible data sets were equally plausible at the start of the Wellesely-Croker game, and the only way in which I ever revised my beliefs was to eliminate those possibilities that were factually inconsistent with the observations.

That sounds perfectly sensible on its own terms. In fact, it even sounds like the hallmark of good deductive reasoning. Like Sherlock Holmes, my approach was to rule out that which is impossible in the hope that what would be left is the truth. Yet as we saw, ruling out the impossible never led me to make a prediction. On its own terms everything I said in my half of the dialogue was entirely correct. An inability to make any predictions is the logical consequence of making "no assumptions". In the end I lost our game because you did make some assumptions and those assumptions turned out to be right. Skill is a real thing, and because you believed in the existence of skill you were able to learn that Wellesley had more of it than Croker. Had you relied on a less sensible assumption to drive your learning you might not have won the game.

Ultimately there are two things you should take away from this. First, as I've said, you cannot avoid making assumptions if you want to learn anything from your data. But second, once you realise that assumptions are necessary it becomes important to make sure you make the right ones! A data analysis that relies on few assumptions is not necessarily better than one that makes many assumptions, it all depends on whether those assumptions are good ones for your data. As we go through the rest of this book I'll often point out the assumptions that underpin a particular statistical technique, and how you can check whether those assumptions are sensible.

Chapter 7

Introduction to probability

> "*[God] has afforded us only the twilight ... of Probability.*"
> – John Locke

Up to this point in the book we've discussed some of the key ideas in experimental design, and we've talked a little about how you can summarise a data set. To a lot of people this is all there is to statistics: collecting all the numbers, calculating averages, drawing pictures, and putting them all in a report somewhere. Kind of like stamp collecting but with numbers. However, statistics covers much more than that. In fact, descriptive statistics is one of the smallest parts of statistics and one of the least powerful. The bigger and more useful part of statistics is that it provides information that lets you make inferences about data.

Once you start thinking about statistics in these terms, that statistics is there to help us draw inferences from data, you start seeing examples of it everywhere. For instance, here's a tiny extract from a newspaper article in the *Sydney Morning Herald* (30 Oct 2010):

> "I have a tough job," the Premier said in response to a poll which found her government is now the most unpopular Labor administration in polling history, with a primary vote of just 23 per cent.

This kind of remark is entirely unremarkable in the papers or in everyday life, but let's have a think about what it entails. A polling company has conducted a survey, usually a pretty big one because they can afford it. I'm too lazy to track down the original survey so let's just imagine that they called 1000 New South Wales (NSW) voters at random, and 230 (23%) of those claimed that they intended to vote for the Australian Labor Party (ALP). For the 2010 Federal election the Australian Electoral Commission reported 4,610,795 enrolled voters in NSW, so the opinions of the remaining 4,609,795 voters (about 99.98% of voters) remain unknown to us. Even assuming that no-one lied to the polling company the only thing we can say with 100% confidence is that the true ALP primary vote is somewhere between 230/4,610,795 (about 0.005%) and 4,610,025/4,610,795 (about 99.83%). So, on what basis is it legitimate for the polling company, the newspaper, and the readership to conclude that the ALP primary vote is only about 23%?

The answer to the question is pretty obvious. If I call 1000 people at random, and 230 of them say they intend to vote for the ALP, then it seems very unlikely that these are the only 230 people out of the entire voting public who actually intend to vote ALP. In other words, we assume that the data collected by the polling company is pretty representative of the population at large. But how representative? Would we be surprised to discover that the true ALP primary vote is actually 24%? 29%? 37%? At this point everyday intuition starts to break down a bit. No-one would be surprised by 24%, and everybody would be surprised by 37%, but it's a bit hard to say whether 29% is plausible. We need some more powerful tools than just looking at the numbers and guessing.

Inferential statistics provides the tools that we need to answer these sorts of questions, and since these kinds of questions lie at the heart of the scientific enterprise, they take up the lions share of every introductory course on statistics and research methods. However, the theory of statistical inference is built on top of **probability theory**. And it is to probability theory that we must now turn. This discussion of probability theory is basically background detail. There's not a lot of statistics per se in this chapter, and you don't need to understand this material in as much depth as the other chapters in this part of the book. Nevertheless, because probability theory does underpin so much of statistics, it's worth covering some of the basics.

7.1 How are probability and statistics different?

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 a six-sided dice twice, 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 "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:

$$P(head) = 0.5$$

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 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(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 $P(heads) \neq 0.5$. 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.

7.2 What does probability mean?

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. 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 of 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 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.

7.2.1 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 $\frac{N_H}{N}$ every time. Table 7.1 shows what I'd get (I did literally flip coins to produce this!).

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

Table 7.1: Coin flips and proportion of heads

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

proportion $\frac{N_H}{N}$ as N increases. Actually, I did it four times just to make sure it wasn't a fluke. The results are shown in Figure 7.1. 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. First, 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 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. First, 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

[1] This 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 redescribe that statement in terms of a sequence of potentially observable events, together with the relative frequencies of different outcomes that appear within that sequence.

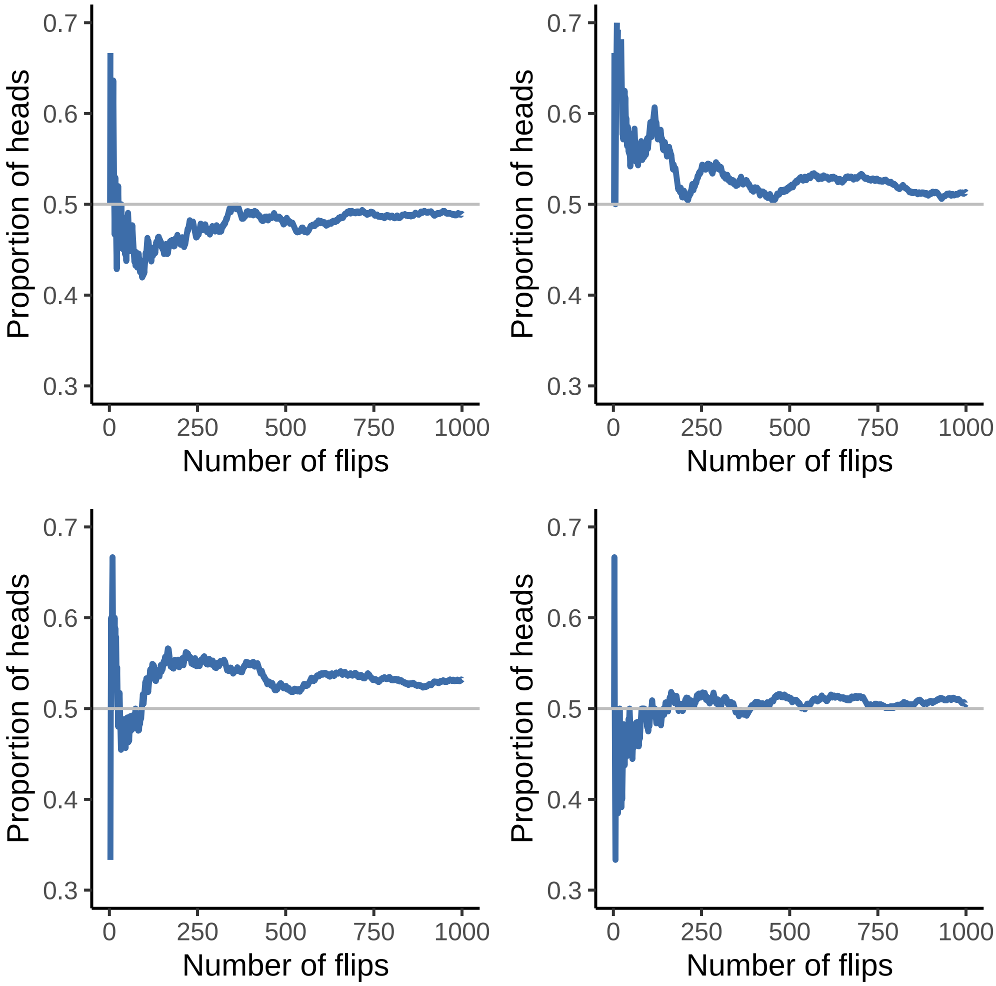


Figure 7.1: An illustration of how frequentist probability works. If you flip a fair coin over and over again the proportion of heads that you have seen eventually settles down and converges to the true probability of 0.5. Each panel shows four different simulated experiments. In each case we pretend 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 had extended the experiment for an infinite number of coin flips they would have

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 one 2 November 2048. There's no infinite sequence of events here, just a one-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 "There is a category of days for which I predict a 60% chance of rain, and 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 counter-intuitive to think of it this way, but you do see frequentists do this sometimes. And it will come up later in this book (e.g. in Section 8.5).

7.2.2 The Bayesian view

The Bayesian view of probability is often called the subjectivist view, and although it has been a minority view among statisticians it has been steadily gaining traction for the last several decades. There are many flavours of Bayesianism, making it 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. If someone offers me a bet that if it rains tomorrow then I win \$5, but if it doesn't rain 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. So 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. Whilst the Bayesian approach requires 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 that we want to assign probabilities to), the Bayesian view is sometimes thought to be too broad (allows too many differences between observers).

7.2.3 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 opt for? 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 difference then 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, p. 311). 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, p. 114). The history of statistics, as you might gather, is not devoid of entertainment.

In any case, whilst 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. Besides, I won't completely ignore the Bayesian perspective. Every now and then I'll add some commentary from a Bayesian point of view, and I'll revisit the topic in more depth in **?@sec-Bayesian-statistics**.

7.3 Basic probability theory

Ideological arguments between Bayesians and frequentists notwithstanding, it turns out that people mostly agree on the rules that probabilities should obey. There are lots of different ways of arriving at these rules. The most commonly used approach is based on the work of Andrey Kolmogorov, one of the great Soviet mathematicians of the 20th century. I won't go into a lot of detail, but I'll try to give you a bit of a sense of how it works. And in order to do so I'm going to have to talk about my trousers.

7.3.1 Introducing probability distributions

One of the disturbing truths about my life is that I only own five pairs of trousers. Three pairs of jeans, the bottom half of a suit, and a pair of tracksuit pants. Even sadder, I've given them names: I call them X_1, X_2, X_3, X_4 and X_5. I really have, that's why they call me Mister Imaginative. Now, on any given day, I pick out exactly one pair of trousers to wear. Not even I'm so stupid as to try to wear two pairs of trousers, and thanks to years of training I never go outside without wearing trousers anymore. If I were to describe this situation using the language of probability theory, I would refer to each pair of trousers (i.e., each X) as an elementary event. The key characteristic of **elementary events** is that every time we make an observation (e.g., every time I put on a pair of trousers) then the outcome will be one and only one of these events. Like I said, these days I always wear exactly one pair of trousers so my trousers satisfy this constraint. Similarly, the set of all possible events is called a **sample space**. Granted, some people would call it a "wardrobe", but that's because they're refusing to think about my trousers in probabilistic terms. Sad.

Okay, now that we have a sample space (a wardrobe), which is built from lots of possible elementary events (trousers), what we want to do is assign a **probability** of one of these elementary events. For an event X, the probability of that event $P(X)$ is a number that lies between 0 and 1. The bigger the value of $P(X)$, the more likely the event is to occur. So, for example, if $P(X) = 0$ it means the event X is impossible (i.e., I never wear those trousers). On the other hand, if $P(X) = 1$ it means that event X is certain to occur (i.e., I always wear those trousers). For probability values in the middle it means that I sometimes wear those trousers. For instance, if $P(X) = 0.5$ it means that I wear those trousers half of the time.

At this point, we're almost done. The last thing we need to recognise is that "something always happens". Every time I put on trousers, I really do end up wearing trousers (crazy, right?). What this somewhat trite statement means, in probabilistic terms, is that the probabilities of the elementary events need to add up to 1. This is known as the **law of total probability**, not that any of us really care. More importantly, if these requirements are satisfied then what we have is a **probability distribution**. For example, Table 7.2 shows an example of a probability distribution.

Each of the events has a probability that lies between 0 and 1, and if we add up the probability of all events they sum to 1. Awesome. We can even draw a nice bar graph (see Section 5.3) to visualise this distribution, as shown in Figure 7.2. And, at this point, we've all achieved something. You've learned what a probability distribution is, and I've finally managed to find a way to create a graph that focuses entirely on my trousers. Everyone wins! The only other thing that I need to point out is that probability theory

Table 7.2: A probability distribution for trouser wearing

Which trousers?	**Label**	**Probability**
Blue jeans	X_1	$P(X_1) = .5$
Grey jeans	X_2	$P(X_2) = .3$
Black jeans	X_3	$P(X_3) = .1$
Black suit	X_4	$P(X_4) = 0$
Blue tracksuit	X_5	$P(X_5) = .1$

allows you to talk about **non elementary events** as well as elementary ones. The easiest way to illustrate the concept is with an example. In the trousers example it's perfectly legitimate to refer to the probability that I wear jeans. In this scenario, the "Dani wears jeans" event is said to have happened as long as the elementary event that actually did occur is one of the appropriate ones. In this case "blue jeans", "black jeans" or "grey jeans". In mathematical terms we defined the "jeans" event E to correspond to the set of elementary events $(X1, X2, X3)$. If any of these elementary events occurs then E is also said to have occurred. Having decided to write down the definition of the E this way, it's pretty straightforward to state what the probability P(E) and, since the probabilities of blue, grey and black jeans respectively are .5, .3 and .1, the probability that I wear jeans is equal to .9. is: we just add everything up. In this particular case

$$P(E) = P(X_1) + P(X_2) + P(X_3)$$

At this point you might be thinking that this is all terribly obvious and simple and you'd be right. All we've really done is wrap some basic mathematics around a few common sense intuitions. However, from these simple beginnings it's possible to construct some extremely powerful mathematical tools. I'm definitely not going to go into the details in this book, but what I will do is list, in Table 7.3, some of the other rules that probabilities satisfy. These rules can be derived from the simple assumptions that I've outlined above, but since we don't actually use these rules for anything in this book I won't do so here.

Table 7.3: Some rules that probabilities satisfy

English	**Notation**	**Formula**
not A	$P(\neg A)$	$1 - P(A)$
A or B	$P(A \cup B)$	$P(A)+P(B)-P(A \cap B)$
A and B	$P(A \cap B)$	$P(A\|B)P(B)$

7.4 The binomial distribution

As you might imagine, probability distributions vary enormously and there's an enormous range of distributions out there. However, they aren't all equally important. In fact, the vast majority of the content in this book relies on one of five distributions: the binomial distribution, the normal distribution, the t distribution, the χ^2 ("chi-square")

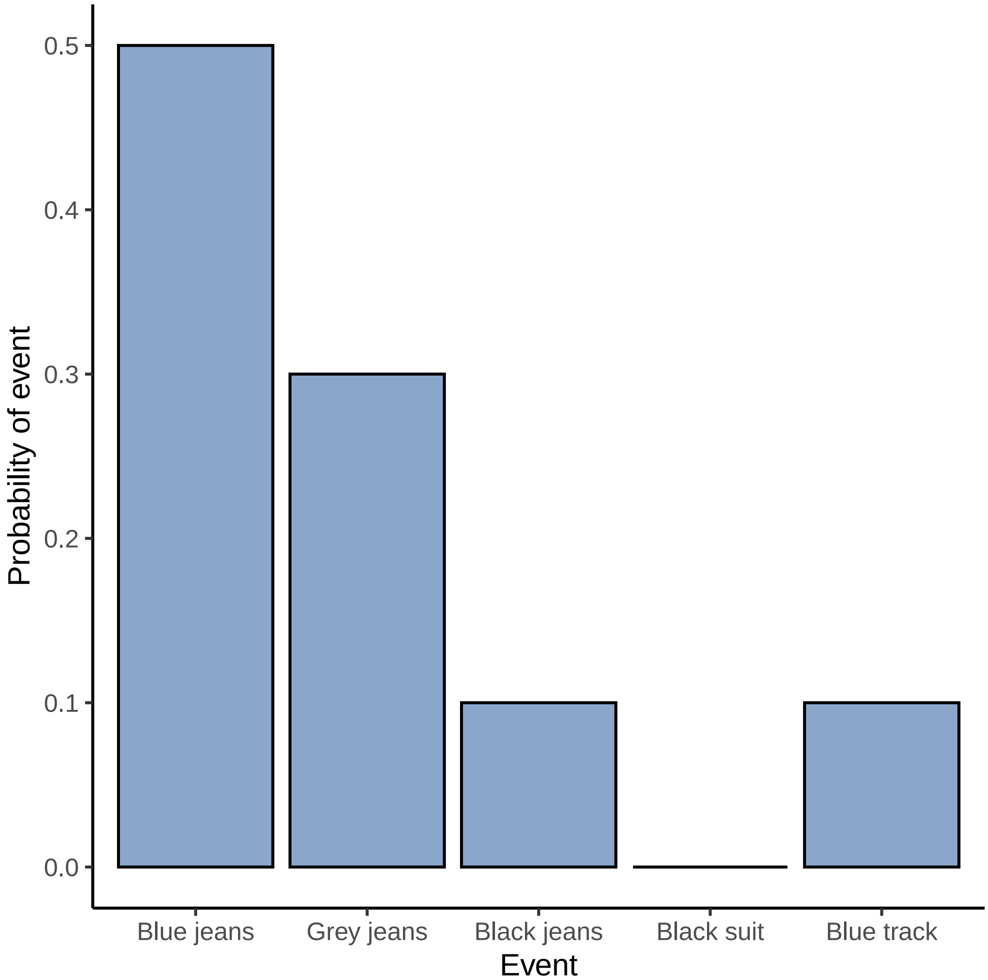


Figure 7.2: A visual depiction of the "trousers" probability distribution. There are five "elementary events", corresponding to the five pairs of trousers that I own. Each event has some probability of occurring – this probability is a number between 0 to 1. The sum of these probabilities is 1

distribution and the F distribution. Given this, what I'll do over the next few sections is provide a brief introduction to all five of these, paying special attention to the binomial and the normal. I'll start with the binomial distribution since it's the simplest of the five.

7.4.1 Introducing the binomial

The theory of probability originated in the attempt to describe how games of chance work, so it seems fitting that our discussion of the **binomial distribution** should involve a discussion of rolling dice and flipping coins. Let's imagine a simple "experiment". In my hot little hand I'm holding 20 identical six-sided dice. On one face of each die there's a picture of a skull, the other five faces are all blank. If I proceed to roll all 20 dice, what's the probability that I'll get exactly 4 skulls? Assuming that the dice are fair, we know that the chance of any one die coming up skulls is 1 in 6. To say this another way, the skull probability for a single die is approximately .167. This is enough information to answer our question, so let's have a look at how it's done.

As usual, we'll want to introduce some names and some notation. We'll let N denote the number of dice rolls in our experiment, which is often referred to as the **size parameter** of our binomial distribution. We'll also use θ to refer to the the probability that a single die comes up skulls, a quantity that is usually called the **success probability** of the binomial.[2] Finally, we'll use X to refer to the results of our experiment, namely the number of skulls I get when I roll the dice. Since the actual value of X is due to chance we refer to it as a **random variable**. In any case, now that we have all this terminology and notation we can use it to state the problem a little more precisely. The quantity that we want to calculate is the probability that $X = 4$ given that we know that $\theta = .167$ and $N = 20$. The general "form" of the thing I'm interested in calculating could be written as

$$P(X|\theta, N)$$

and we're interested in the special case where $X = 4, \theta = .167$ and $N = 20$.

[Additional technical detail[3]]

Yeah, yeah. I know what you're thinking: notation, notation, notation. Really, who cares? Very few readers of this book are here for the notation, so I should probably move on and talk about how to use the binomial distribution. I've included the formula

[2]Note that the term "success" is pretty arbitrary and doesn't actually imply that the outcome is something to be desired. If θ referred to the probability that any one passenger gets injured in a bus crash I'd still call it the success probability, but that doesn't mean I want people to get hurt in bus crashes!

[3]For those readers who know a little calculus, I'll give a slightly more precise explanation. In the same way that probabilities are non-negative numbers that must sum to 1, probability densities are non-negative numbers that must integrate to 1 (where the integral is taken across all possible values of X). To calculate the probability that X falls between a and b we calculate the definite integral of the density function over the corresponding range, $\int \_a^b p(x) dx$. If you don't remember or never learned calculus, don't worry about this. It's not needed for this book.

for the binomial distribution in a footnote,[4] since some readers may want to play with it themselves, but since most people probably don't care that much and because we don't need the formula in this book, I won't talk about it in any detail. Instead, I just want to show you what the binomial distribution looks like.

To that end, Figure 7.3 plots the binomial probabilities for all possible values of X for our dice rolling experiment, from $X = 0$ (no skulls) all the way up to $X = 20$ (all skulls). Note that this is basically a bar chart, and is no different to the "trousers probability" plot I drew in Figure 7.2. On the horizontal axis we have all the possible events, and on the vertical axis we can read off the probability of each of those events. So, the probability of rolling 4 skulls out of 20 is about 0.20 (the actual answer is 0.2022036, as we'll see in a moment). In other words, you'd expect that to happen about 20% of the times you repeated this experiment.

To give you a feel for how the binomial distribution changes when we alter the values of θ and N, let's suppose that instead of rolling dice I'm actually flipping coins. This time around, my experiment involves flipping a fair coin repeatedly and the outcome that I'm interested in is the number of heads that I observe. In this scenario, the success probability is now $\theta = \frac{1}{2}$. Suppose I were to flip the coin $N = 20$ times. In this example, I've changed the success probability but kept the size of the experiment the same. What does this do to our binomial distribution? Well, as Figure 7.4 shows, the main effect of this is to shift the whole distribution, as you'd expect. Okay, what if we flipped a coin $N = 100$ times? Well, in that case we get Figure 7.4 (b). The distribution stays roughly in the middle but there's a bit more variability in the possible outcomes.

7.5 The normal distribution

While the binomial distribution is conceptually the simplest distribution to understand, it's not the most important one. That particular honour goes to the normal distribution, also referred to as "the bell curve" or a "Gaussian distribution". A **normal distribution** is described using two parameters: the mean of the distribution µ and the standard deviation of the distribution σ. The notation that we sometimes use to say that a variable X is normally distributed is as follows:

$$X \sim Normal(\mu, \sigma)$$

[Additional technical detail[5]]

[4] In the equation for the binomial, $X!$ is the factorial function (i.e., multiply all whole numbers from 1 to X):

$$P(X|\theta, N) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{N-X}$$

If this equation doesn't make a lot of sense to you, don't worry too much about it.

[5] As was the case with the binomial distribution, I have included the formula for the normal distribution in this book, because I think it's important enough that everyone who learns statistics should at least look at it, but since this is an introductory text I don't want to focus on it, so I've tucked it away in this footnote:

$$p(X|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

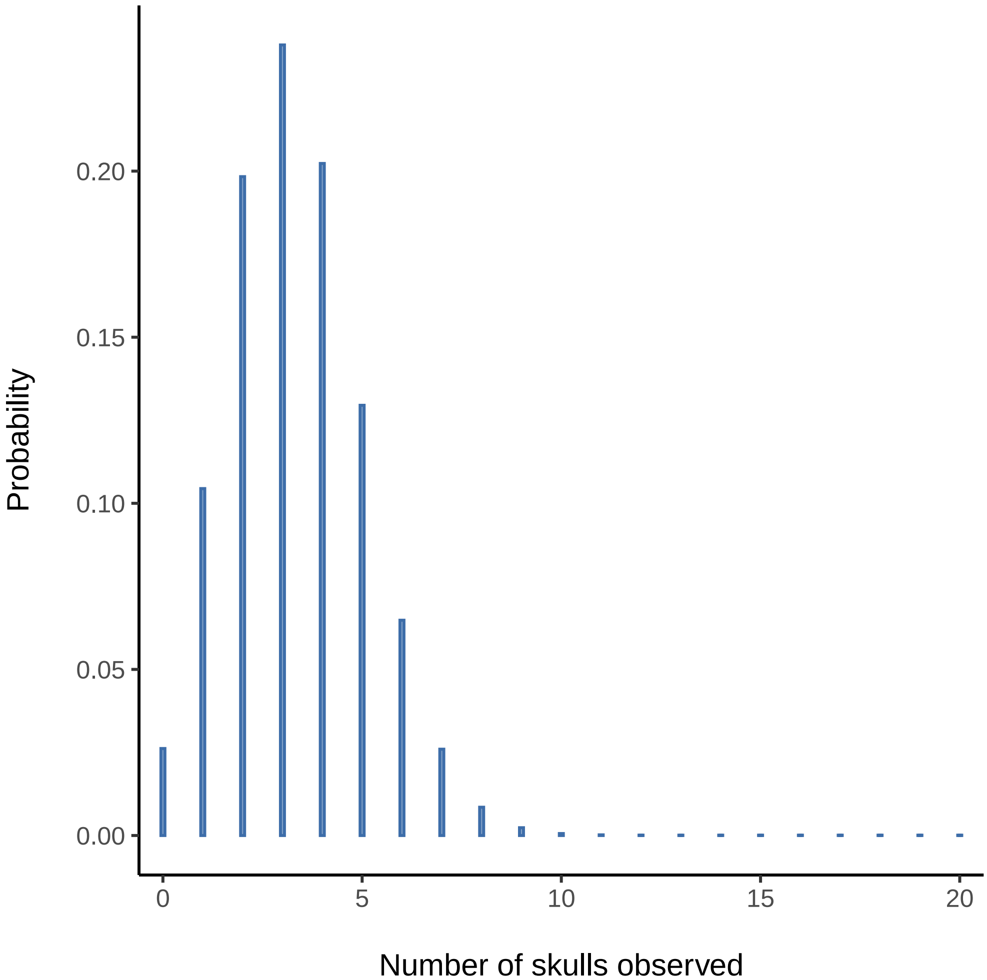


Figure 7.3: The binomial distribution with size parameter of $N = 20$ and an underlying success probability of $\theta = \frac{1}{6}$. Each vertical bar depicts the probability of one specific outcome (i.e., one possible value of X). Because this is a probability distribution, each of the probabilities must be a number between 0 and 1, and the heights of the bars must sum to 1 as well

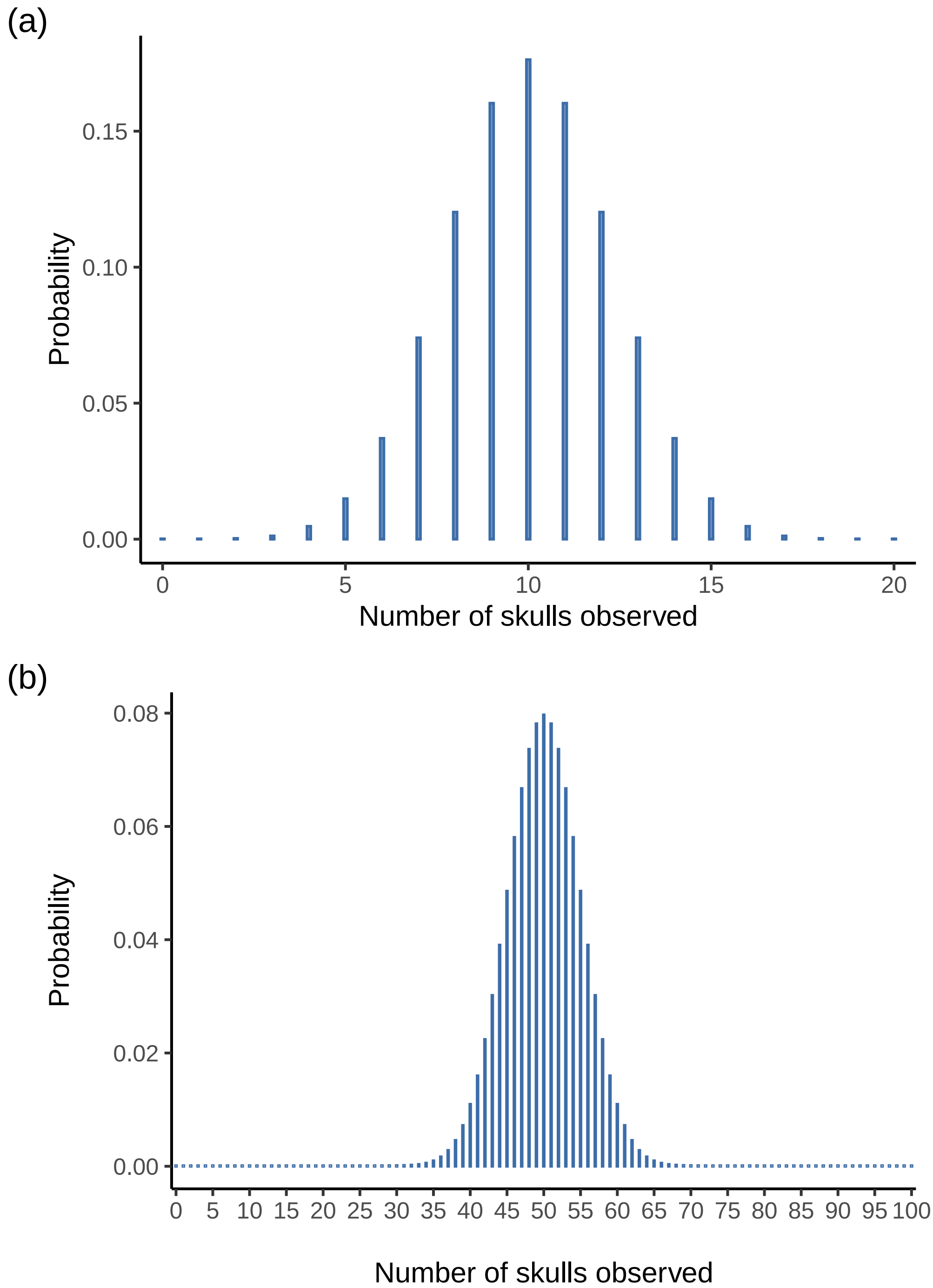


Figure 7.4: Two binomial distributions, involving a scenario in which I flip a fair coin so the underlying success probability is $\theta = \frac{1}{2}$. In panel (a), I flipped the coin $N = 20$ times. In panel (b) the coin was flipped $N = 100$ times

Let's try to get a sense for what it means for a variable to be normally distributed. To that end, have a look at Figure 7.5 which plots a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. You can see where the name "bell curve" comes from; it looks a bit like a bell. Notice that, unlike the plots that I drew to illustrate the binomial distribution, the picture of the normal distribution in Figure 7.5 shows a smooth curve instead of "histogram-like" bars. This isn't an arbitrary choice, the normal distribution is continuous whereas the binomial is discrete. For instance, in the die rolling example from the last section it was possible to get 3 skulls or 4 skulls, but impossible to get 3.9 skulls. The figures that I drew in the previous section reflected this fact. In Figure 7.3, for instance, there's a bar located at $X = 3$ and another one at $X = 4$ but there's nothing in between. Continuous quantities don't have this constraint. For instance, suppose we're talking about the weather. The temperature on a pleasant Spring day could be 23 degrees, 24 degrees, 23.9 degrees, or anything in between since temperature is a continuous variable. And so a normal distribution might be quite appropriate for describing Spring temperatures.[6]

With this in mind, let's see if we can't get an intuition for how the normal distribution works. First, let's have a look at what happens when we play around with the parameters of the distribution. To that end, Figure 7.6 plots normal distributions that have different means but have the same standard deviation. As you might expect, all of these distributions have the same "width". The only difference between them is that they've been shifted to the left or to the right. In every other respect they're identical. In contrast, if we increase the standard deviation while keeping the mean constant, the peak of the distribution stays in the same place but the distribution gets wider, as you can see in Figure 7.7. Notice, though, that when we widen the distribution the height of the peak shrinks. This has to happen, in the same way that the heights of the bars that we used to draw a discrete binomial distribution have to sum to 1, the total area under the curve for the normal distribution must equal 1. Before moving on, I want to point out one important characteristic of the normal distribution. Irrespective of what the actual mean and standard deviation are, 68.3% of the area falls within 1 standard deviation of the mean. Similarly, 95.4% of the distribution falls within 2 standard deviations of the mean, and (99.7%) of the distribution is within 3 standard deviations. This idea is illustrated in Figure 7.8; see also Figure 7.9.

7.5.1 Probability density

There's something I've been trying to hide throughout my discussion of the normal distribution, something that some introductory textbooks omit completely. They might be right to do so. This "thing" that I'm hiding is weird and counter-intuitive even by the admittedly distorted standards that apply in statistics. Fortunately, it's not something that you need to understand at a deep level in order to do basic statistics. Rather, it's something that starts to become important later on when you move beyond the basics. So, if it doesn't make complete sense, don't worry too much, but try to make sure that you follow the gist of it.

Throughout my discussion of the normal distribution there's been one or two things that

[6] In practice, the normal distribution is so handy that people tend to use it even when the variable isn't actually continuous. As long as there are enough categories (e.g., Likert-scale responses to a questionnaire), it's pretty standard practice to use the normal distribution as an approximation. This works out much better in practice than you'd think.

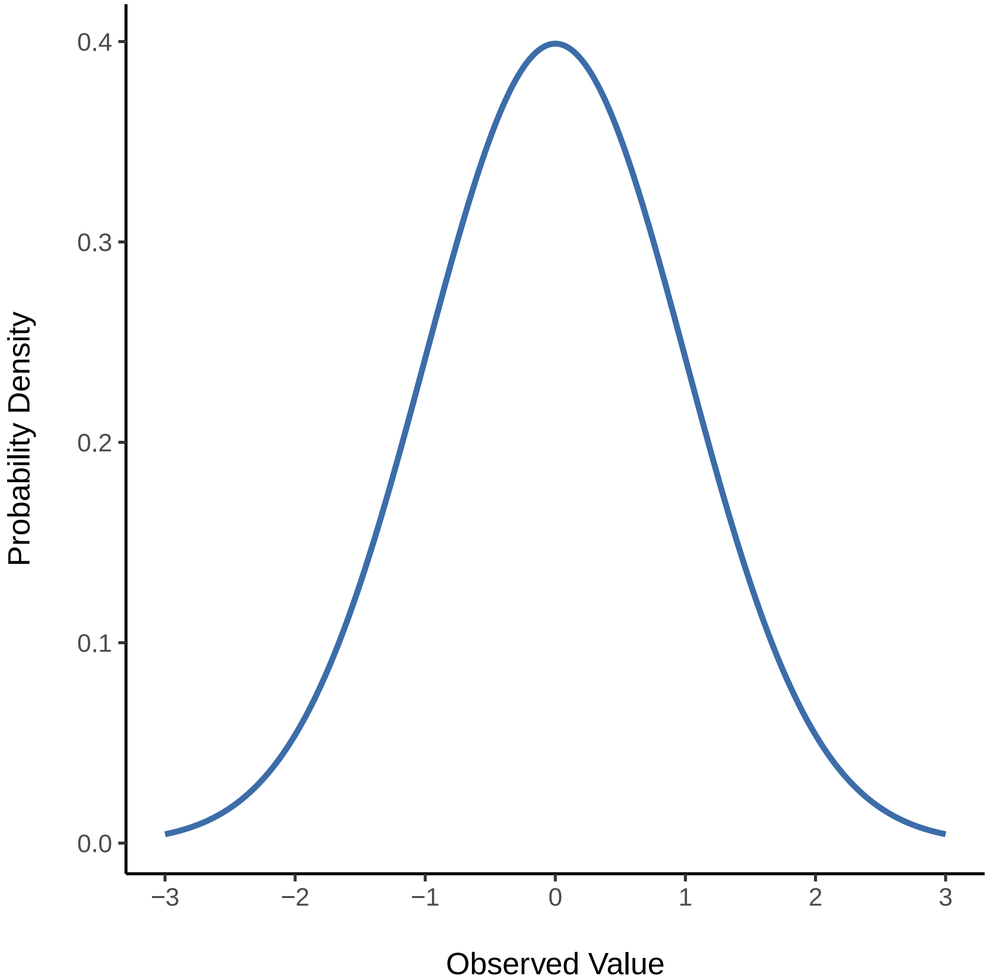


Figure 7.5: The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The x-axis corresponds to the value of some variable, and the y-axis tells us something about how likely we are to observe that value. However, notice that the y-axis is labelled *Probability Density* and not *Probability*. There is a subtle and somewhat frustrating characteristic of continuous distributions that makes the y-axis behave a bit oddly - the height of the curve here is not actually the probability of observing a particular x value. On the other hand, it is true that the heights of the curve tells you which x values are more likely (the higher ones!). (see Probability density section for all the annoying details)

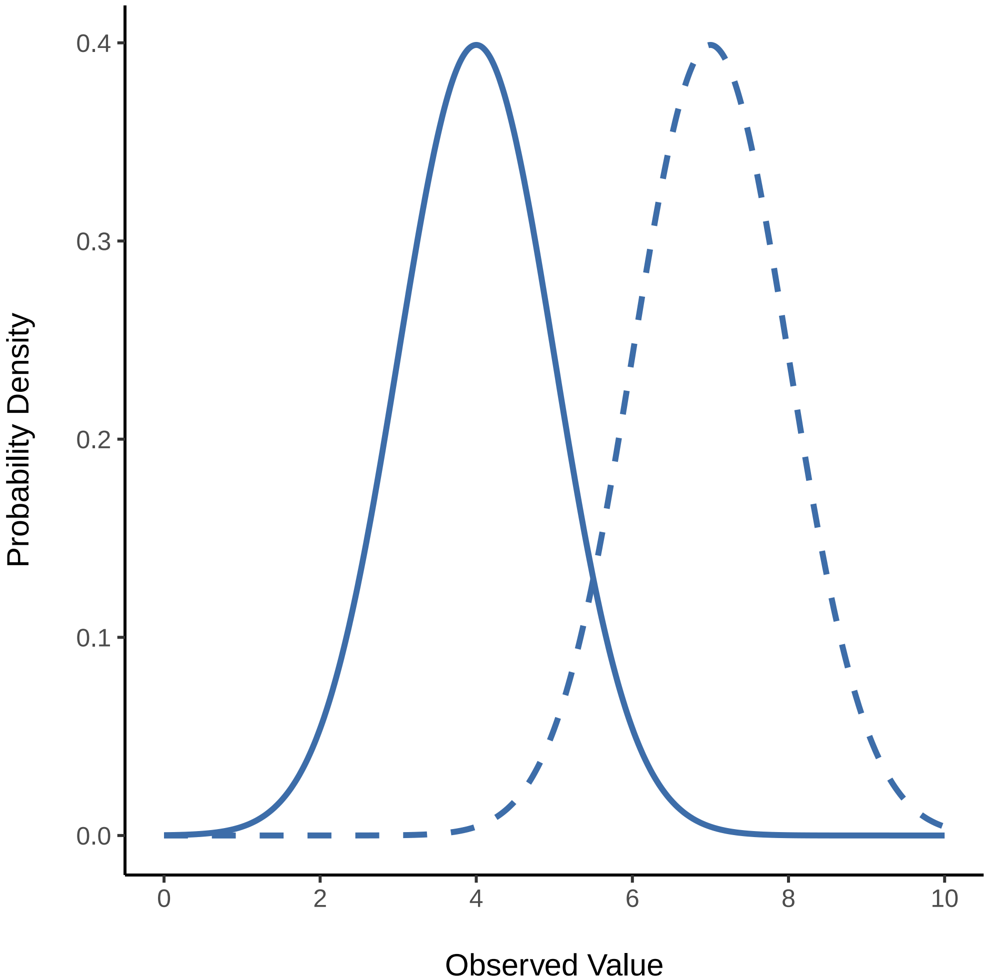


Figure 7.6: An illustration of what happens when you change the mean of a normal distribution. The solid line depicts a normal distribution with a mean of $\mu = 4$. The dashed line shows a normal distribution with a mean of $\mu = 7$. In both cases, the standard deviation is $\sigma = 1$. Not surprisingly, the two distributions have the same shape, but the dashed line is shifted to the right

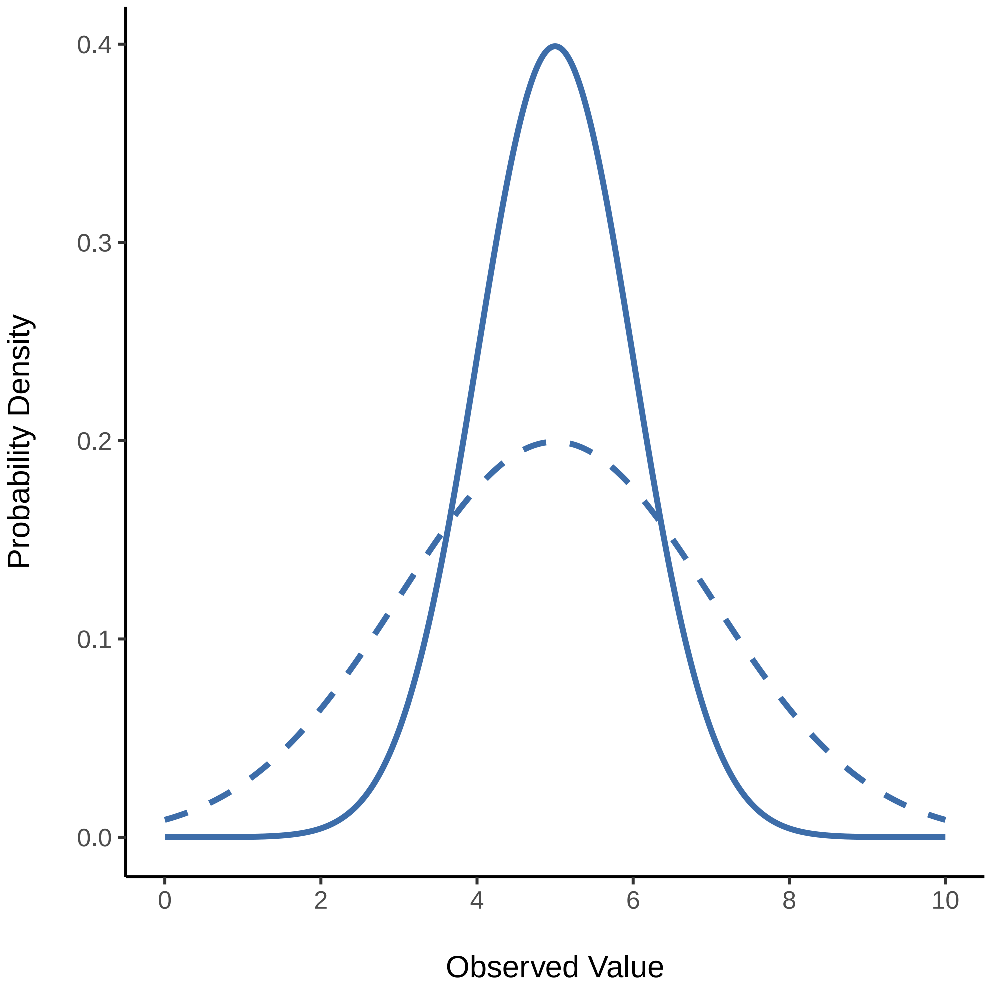


Figure 7.7: An illustration of what happens when you change the the standard deviation of a normal distribution. Both distributions plotted in this figure have a mean of $\mu = 5$, but they have different standard deviations. The solid line plots a distribution with standard deviation $\sigma = 1$, and the dashed line shows a distribution with standard deviation $\sigma = 2$. As a consequence, both distributions are "centred" on the same spot, but the dashed line is wider than the solid one

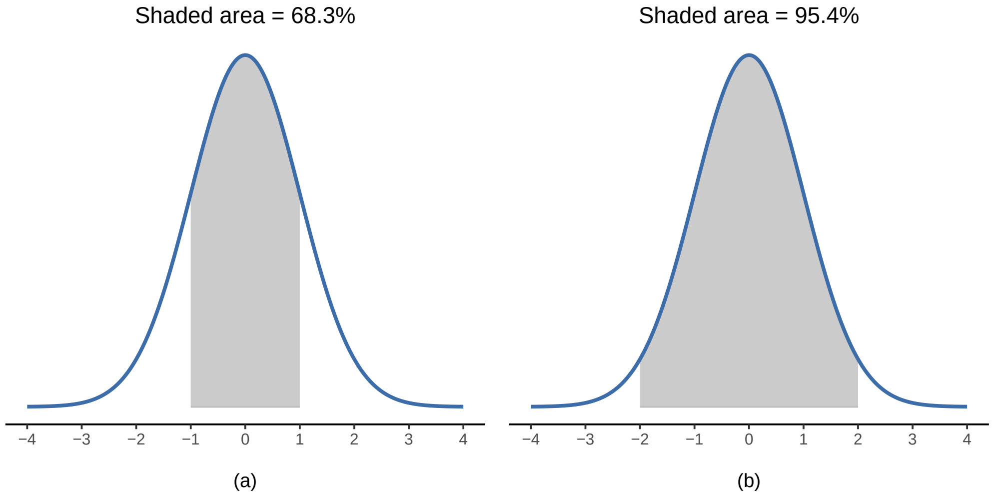


Figure 7.8: The area under the curve tells you the probability that an observation falls within a particular range. The solid lines plot normal distributions with mean $\mu = 0$ and standard deviation $\sigma = 1$. The shaded areas illustrate "areas under the curve" for two important cases. In panel (a), we can see that there is a 68.3% chance that an observation will fall within one standard deviation of the mean. In panel (b), we see that there is a 95.4% chance that an observation will fall within two standard deviations of the mean

don't quite make sense. Perhaps you noticed that the y-axis in these figures is labelled "Probability Density" rather than density. Maybe you noticed that I used $p(X)$ instead of $P(X)$ when giving the formula for the normal distribution.

As it turns out, what is presented here isn't actually a probability, it's something else. To understand what that something is you have to spend a little time thinking about what it really means to say that X is a continuous variable. Let's say we're talking about the temperature outside. The thermometer tells me it's 23 degrees, but I know that's not really true. It's not exactly 23 degrees. Maybe it's 23.1 degrees, I think to myself. But I know that that's not really true either because it might actually be 23.09 degrees. But I know that... well, you get the idea. The tricky thing with genuinely continuous quantities is that you never really know exactly what they are.

Now think about what this implies when we talk about probabilities. Suppose that tomorrow's maximum temperature is sampled from a normal distribution with mean 23 and standard deviation 1. What's the probability that the temperature will be exactly 23 degrees? The answer is "zero", or possibly "a number so close to zero that it might as well be zero". Why is this? It's like trying to throw a dart at an infinitely small dart board. No matter how good your aim, you'll never hit it. In real life you'll never get a value of exactly 23. It'll always be something like 23.1 or 22.99998 or suchlike. In other words, it's completely meaningless to talk about the probability that the temperature is exactly 23 degrees. However, in everyday language if I told you that it was 23 degrees outside and it turned out to be 22.9998 degrees you probably wouldn't call me a liar. Because in everyday language "23 degrees" usually means something like "somewhere

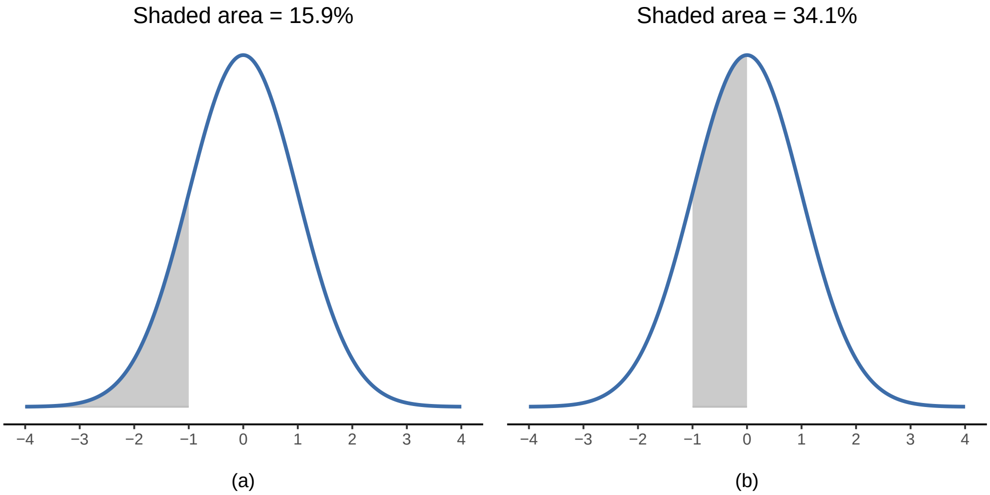


Figure 7.9: Two more examples of the "area under the curve idea". There is a 15.9% chance that an observation is one standard deviation below the mean or smaller (panel (a)), and a 34.1% chance that the observation is somewhere between one standard deviation below the mean and the mean (panel (b)). Notice that if you add these two numbers together you get 15.9% + 34.1% = 50%. For normally distributed data, there is a 50% chance that an observation falls below the mean. And of course that also implies that there is a 50% chance that it falls above the mean

between 22.5 and 23.5 degrees". And while it doesn't feel very meaningful to ask about the probability that the temperature is exactly 23 degrees, it does seem sensible to ask about the probability that the temperature lies between 22.5 and 23.5, or between 20 and 30, or any other range of temperatures.

The point of this discussion is to make clear that when we're talking about continuous distributions it's not meaningful to talk about the probability of a specific value. However, what we can talk about is the probability that the value lies within a particular range of values. To find out the probability associated with a particular range what you need to do is calculate the "area under the curve". We've seen this concept already, in Figure 7.8 the shaded areas shown depict genuine probabilities (e.g., in Figure 7.8) it shows the probability of observing a value that falls within 1 standard deviation of the mean).

Okay, so that explains part of the story. I've explained a little bit about how continuous probability distributions should be interpreted (i.e., area under the curve is the key thing). But what does the formula for $p(x)$ that I described earlier actually mean? Obviously, $P(x)$ doesn't describe a probability, but what is it? The name for this quantity $P(x)$ is a **probability density**, and in terms of the plots we've been drawing it corresponds to the height of the curve. The densities themselves aren't meaningful in and of themselves, but they're "rigged" to ensure that the area under the curve is always interpretable as genuine probabilities. To be honest, that's about as much as you really need to know for now.[7]

[7]For those readers who know a little calculus, I'll give a slightly more precise explanation. In the

7.6 Other useful distributions

The normal distribution is the distribution that statistics makes most use of (for reasons to be discussed shortly), and the binomial distribution is a very useful one for lots of purposes. But the world of statistics is filled with probability distributions, some of which we'll run into in passing. In particular, the three that will appear in this book are the t distribution, the χ^2 distribution and the F distribution. I won't give formulas for any of these, or talk about them in too much detail, but I will show you some pictures: Figure 7.10, Figure 7.11 and Figure 7.12.

- The t distribution is a continuous distribution that looks very similar to a normal distribution, see Figure 7.10. Note that the "tails" of the t distribution are "heavier" (i.e., extend further outwards) than the tails of the normal distribution). That's the important difference between the two. This distribution tends to arise in situations where you think that the data actually follow a normal distribution, but you don't know the mean or standard deviation. We'll run into this distribution again in **?@sec-Comparing-two-means**.
- The χ^2 distribution is another distribution that turns up in lots of different places. The situation in which we'll see it is when doing categorical data analysis in **?@sec-Categorical-data-analysis**, but it's one of those things that actually pops up all over the place. When you dig into the maths (and who doesn't love doing that?), it turns out that the main reason why the χ^2 distribution turns up all over the place is that if you have a bunch of variables that are normally distributed, square their values and then add them up (a procedure referred to as taking a "sum of squares"), this sum has a χ^2 distribution. You'd be amazed how often this fact turns out to be useful. Anyway, here's what a χ^2 distribution looks like: Figure 7.11.
- The F distribution looks a bit like a χ^2 distribution, and it arises whenever you need to compare two χ^2 distributions to one another. Admittedly, this doesn't exactly sound like something that any sane person would want to do, but it turns out to be very important in real-world data analysis. Remember when I said that χ^2 turns out to be the key distribution when we're taking a "sum of squares"? Well, what that means is if you want to compare two different "sums of squares", you're probably talking about something that has an F distribution. Of course, as of yet I still haven't given you an example of anything that involves a sum of squares, but I will in **?@sec-Comparing-several-means-one-way-ANOVA**. And that's where we'll run into the F distribution. Oh, and there's a picture in Figure 7.12.

Okay, time to wrap this section up. We've seen three new distributions: χ^2, t and F. They're all continuous distributions, and they're all closely related to the normal distribution. The main thing for our purposes is that you grasp the basic idea that these distributions are all deeply related to one another, and to the normal distribution. Later on in this book we're going to run into data that are normally distributed, or at

same way that probabilities are non-negative numbers that must sum to 1, probability densities are non-negative numbers that must integrate to 1 (where the integral is taken across all possible values of X). To calculate the probability that X falls between a and b we calculate the definite integral of the density function over the corresponding range, $\int \_a^b p(x)dx$. If you don't remember or never learned calculus, don't worry about this. It's not needed for this book.

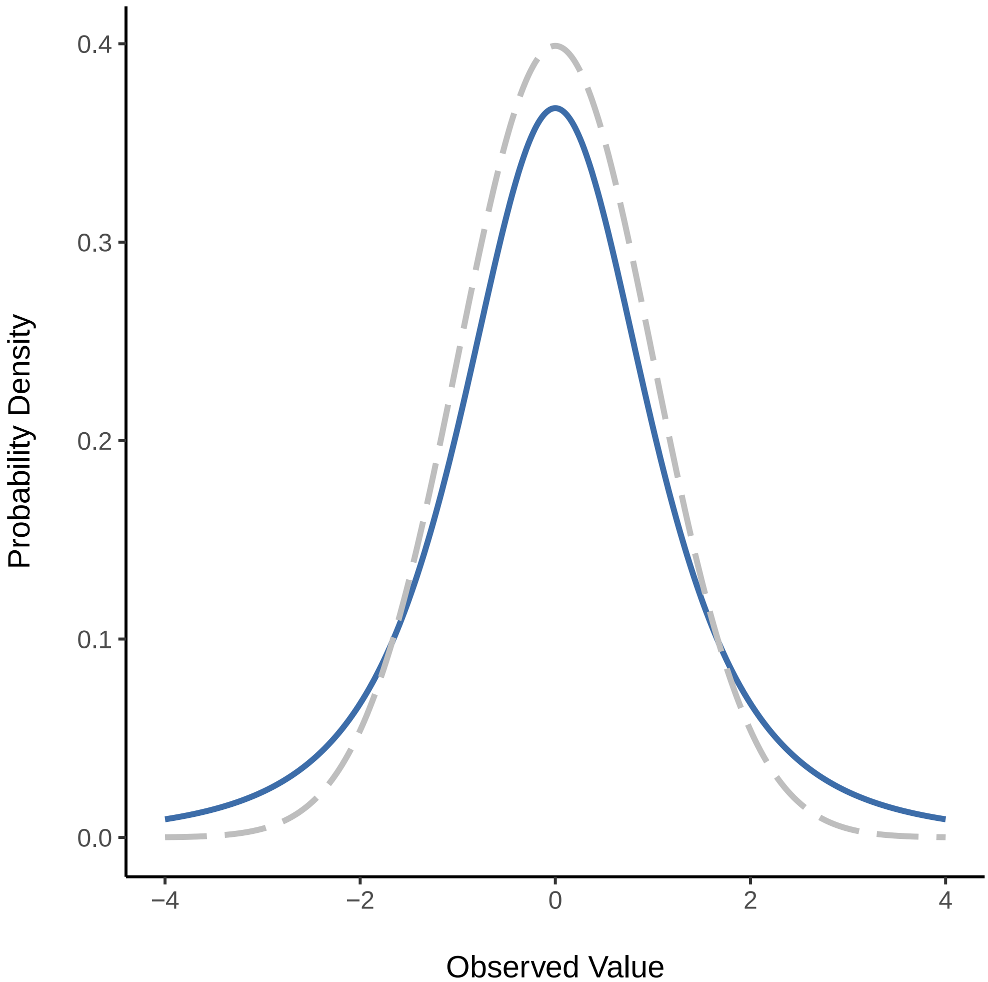


Figure 7.10: A t distribution with 3 degrees of freedom (solid line). It looks similar to a normal distribution, but it is not quite the same. For comparison purposes I have plotted a standard normal distribution as the dashed line

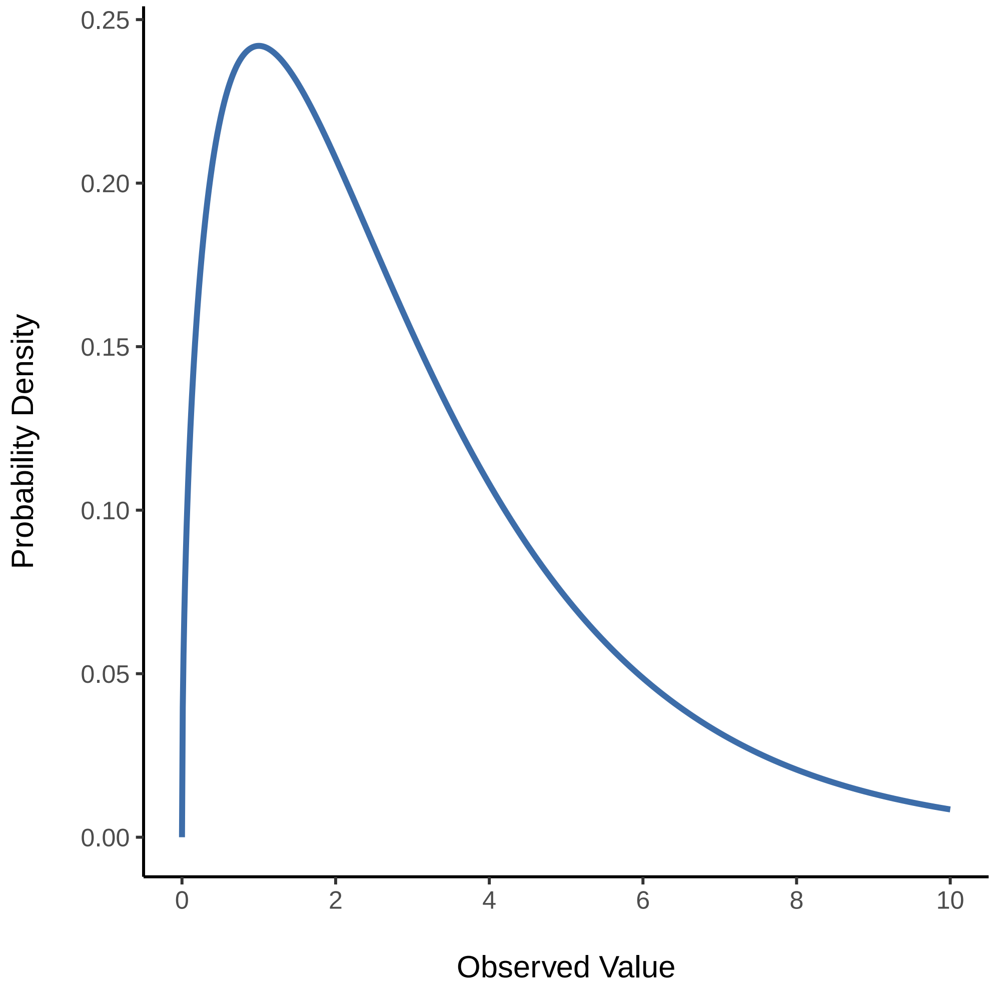


Figure 7.11: χ^2 distribution with 3 degrees of freedom. Notice that the observed values must always be greater than zero, and that the distribution is pretty skewed. These are the key features of a chi-square distribution

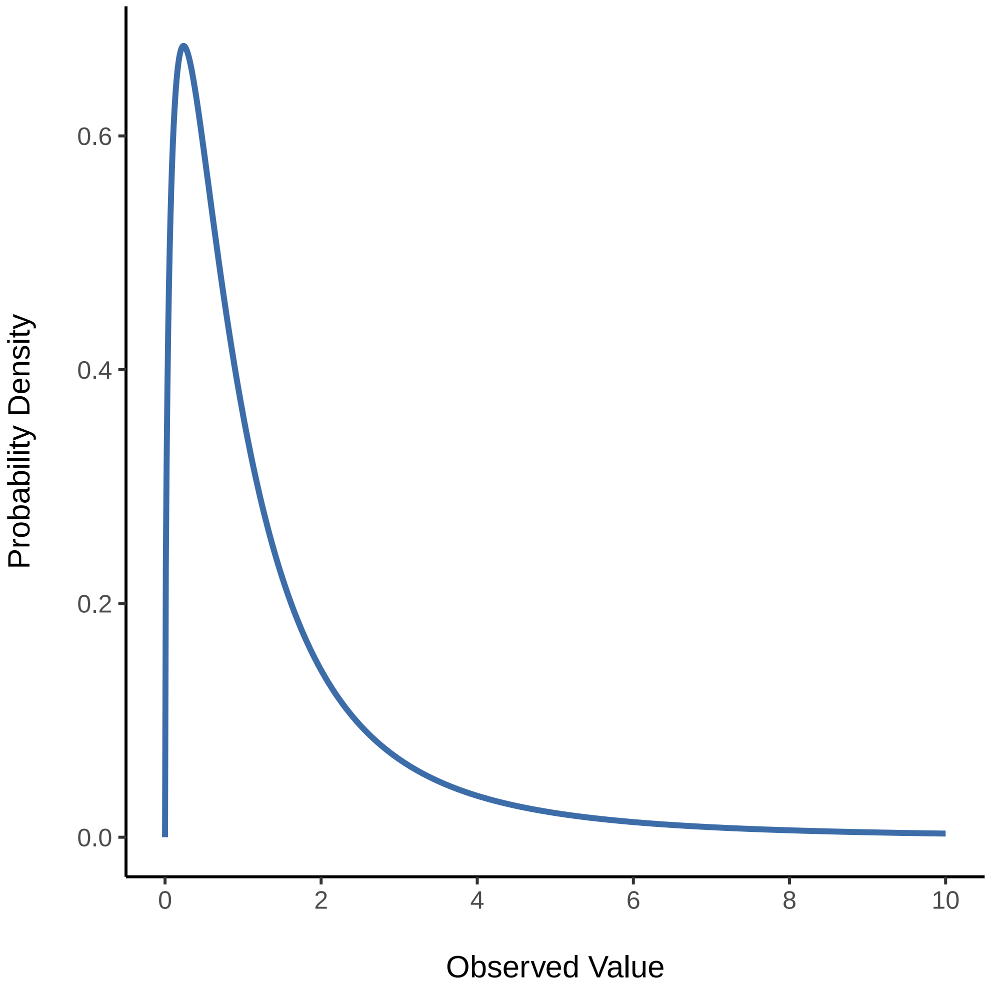


Figure 7.12: An F distribution with 3 and 5 degrees of freedom. Qualitatively speaking, it looks pretty similar to a chi-square distribution, but they are not quite the same in general

least assumed to be normally distributed. What I want you to understand right now is that, if you make the assumption that your data are normally distributed, you shouldn't be surprised to see χ^2, t and F distributions popping up all over the place when you start trying to do your data analysis.

7.7 Summary

In this chapter we've talked about probability. We've talked about what probability means and why statisticians can't agree on what it means. We talked about the rules that probabilities have to obey. And we introduced the idea of a probability distribution and spent a good chunk of the chapter talking about some of the more important probability distributions that statisticians work with. The section-by-section breakdown looks like this:

- Probability theory versus statistics: How are probability and statistics different?.
- The frequentist view versus The Bayesian view of probability.
- Basic probability theory.
- The binomial distribution, The normal distribution, and Other useful distributions.

As you'd expect, my coverage is by no means exhaustive. Probability theory is a large branch of mathematics in its own right, entirely separate from its application to statistics and data analysis. As such, there are thousands of books written on the subject and universities generally offer multiple classes devoted entirely to probability theory. Even the "simpler" task of documenting standard probability distributions is a big topic. I've described five standard probability distributions in this chapter, but sitting on my bookshelf I have a 45-chapter book called "Statistical Distributions" (M. Evans et al., 2011) that lists a lot more than that. Fortunately for you, very little of this is necessary. You're unlikely to need to know dozens of statistical distributions when you go out and do real-world data analysis, and you definitely won't need them for this book, but it never hurts to know that there's other possibilities out there.

Picking up on that last point, there's a sense in which this whole chapter is something of a digression. Many undergraduate psychology classes on statistics skim over this content very quickly (I know mine did), and even the more advanced classes will often "forget" to revisit the basic foundations of the field. Most academic psychologists would not know the difference between probability and density, and until recently very few would have been aware of the difference between Bayesian and frequentist probability. However, I think it's important to understand these things before moving onto the applications. For example, there are a lot of rules about what you're "allowed" to say when doing statistical inference and many of these can seem arbitrary and weird. However, they start to make sense if you understand that there is this Bayesian vs. frequentist distinction. Similarly, in **?@sec-Comparing-two-means** we're going to talk about something called the t-test, and if you really want to have a grasp of the mechanics of the t-test it really helps to have a sense of what a t-distribution actually looks like. You get the idea, I hope.

Chapter 8

Estimating unknown quantities from a sample

At the start of the last chapter I highlighted the critical distinction between descriptive statistics and *inferential statistics*. As discussed in Chapter 4, 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". Now that we have a foundation in probability theory 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 sampling theory first because estimation theory doesn't make sense until you understand sampling. As a consequence, this chapter divides naturally into two parts, the first three sections are focused on sampling theory, and the last two sections make use of sampling theory to discuss how statisticians think about estimation.

8.1 Samples, populations and sampling

In the Prelude to part IV 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 is 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 foundation 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 an experiment with some number of participants, a polling company might have phoned some number of people to ask questions about voting intentions, and so on. In this way the data set available to us is finite and incomplete. We can't possibly get every person in the world to do our experiment, for example a polling company doesn't have the time or the money to ring up every voter in the country. In our earlier discussion of descriptive statistics in Chapter 4 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.

8.1.1 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 with the data does depend on the population about which you want to make statements.

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 time of the study, millions of people. The sample was a set of 1000 people who all belong to that population. In most studies the situation is much less straightforward. In a typical 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 to part IV. The sample here is a specific sequence of 12 wins and 0 losses for Wellesley. What is the population? Again, it's not obvious what the population is.

- 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?

8.1.2 Simple random samples

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 8.1. 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 8.1. 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 is what we refer to as a *random process*.[1] 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 five-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 8.2. Now consider the evidential 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

[1] The 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.

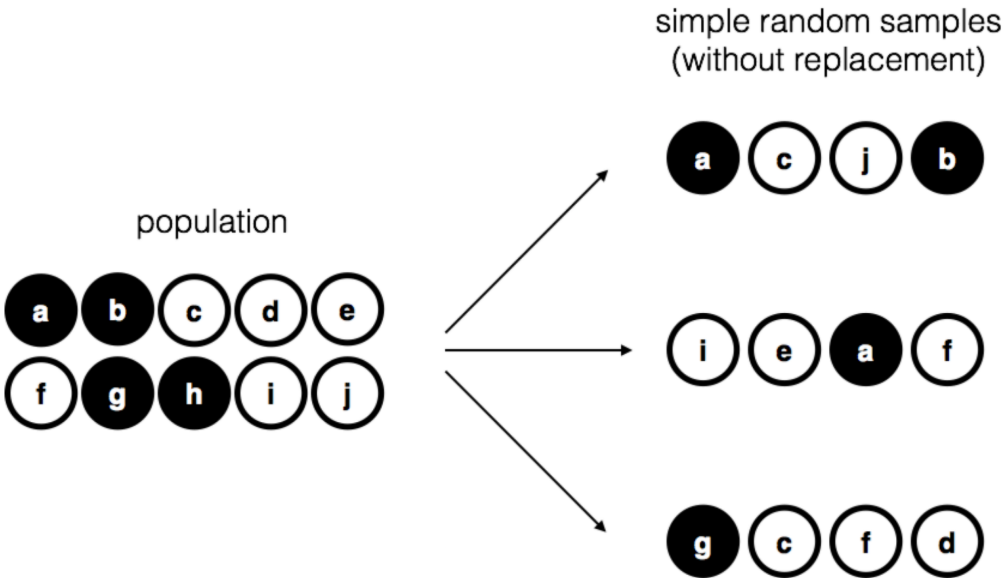


Figure 8.1: Simple random sampling without replacement from a finite population

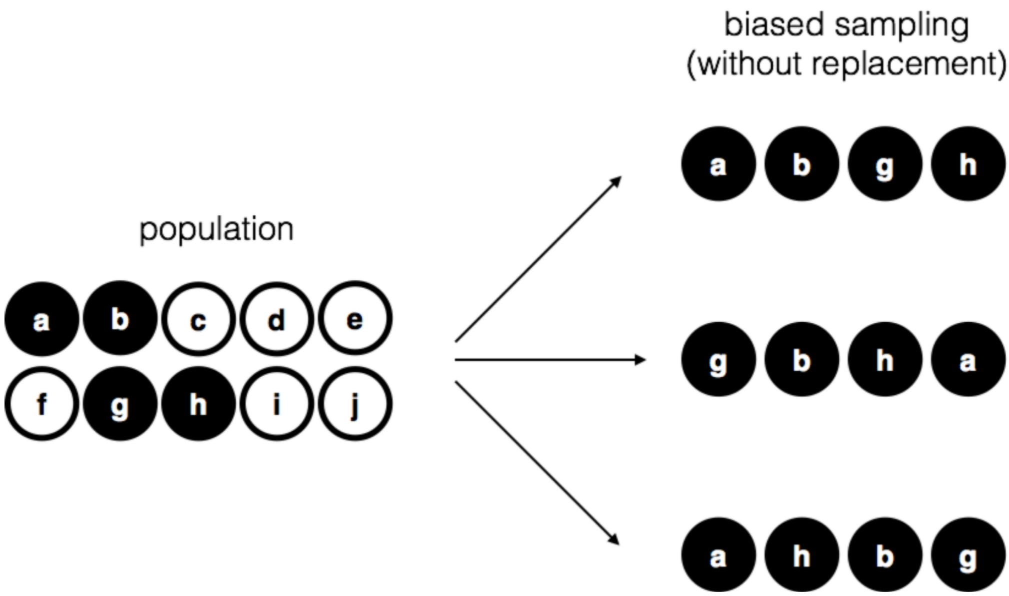


Figure 8.2: Biased sampling without replacement from a finite population

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 8.3.

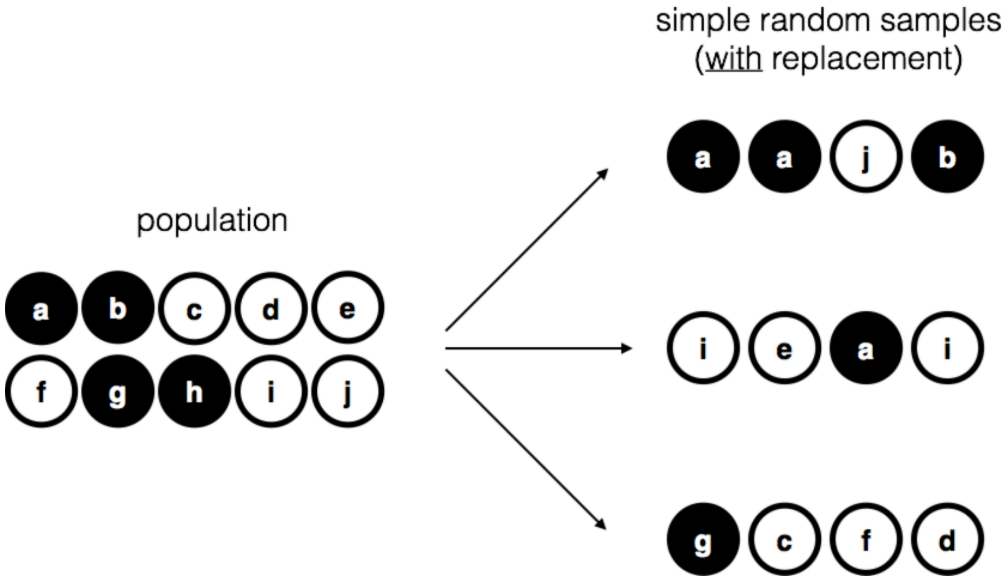


Figure 8.3: Simple random sampling *with* replacement from a finite population

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 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.

8.1.3 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 sub-populations, 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 sub-populations are rare. For instance, when studying schizophrenia it would be much better to divide the population into two[2] 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 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 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 issue. 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. First, reliance on undergraduate psychology students automatically means that your data are restricted to a single sub-population. Second, the students usually get to pick which studies they participate in, so the sample is a self selected subset of psychology students and 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.

[2]Nothing 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.

8.1.4 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 Figure 8.1 and Figure 8.2. 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, and 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 much more than people who haven't grown up in a similar environment. This could therefore lead to a misleading picture of what working memory capacity is.

There are two points hidden in this discussion. First, 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. Second, 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.

8.1.5 Population parameters and sample statistics

Okay. Setting aside the thorny methodological issues associated with obtaining a random sample, 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 2.1), statisticians operationalise the concept of a "population" in terms of mathematical objects that they know how to work with. You've already come across these objects in Chapter 7. 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 8.4 (a). 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.

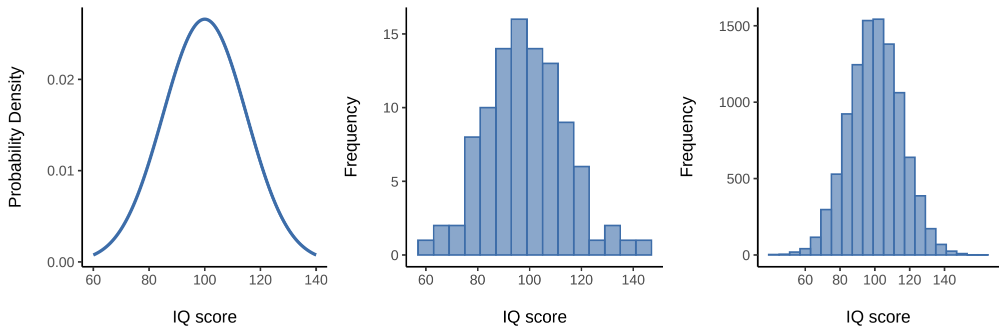


Figure 8.4: The population distribution of IQ scores (panel (a)) and two samples drawn randomly from it. In panel (b) we have a sample of 100 observations, and panel (c) we have a sample of 10,000 observations

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 8.4 (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 8.4 (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 Estimating population parameters using your sample statistics and also Estimating a confidence interval but before we get to that there's a few more ideas in sampling theory that you need to know about

8.2 The law of large numbers

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 as 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 jamovi. The *IQsim.omv* file is a jamovi data file. In this file I have generated 10,000 random numbers sampled from a normal distribution for a population with mean = 100 and sd = 15. This was done by computing a new variable using the = NORM(100,15) function. A histogram and density plot shows that this 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 99.68 and the standard deviation is 14.90. These values are now very close to the true population. See Figure 8.5.

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, p. 65).

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**.

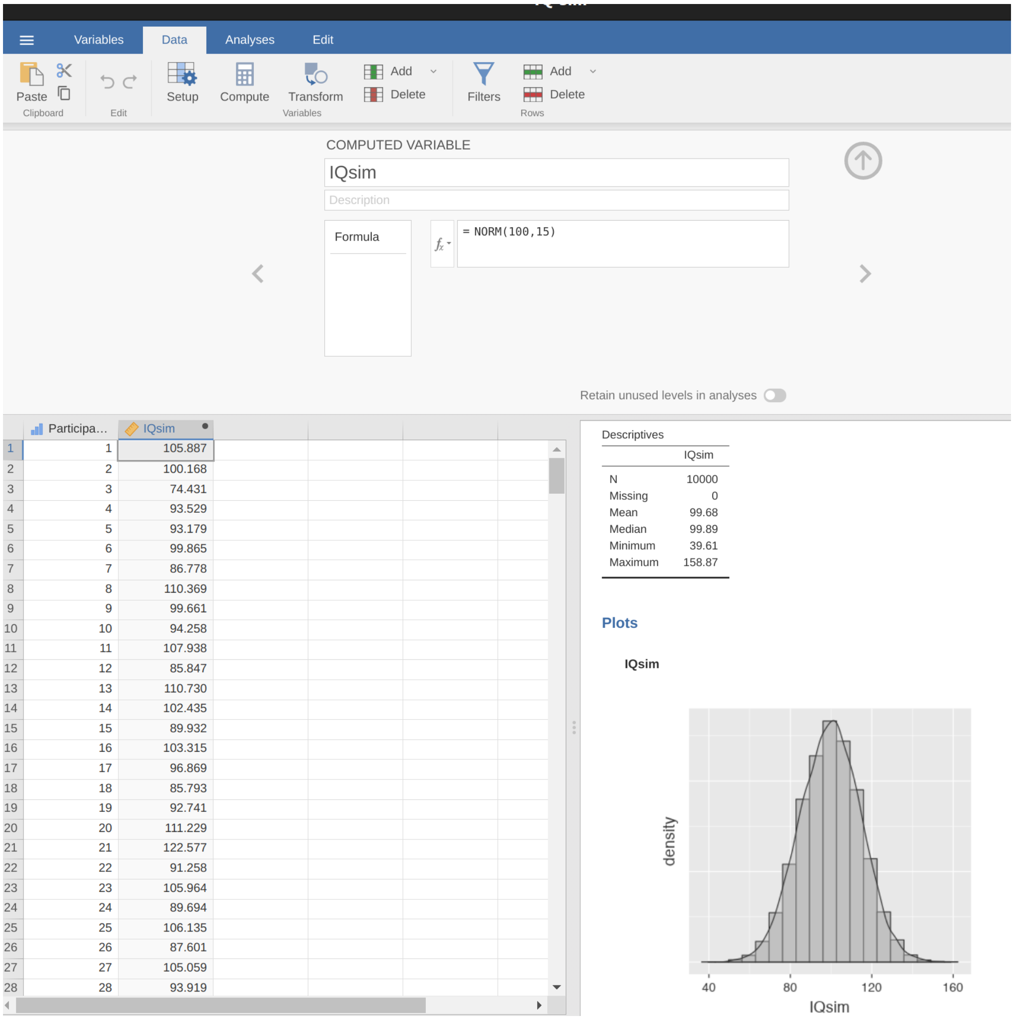


Figure 8.5: A random sample drawn from a normal distribution using jamovi

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 $N \longrightarrow \infty$), the sample mean approaches the population mean $\bar{X} \longrightarrow \mu$)[3]

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 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.

8.3 Sampling distributions and the central limit theorem

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, p. 80).

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!

[3]Technically, 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.

8.3.1 Sampling distribution of the mean

With this in mind, let's abandon the idea that our studies will have sample sizes of 10,000 and consider instead 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 jamovi = NORM(100,15) function, but I only need 5 participant IDs this time, not 10,000. These are the five numbers that jamovi generated:

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 and I randomly sample 5 new people and measure their IQ. Again, jamovi allows me to simulate the results of this procedure, and generates these five numbers:

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 8.1, and as you can see the sample mean varies from one replication to the next.

Table 8.1: Ten replications of the IQ experiment, each with a sample size of ($N = 5$)

	Person 1	**Person 2**	**Person 3**	**Person 4**	**Person 5**	**Sample Mean**
Rep. 1	90	82	94	99	110	95.0
Rep. 2	78	88	111	111	117	101.0
Rep. 3	111	122	91	98	86	101.6
Rep. 4	98	96	119	99	107	103.8
Rep. 5	105	113	103	103	98	104.4
Rep. 6	81	89	93	85	114	92.4
Rep. 7	100	93	108	98	133	106.4
Rep. 8	107	100	105	117	85	102.8
Rep. 9	86	119	108	73	116	100.4
Rep. 10	95	126	112	120	76	105.8

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 8.1, 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. Well that's exactly what I did, and you can see the results in Figure 8.6. 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! (Table 8.1). This distribution has a special name in statistics, it's called the **sampling distribution of the mean**.

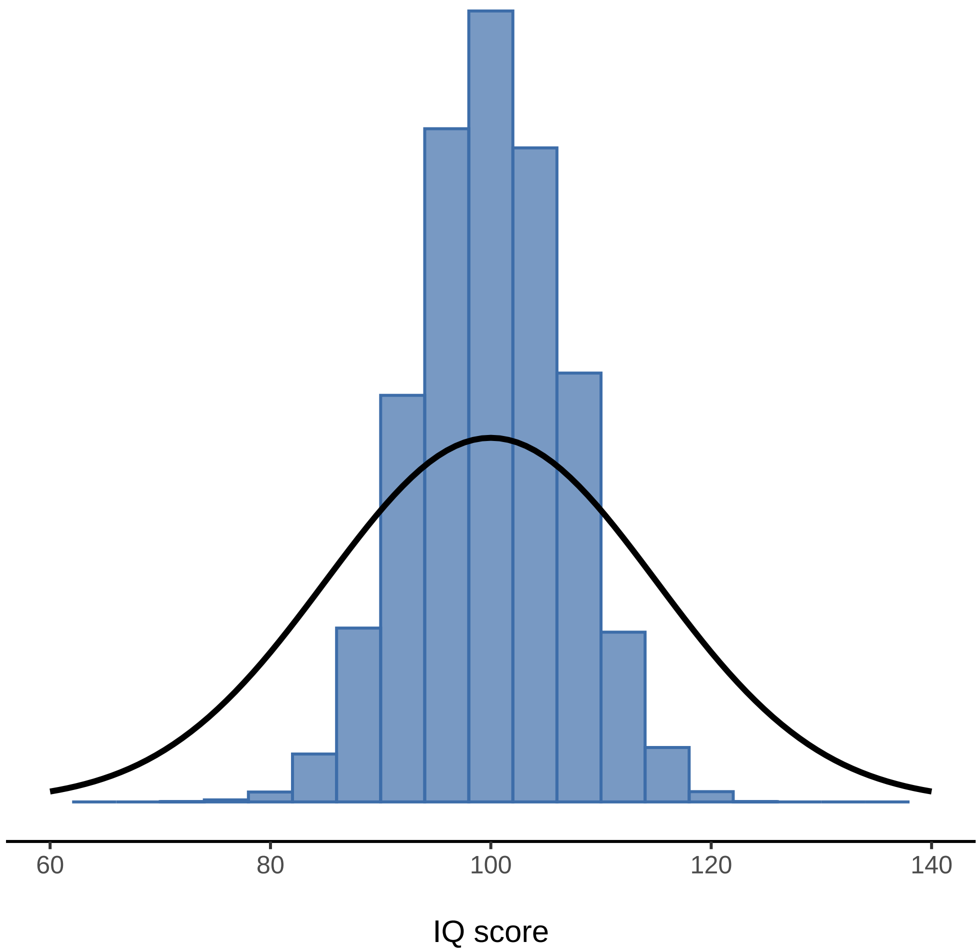


Figure 8.6: The sampling distribution of the mean for the "five IQ scores experiment". If you sample 5 people at random and calculate their average IQ you will almost certainly get a number between 80 and 120, even though there are quite a lot of individuals who have IQs above 120 or below 80. For comparison, the black line plots the population distribution of IQ scores

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 8.6 tells us, though, is that the "five IQ scores" experiment is not very accurate. If I repeat the experiment, the sampling distribution tells me that I can expect to see a sample mean anywhere between 80 and 120.

8.3.2 Sampling distributions exist for any sample statistic!

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 8.7. 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.

8.3.3 The central limit theorem

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 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 Figure 8.8, showing that 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. To do this, I started with the "ramped" distribution shown in the histogram in Figure 8.9. As you can see by comparing the triangular shaped histogram

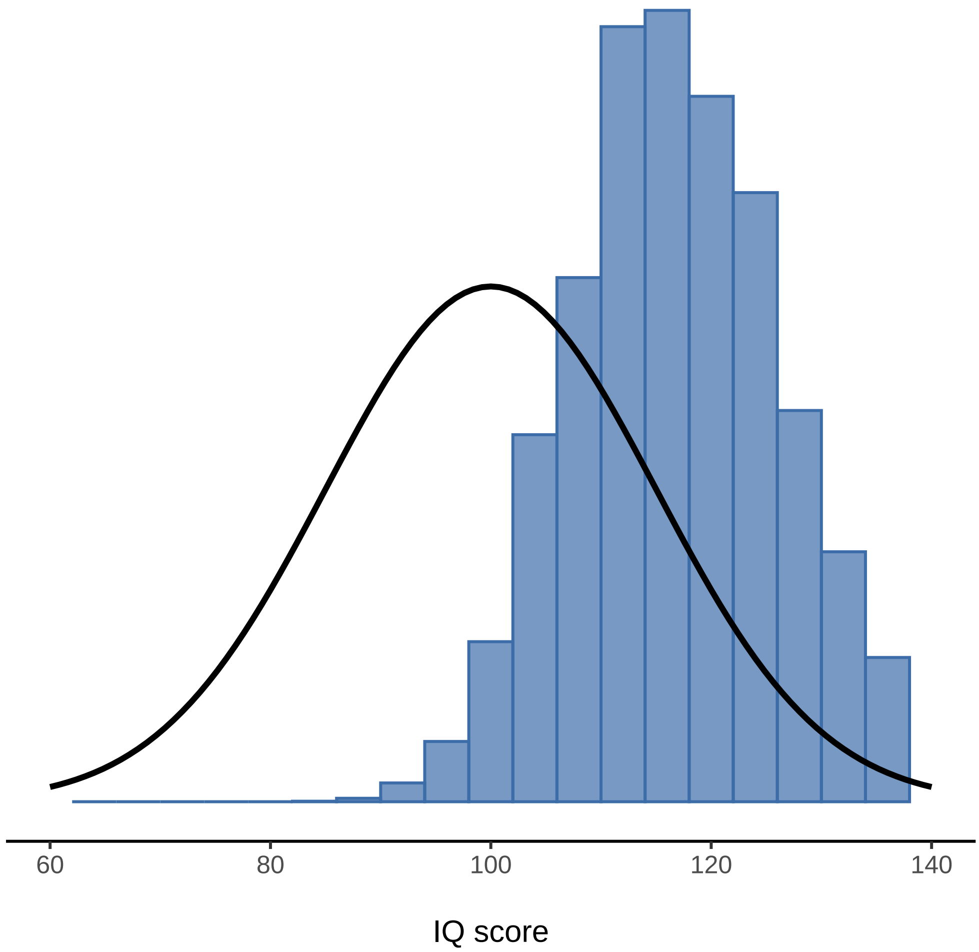


Figure 8.7: The sampling distribution of the maximum for the "five IQ scores experiment". If you sample 5 people at random and select the one with the highest IQ score you will probably see someone with an IQ between 100 and 140

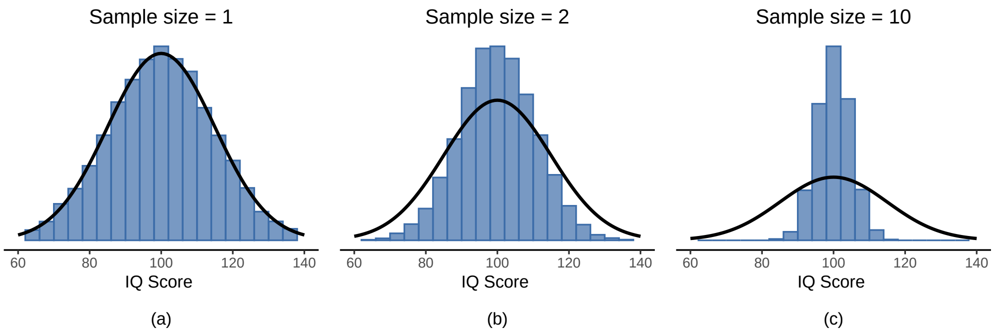


Figure 8.8: An illustration of the how sampling distribution of the mean depends on sample size. In each panel I generated 10,000 samples of IQ data and calculated the mean IQ observed within each of these data sets. The histograms in these plots show the distribution of these means (i.e., the sampling 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. In panel (a), each data set contained only a single observation, so the mean of each sample is the IQ score of just one person. As a consequence, the sampling distribution of the mean is of course identical to the population distribution of IQ scores. However, when we raise the sample size to 2 the mean of any one sample tends to be closer to the population mean than the IQ score of any one person, and so the histogram (i.e., the sampling distribution) is a bit narrower than the population distribution. By the time we raise the sample size to 10 (panel (c)), we can see that the distribution of sample means tend to be fairly tightly clustered around the true population mean

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 simulated 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 8.9 (b) plots the histogram of these sample means (i.e., the sampling distribution of the mean for $N = 2$). This time, the histogram produces a χ^2-shaped distribution. It's still not normal, but it's a lot closer to the black line than the population distribution in Figure 8.9 (a). When I increase the sample size to $N = 4$, the sampling distribution of the mean is very close to normal (Figure 8.9 (c)), 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!

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

$$SEM = \frac{\sigma}{\sqrt{N}}$$

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.[4]

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] As 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.

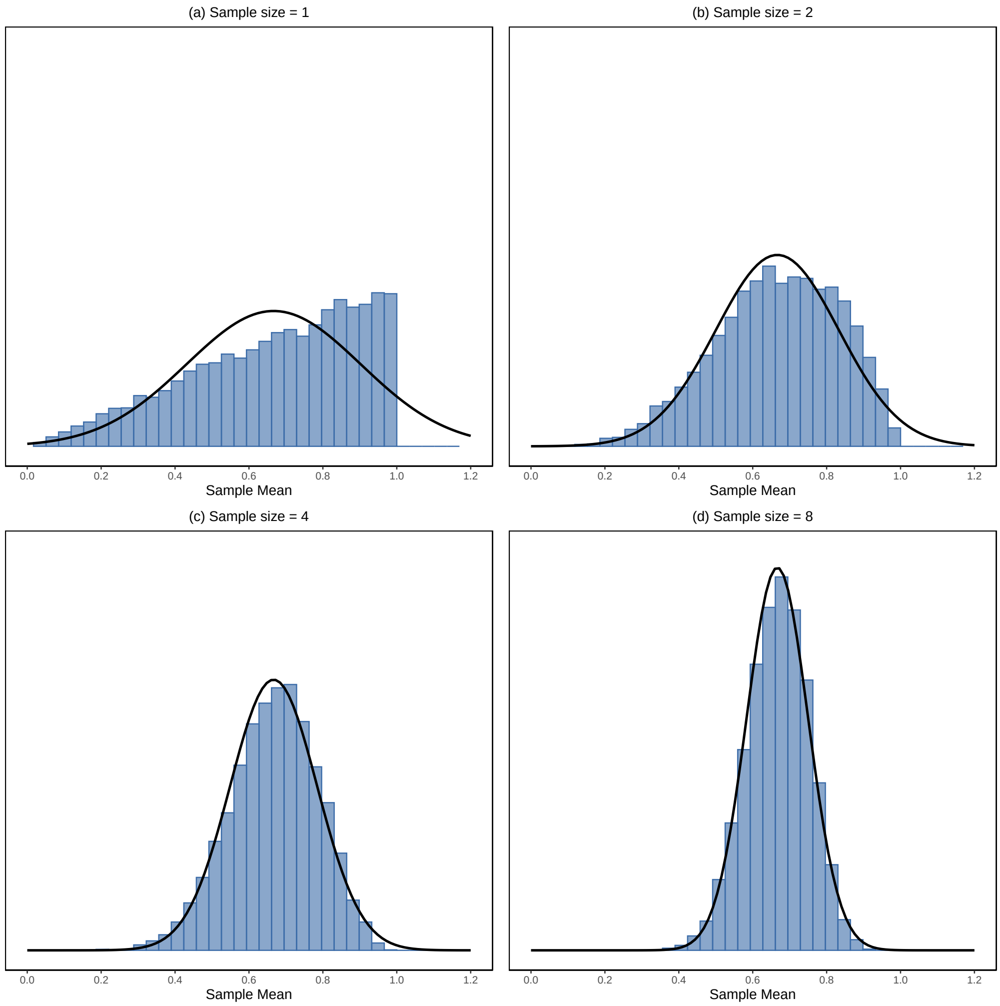


Figure 8.9: A demonstration of the central limit theorem. In panel (a), we have a non-normal population distribution, and panels (b)-(d) show the sampling distribution of the mean for samples of size 2,4 and 8 for data drawn from the distribution in panel (a). As you can see, even though the original population distribution is non-normal the sampling distribution of the mean becomes pretty close to normal by the time you have a sample of even four observations

8.4 Estimating population parameters

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 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.[5] 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?

8.4.1 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.

[5] Please 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 2, 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.

In this example estimating the unknown population parameter is straightforward. I calculate the sample mean and I use that as my **estimate of the 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 use different notation to refer to them. For instance, if the 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 (Table 8.2):

Table 8.2: Notation for the mean

Symbol	**What is it?**	**Do we know what it is?**
$\hat{X}$	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 in simple random samples

8.4.2 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:

$$20$$

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 simulate the results of some experiments. With that in mind, let's return to our IQ studies. Suppose the true population mean IQ is 100 and the standard deviation is 15. First I'll conduct an experiment in which I measure $N = 2$ IQ scores and I'll 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 8.10. 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 8.8 (b) when we plotted the sampling distribution of the mean, where the population mean is 100 and the average of the sample means is also 100.

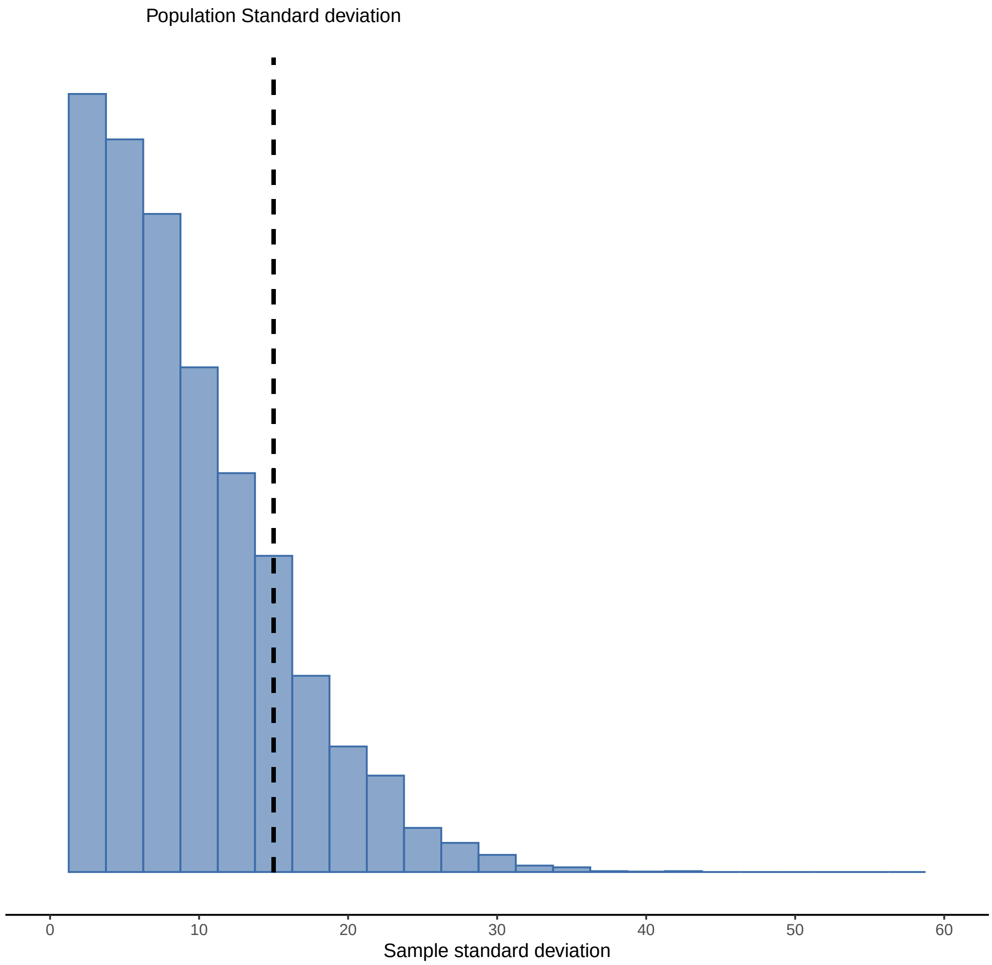


Figure 8.10: The sampling distribution of the sample standard deviation for a "two IQ scores" experiment. The true population standard deviation is 15 (dashed line), but as you can see from the histogram the vast majority of experiments will produce a much smaller sample standard deviation than this. On average, this experiment would produce an estimated standard deviation of only 8.4, well below the true value! In other words, the sample standard deviation is a biased estimate of the population standard deviation

Now let's extend the simulation. Instead of restricting ourselves to the situation where $N = 2$, 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 8.11. 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.[6] 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 $\hat{\sigma}$ we should make sure our guess is a little bit larger than the sample standard deviation s.

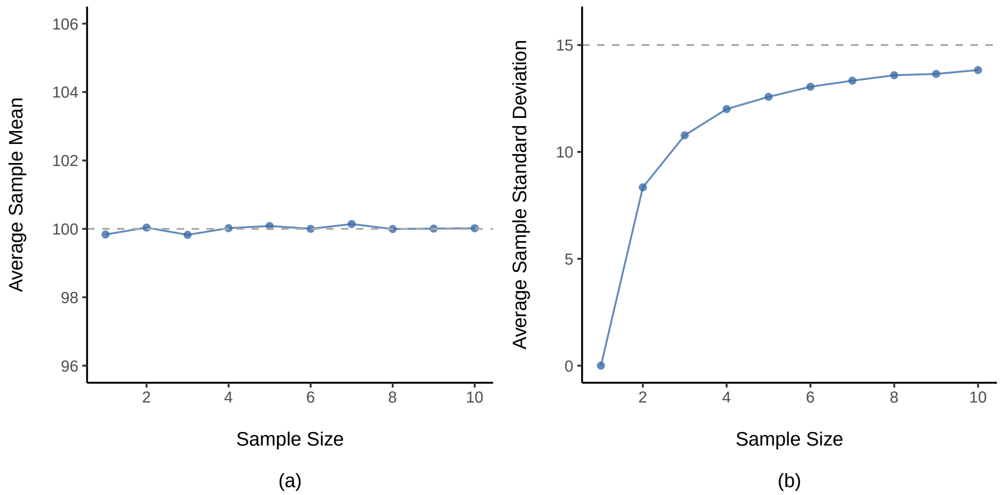


Figure 8.11: An illustration of the fact that the sample mean is an unbiased estimator of the population mean (panel a), but the sample standard deviation is a biased estimator of the population standard deviation (panel b). For the figure I generated 10,000 simulated data sets with 1 observation each, 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

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 the section on Estimating population parameters, the sample variance is defined to be the average

[6] I 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.

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.

This is an unbiased estimator of the population variance σ. Moreover, this finally answers the question we raised in Estimating population parameters. Why did jamovi give us slightly different answers for variance? It's because jamovi 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 is unbiased, and when we use jamovi's built in standard deviation function, what it's doing is calculating $\hat{\sigma}$ not s.[7]

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 just 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 (Table 8.3 and Table 8.4).

[7] Dividing by $N-1$ gives us an unbiased estimate of the population variance:

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

and similarly for standard deviation:

$$\hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2}$$

Okay, I'm hiding something else here. In a bizarre and counter-intuitive 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.

Table 8.3: Notation for standard deviation

Symbol	What is it?	Do we know what it is?
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

Table 8.4: Notation for variance

Symbol	What is it?	Do we know what it is?
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

8.5 Estimating a confidence interval

> "*Statistics means never having to say you're certain*"
> – Unknown origin [8]

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 participants is $\bar{X}$. We

[8] This quote appears on a great many t-shirts and websites, and even gets a mention in a few academic papers (e.g., https://jse.amstat.org/v10n3/friedman.html), but I've never found the original source.

know from our discussion of The central limit theorem that the sampling distribution of the mean is approximately normal. We also know from our discussion of the normal distribution in Section 7.5 that there is a 95% chance that a normally-distributed quantity will fall within about two standard deviations of the true mean.

To be more precise, the more correct answer is that there is a 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 $\bar{X}$ that we have actually observed lies within 1.96 standard errors of the population mean.

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 would have used 1.04 as the magic number rather than 1.96.

[Additional technical detail[9]]

8.5.1 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 the the section What does probability mean?, you'll notice that talking about personal belief and confidence is a Bayesian idea. 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

[9] Mathematically, we write this as:

$$\mu - (1.96 \times SEM) \leq \bar{X} \leq \mu + (1.96 \times SEM)$$

where the SEM is equal to $\frac{\sigma}{\sqrt{N}}$ 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:

$$\bar{X} - (1.96 \times SEM) \leq \mu \leq \bar{X} + (1.96 \times SEM)$$

What this is telling is 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:

$$CI_{95} = \bar{X} \pm (1.96 \times \frac{\sigma}{\sqrt{N}})$$

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 8.12, 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). We'd expect that around 95 of our confidence intervals would contain the true population mean, and that's what we found in Figure 8.12. 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 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 **?@sec-Bayesian-statistics**.

8.5.2 Calculating confidence intervals in jamovi

jamovi includes a simple way to calculate confidence intervals for the mean as part of the 'Descriptives' functionality. Under 'Descriptives'-'Statistics' there is a check box for both the 'Std. error of Mean' and 'Confidence interval for the mean', so you can use this to find out the 95% confidence interval (which is the default). So, for example, if I load the *IQsim.omv* file, check 'Confidence interval for the mean', I can see the confidence interval associated with the simulated mean IQ: Lower 95% CI = 99.39 and Upper 95% CI = 99.97 So, in our simulated large sample data with N = 10,000, the mean IQ score is 99.68 with a 95% CI from 99.39 to 99.97.

When it comes to plotting confidence intervals in jamovi, you can specify that the mean is included as an option in a box plot. Moreover, when we get onto learning about specific statistical tests, for example in **?@sec-Comparing-several-means-one-way-ANOVA**, we will see that we can also plot confidence intervals as part of the data analysis. That's pretty cool, so we'll show you how to do that later on.

8.6 Summary

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

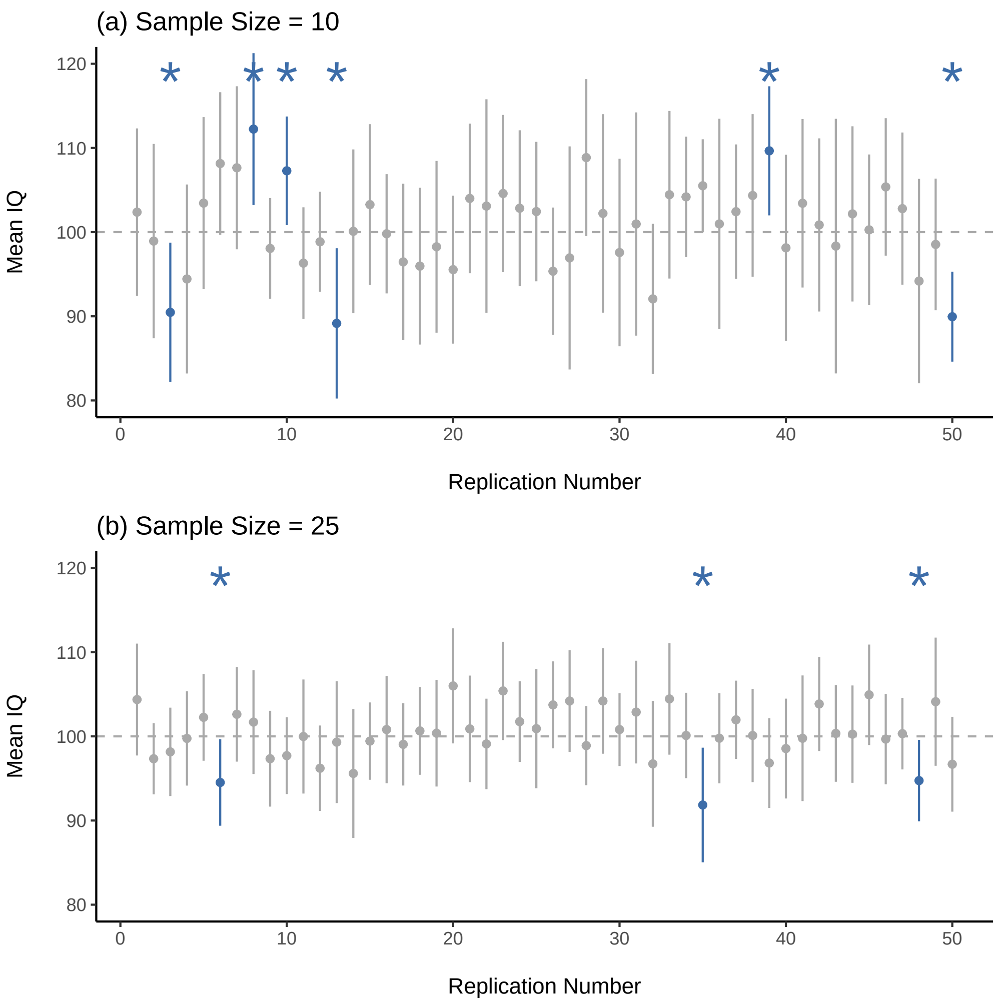


Figure 8.12: 95% confidence intervals. The top panel (a) shows 50 simulated replications of an experiment in which we measure the IQs of 10 people. The dot marks the location of the sample mean and the line shows the 95% confidence interval. Most of the 50 confidence intervals do contain the true mean (i.e., 100), but a few – in blue and 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

construct estimates of the population parameters. The section breakdown looks like this:

- Basic ideas about Samples, populations and sampling.
- Statistical theory of sampling: The law of large numbers and Sampling distributions and the central limit theorem.
- Estimating population parameters. Means and standard deviations.
- Estimating a confidence interval.

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.

Adair, G. (1984). The hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, *69*, 334–345.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, *187*, 398–404.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, *11*, 295–306.

Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions (3rd ed)*. Wiley.

Fisher, R. A. (1922). On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society A*, *222*, 309–368.

Hróbjartsson, A., & Gøtzsche, P. (2010). Placebo interventions for all clinical conditions. *Cochrane Database of Systematic Reviews*, *1*. https://doi.org//10.1002/14651858.CD003974.pub3

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), 697–701.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Keynes, J. M. (1923). *A tract on monetary reform*. Macmillan; Company.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *Public Library of Science One*, *9*, 1–8.

Meehl, P. H. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Pfungst, O. (1911). *Clever hans (the horse of mr. Von osten): A contribution to experimental animal and human psychology* (C. L. Rahn, Trans.). Henry Holt.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. Appleton.

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.

Wilkinson, L., Wills, D., Rope, D., Norton, A., & Dubbs, R. (2006). *The grammar of graphics*. Springer.

jamovi Stats. Open. Now.

This textbook covers the contents of an introductory statistics class, as typically taught to undergraduate psychology, health or social science students. The book covers how to get started in jamovi as well as giving an introduction to data manipulation. From a statistical perspective, the book discusses descriptive statistics and graphing first, followed by chapters on probability theory, sampling and estimation, and null hypothesis testing. After introducing the theory, the book covers the analysis of contingency tables, correlation, *t*-tests, regression, ANOVA and factor analysis. Bayesian statistics are touched on at the end of the book.

Citation: Navarro DJ and Foxcroft DR (2022). learning statistics with jamovi: a tutorial for psychology students and other beginners. (Version 0.75). [DOI: 10.24384/hgc3-7p15](https://dx.doi.org/10.24384/hgc3-7p15)