Answering questions with data

Mallory Barnes

2023-10-09

This comprehensive resource offers a free, accessible textbook for environmental science students embarking on introductory statistics. The package includes a practical lab manual and a dedicated course website, all provided under a CC BY-SA 4.0 license.

Table of contents

Preface

First Draft (version 0.0 = August 14th, 2023)

Welcome to the first edition of this Open Educational Resource (OER) textbook, specifically adapted to the needs of Environmental Science students enrolled in the SPEA E-538 statistics course at Indiana University (IU).

This textbook is an adaptation of a thorough introductory statistics textbook originally developed for undergraduate Psychology students by Matthew Crump and colleagues (refer to Acknowledgements for more details). As part of IU's Course Materials Fellowship Program (CMFP), I've had the opportunity to mold this material, refining it to serve as a specialized resource for students studying Environmental Science.

Online Textbook:https://malloryb.github.io/statistics_E538/

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

All resources are released under a creative commons licence CC BY-SA 4.0. Click the link to read more about the license, or read more below in the license section.

Acknowledgements

I wish to express my deepest appreciation to the contributors of the original textbook, without whom this adaptation would not have been possible. I am deeply grateful for the expertise and vision of Matthew Crump, Alla Chavarga, Anjali Krishnan, Jeffrey Suzuki, and Stephen Volz. Their exceptional groundwork laid the foundation for this project.

My heartfelt thanks also go to the Course Materials Fellowship Program (CMFP) at Indiana University (IU), which has been instrumental in supporting this adaptation effort. The CMFP is an initiative designed to incentivize the discovery, implementation, and creation of cost-effective course materials. Its aim is to foster the use of Open Educational Resources (OERs)—freely accessible and customizable learning materials that make education more equitable and accessible.

I am particularly grateful to Sarah Hare, the Bloomington lead for the CMFP program, and Adam Mazel, a digital publishing librarian at IU. Their guidance, expertise, and steadfast

support have been crucial to this project's success. Their contributions to the advancement of affordable and accessible education are truly commendable.

CC BY-SA 4.0 license

This license means that you are free to:

- Share: copy and redistribute the material in any medium or format
- Adapt: remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

- Attribution: You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- ShareAlike: If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
- No additional restrictions: You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Copying the textbook

This textbook was written in R-Studio, using R Markdown, and compiled into a web-book format using the bookdown package. In general, I thank the larger R community for all of the amazing tools they made, and for making those tools open, so that I could use them to make this thing.

All of the source code for compiling the book is available from the GitHub repository for the original textbook:

https://github.com/CrumpLab/statistics

and my github repository:

https://github.com/malloryb/statistics_E538

In principle, anybody could fork or otherwise download the E-538 textbook, or the original textbook, which is what I did. Load the Rproj file in R-studio, and then compile the entire book. Then, the individual .rmd files for each chapter could be edited for content and style to better suit your needs.

If you want to contribute to this version of the textbook, you could make pull requests on GitHub, or discuss issues and request on the issues tab.

The vision behind this textbook

The goal of this textbook is twofold. First, it aims to make complex statistical concepts accessible and digestible for students in Environmental Science. Second, it hopes to promote the usage of open-source tools like R, recognizing their value in today's data-driven world.

1 Why Statistics?

Adapted to environmental science by Mallory Barnes. Portions adapted nearly verbatim from Chapters 1 and 2 in Navarro, D. J. "Learning Statistics with R." https://compcogscisydney.org/learning-statistics-with-r/

> To call in statisticians after the experiment is done may be no more than asking them to perform a post-mortem examination: They may be able to say what the experiment died of. —Sir Ronald Fisher

1.1 On the Importance of Statistics in Environmental Science

Many students may find it surprising that statistics plays a substantial role in the study of environmental science. It is an essential tool for analyzing complex ecological data, predicting environmental trends, and making informed decisions about natural resource management. Despite its significance, statistics may not be the favorite part of every student's environmental science education. After all, if statistics were the primary interest, students might find themselves in a dedicated statistics course rather than an environmental science one.

However, the reality is that statistics is integral to understanding various environmental processes and phenomena. There's likely a segment of the student population that might be initially reluctant to embrace the statistical component of the subject. Recognizing this, it's valuable to start by addressing some common questions and concerns about the role of statistics in environmental science, aiming to demonstrate its relevance and importance in today's world. This understanding can help students approach statistical concepts with more confidence and see them as essential tools in their scientific toolkit.

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 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?*

You might think this question is a bit simple, but sometimes the best questions are the ones that seem the most obvious. So why do scientists crunch numbers instead of just using their gut feeling?

Imagine you're trying to figure out if a new factory is harming the fish in a nearby river. You could take a quick look, see some sick fish, and think, "Yep, that factory is the problem." But what if there's more to the story?

That's where statistics come into play. Statistics is like a detective's toolkit for scientists. It helps them sift through the clues and find the real cause. Without statistics, scientists might jump to conclusions. They might miss something important. Or, they might be swayed by their own opinions or what's easiest to believe.

Common sense is a bit like trying to solve a jigsaw puzzle in the dark. You might get some pieces right, but you're likely to miss the whole picture. Statistics turns on the lights. It helps scientists see things they might overlook and steer clear of pitfalls and biases that can trip them up.

Sure, our gut feelings can be handy in daily life. But when it comes to understanding the complexity of nature and the environment, we need more than just a hunch. The world is changing, and our instincts aren't always up to speed.

Statistics gives scientists the tools they need to make wise guesses about the world. It helps them go beyond what they can see or feel to uncover truths about how nature works. And in a world where we're trying to tackle big problems like climate change and pollution, having the right tools is crucial.

So next time you wonder why environmental scientists are so keen on numbers and charts, remember: it's not because they don't trust themselves. It's because they want to get to the heart of the matter, without letting anything get in the way.

1.1.1 The curse of belief bias

People are mostly pretty smart. We're certainly 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., 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:

- No cigarettes are inexpensive (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:

- No addictive things are inexpensive (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:

- No addictive things are inexpensive (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:

- No cigarettes are inexpensive (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 like this:

	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, Evans, Barston, and Pollard (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:

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

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:

	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!)

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 judgement 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 if we want to "believe in the data" instead, 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 that the gender breakdown of their admissions, which looked like this:

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

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

Extra

Earlier versions of these notes incorrectly suggested that they actually were sued – apparently that's not true. There's a nice commentary on this here: https://www.refsmmat.com/posts/2016-05-08-simpsons-paradox-berkeley.html.

When people started looking more carefully at the admissions data (Bickel, Hammel, and O'Connell 1975) they told a rather different story. Specifically, when they looked at it on a department by department basis, it turned out that most of the departments actually had a slightly *higher* success rate for female applicants than for male applicants. The table below shows the admission figures for the six largest departments (with the names of the departments removed for privacy reasons):

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., engineering, chemistry) tended to admit a high percentage of the qualified applicants, whereas others (e.g., English) tended to reject most of the candidates, even if they were high quality. So, among the six departments shown above, notice that department A is the most generous, followed by B, C, D, E and F in that order. Next, notice that males and females tended to apply to different departments. If we rank the departments in terms of the total number of male applicants, we get **A**>**B**>D>C>F>E (the "easy" departments are in bold). On the whole, males tended to apply to the departments that had high admission rates. Now compare this to how the female applicants distributed themselves. Ranking the departments in terms of the total number of female applicants produces a quite different ordering C>E>D>F>**A**>**B**. In other words, what these data seem to be suggesting is that the female applicants tended to apply to "harder" departments. And in fact, if we look at Figure **??** 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 behavior of all the departments, it turned out that the actual departments were, if anything, slightly biased in favor of women. The gender bias in total admissions was caused by the fact that women tended to self-select for harder departments. From a legal perspective, that would probably put the university in the clear. Postgraduate admissions are determined at the level of the individual department (and there are good reasons to do that), and at the level of individual departments, the decisions are more or less unbiased (the weak bias in favor 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 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

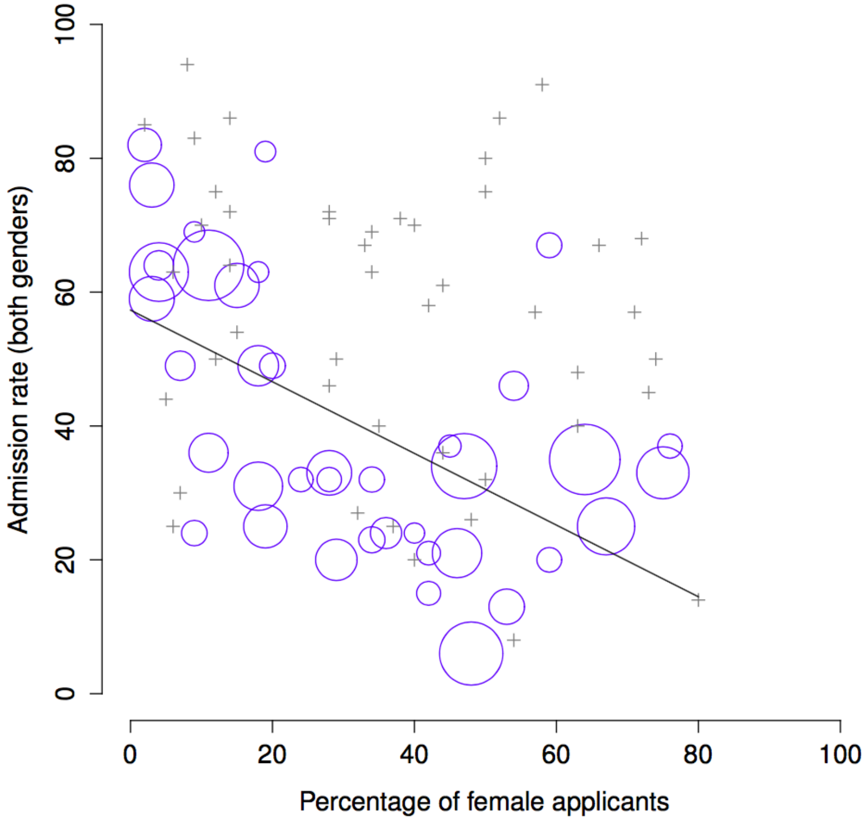


Figure 1.1: The Berkeley 1973 college admissions data. This figure plots the admission rate for the 85 departments that had at least one female applicant, as a function of the percentage of applicants that were female. The plot is a redrawing of Figure 1 from Bickel et al. (1975). Circles plot departments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

that the reason for why the humanities departments have low admission rates is because the government doesn't want to fund the humanities (spots in Ph.D. programs, for instance, are often tied to government funded research projects). Does that constitute a gender bias? Or just an unenlightened view of the value of the humanities? What if someone at a high level in the government cut the humanities funds because they felt that the humanities are "useless chick stuff". That seems pretty *blatantly* gender biased. None of this falls within the purview of statistics, but it matters to the research project. If you're interested in the overall structural effects of subtle gender biases, then you probably want to look at *both* the aggregated and disaggregated data. If you're interested in the decision making process at Berkeley itself then you're probably only interested in the disaggregated data.

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

1.3 Statistics in Environmental Science: Why All the Numbers?

You may wonder why your environmental science class is suddenly filled with graphs, equations, and statistical terms. Why does environmental science have so much statistics? Let me break it down for you!

Why Does Environmental Science Need Statistics?

It's a bit of a complex story, but let's face it: nature is complex too! Unlike physics, where you might be studying straightforward things like electrons, in environmental science, we're dealing with living ecosystems, unpredictable weather, and human behavior.

Think of nature as a giant puzzle with pieces constantly changing shape. We use statistics to piece it together. We're not dealing with simple objects; we're dealing with the flow of rivers, the growth of forests, and the migration of animals. It's a whole different ballgame.

In a sense, environmental scientists have to be part detective, part mathematician. We need to understand statistics because we're solving harder problems than some other fields. We're trying to make sense of the messy, beautiful, and intricate web of life.

Can't Someone Else Handle the Statistics?

Sure, you could ask someone else to do the math, but here's the catch: understanding statistics is crucial for understanding environmental problems and solutions.

1. **The Design Connection:** Want to research how pollution affects fish? You'll need to understand the statistics behind your study. Statistics and research design are like two sides of the same coin.

2. **Understanding the Science:** Want to read up on climate change? You'll find a lot of statistics in those papers. To truly get what's going on, you need to know what those numbers are saying.
3. **The Money Matter:** Hiring a statistician for every project? That'll cost you. Being self-sufficient in statistics means you can tackle more projects with fewer resources.

But What if I Don't Care About Jobs, Research, or Conservation? Why Do I Need Statistics?

Well, first of all, you've got me curious! But even if you're just an everyday citizen, statistics matter. You're surrounded by data, whether it's weather forecasts, pollution reports, or wildlife trends.

Knowing a bit about statistics is like having a decoder ring for the modern world. It helps you understand the news, make informed decisions, and even argue a point more convincingly.

So yes, statistics might feel like a strange guest in your environmental science class. But trust me, it's a guest you want to get to know. Because understanding statistics isn't just about crunching numbers; it's about understanding our world and how we can make it better.

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; 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. Perhaps, 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 :).

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.

1.6 A brief introduction to research design

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 and Stanley (1963) for the discussion of study design, and Stevens (1946) for the discussion of scales of measurement. Later versions will attempt to be more precise in the citations.

1.7 Introduction to measurements

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 behavior or the human mind. What do I mean by "measurement"?

1.7.1 Some thoughts about measurements

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 measurement:

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

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

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

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

1.7.2 Operationalization: defining your measurement

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

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

Operationalization is a tricky business, and there's no "one, true way" to do it. The way in which you choose to operationalize 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, operationalization needs to be thought through on a case by case basis. Nevertheless, while there 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 behavioral observation or a brain scan could all count as a measure.

- **An operationalization**. The term “operationalization” refers to the logical connection between the measure and the theoretical construct, or to the process by which we try to derive a measure from a theoretical construct.
- **A variable**. Finally, a new term. A variable is what we end up with when we apply our measure to something in the world. That is, variables are the actual “data” that we end up with in our data sets.

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

1.8 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 it’s very 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**.

1.8.1 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 color”. Eyes can be blue, green and brown, among other possibilities, but none of them is any “better” than any other one. As a result, it would feel really weird to talk about an “average eye color”. 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”, etc. For now, let’s suppose that these four are the only possibilities, and suppose that when I ask 100 people how they got to work today, and I get this:

Transportation	Number of people
(1) Train	12
(2) Bus	30

Transportation	Number of people
(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 this and nothing really changes.

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

1.8.2 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 *don't* 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 an more psychologically interesting example. Suppose I'm interested in people's attitudes to climate change, and I ask them to pick one of these four 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 the 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…

(3) Temperatures are rising, but not because of humans

(4) Temperatures are rising, because of human activity

(5) Temperatures are not rising

(6) 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 following answers:

	Number
(1) Temperatures are rising, because of human activity	51
(2) Temperatures are rising, but we don't know why	20
(3) Temperatures are rising, but not because of humans	10
(4) Temperatures are not rising	19

When analyzing these data, it seems quite reasonable to try to group (1), (2) and (3) together, and say that 81 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 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 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 that sounds like gibberish to me!

1.8.3 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.

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 insane 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.

1.8.4 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".

1.8.5 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. The difference between these is as follows:

- 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.

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 it's possible for Cameron's response time to lie in between, by taking 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 say that RT is continuous.

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. The table summarizes 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.

Table 1.9: The relationship between the scales of measurement and the discrete/continuity distinction. Cells with an x correspond to things that are possible.

	continuous	discrete
nominal		x
ordinal		x
interval	x	x
ratio	x	x

1.8.6 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: 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 be 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 if it were interval scale. It's not interval scale, but in practice it's close enough that we usually think of it as being **quasi-interval scale**.

1.9 Assessing the reliability of a measurement

At this point we've thought a little bit about how to operationalize 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.

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 mom" is very unreliable: some days she tells me I'm a bit thick, and other days she tells me I'm a complete moron. 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 potatoes when I step on and off of 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, while my mom'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. So that would be an *unreliable but valid* measure. Of course, to some extent, notice that 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 a 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.

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, *Inroduction to Environmental Science*, has an assessment structure that has a presentation component and an exam component (plus other things). The exam component is *intended* to measure something different from the presentation 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 measure the same thing!

1.10 The role of variables: predictors and outcomes

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

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

For example, in an experiment the IV refers to the **manipulation**, and the DV refers to the **measurement**. So, we could use **manipulated variable** (independent variable) and **measured variable** (dependent variable).

Table 1.10: The terminology used to distinguish between different roles that a variable can play when analyzing a data set.

role of the variable	classical name	modern name
"to be explained"	dependent variable (DV)	Measurement
"to do the explaining"	independent variable (IV)	Manipulation

We could also use **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). This is summarized in the table:

Table 1.11: The terminology used to distinguish between different roles that a variable can play when analyzing a data set.

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

1.11 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.

1.11.1 Experimental research

The key features of **experimental research** is that the researcher controls all aspects of the study, especially what participants experience during the study. In particular, the researcher manipulates or varies something (IVs), and then allows the outcome variable (DV) to vary naturally. The idea here is to deliberately vary the something in the world (IVs) to see if it has any causal effects on the outcomes. Moreover, in order to ensure that there's no chance that something other than the manipulated variable 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 **randomization**: 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 randomization later in this course, but for now, it's enough to say that what randomization does is minimize (but not eliminate) the chances that there are any systematic difference between groups.

Let's consider a very simple, completely unrealistic and grossly unethical example. Suppose you wanted to find out if smoking causes lung cancer. One way to do this would be to find people who smoke and people who don't smoke, and look to see if smokers have a higher rate of lung cancer. This is *not* a proper experiment, since the researcher doesn't have a lot of control over who is and isn't a smoker. And this really matters: for instance, it might be that people who choose to smoke cigarettes also tend to have poor diets, or maybe they tend to work in asbestos mines, or whatever. The point here is that the groups (smokers and non-smokers) actually differ on lots of things, not *just* smoking. So it might be that the higher incidence of lung cancer among smokers is caused by something else, not by smoking per se. In technical terms, these other things (e.g. diet) are called "confounds", and we'll talk about those in just a moment.

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

1.11.2 Non-experimental research

Non-experimental research in environmental science often encompasses studies in which researchers cannot exert full control over variables as they might in a typical experimental setup. In environmental science, this issue is particularly pronounced, as there is no "control planet" to allow for perfect experimental conditions. This unique challenge leads to a reliance on alternative approaches, such as time series analysis, to understand complex environmental phenomena.

One distinction worth making between two types of non-experimental research is the difference between **quasi-experimental research** and **case studies**. Much of the research in environmental science fits into a quasi-experimental design. Here, researchers might wish to study the effects of industrial pollution on a river system but cannot directly control all the variables, such as the quantity or type of pollutants being emitted by various industries. Instead, they must observe existing conditions and make careful comparisons, possibly using statistical tools to account for confounding variables.

Time series analysis becomes vital in this context, as many environmental processes unfold over extended periods. For example, tracking changes in global temperatures or sea level requires years of consistent data collection. The complexity and variability inherent in environmental systems make drawing definitive conclusions difficult, but sophisticated statistical methods can help isolate specific effects.

Environmental science also frequently relies on detailed **case studies**. These investigations focus on particular events or locations to provide in-depth insights into environmental processes and their impacts on ecosystems, human health, and societal structures. Case studies might explore the aftermath of a natural disaster, the unique ecology of a threatened habitat, or the social and economic consequences of environmental policies.

Though they often lack the broad generalizability of more controlled experimental research, case studies in environmental science offer valuable opportunities to understand the complexity of real-world situations. They can uncover nuances and subtleties that might be overlooked in a broader experimental or quasi-experimental approach.

The absence of a control planet and the complexity of environmental systems necessitates flexible and often non-traditional approaches to scientific inquiry in environmental science. While these methods, including quasi-experimental designs and case studies, may not offer the same level of control as experimental designs, they are indispensable tools in the field. By leveraging time series data, employing careful statistical analysis, and embracing the rich insights offered by focused case studies, researchers can continue to expand our understanding of the biosphere and its interactions with human activities.

1.12 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, while 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's lots of different kinds of validity, each of which raises it's own issues, and not all forms of validity are relevant to all studies. I'm going to talk about five different types:

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

To give you 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.

1.12.1 Internal validity

Internal validity refers to the extent to which you are able 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… the big problem that you have 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.

1.12.2 External validity

External validity relates to the **generalizability** 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. So, if it turns out that the results don't actually generalize 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 psych 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 populace in some *relevant* respect, then we may start worrying about a lack of external validity.

That said, it is absolutely critical to realize 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 is threatened by the choice of population if (a) the population from which you sample your participants is very narrow (e.g., psych 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 italicized 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 psych 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's 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 very 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 the 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.

1.12.3 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 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 theater 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.

1.12.4 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 verbalize, so there might be something to worry about even if you can't quite say why. In other words, when someone you trust criticizes 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 much.

- Often (very often), completely uninformed people will also have a "hunch" that your research is crap. And they'll criticize it on the internet or something. On close inspection, you'll often 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.

1.12.5 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 rigor 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. An simple example would be eyewitness identification studies. Most of these studies tend to be done in a university setting, often with 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 that the witness being scared, and there's 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.

1.13 Confounds, artifacts 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 *confounds* and *artifact*. These two terms are defined in the following way:

- **Confound**: A confound is an additional, often unmeasured variable that turns out to be related to both the predictors and the outcomes. The existence of confounds 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, etc.
- **Artifact**: A result is said to be "artifactual" if it only holds in the special situation that you happened to test in your study. The possibility that your result is an artifact describes a threat to your external validity, because it raises the possibility that you can't generalize your results to the actual population that you care about.

As a general rule confounds 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 confounds working their way into your study. Experimental research tends to be much less vulnerable to confounds: the more control you have over what happens during the study, the more you can prevent confounds from appearing.

However, there's always swings and roundabouts, and when we start thinking about artifacts rather than confounds, the shoe is very firmly on the other foot. For the most part, artifactual results tend to be a concern for experimental studies than for non-experimental studies. To see this, it helps to realize that the reason that a lot of studies are non-experimental is precisely because what the researcher is trying to do is examine human behavior in a more naturalistic context. By working in a more real-world context, you lose experimental control (making yourself vulnerable to confounds) but because you tend to be studying human psychology "in the wild" you reduce the chances of getting an artifactual 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 than you wanted to study: which is more or less the definition of an artifact.

Be warned though: the above is a rough guide only. It's absolutely possible to have confounds in an experiment, and to get artifactual results with non-experimental studies. This can happen for all sorts of reasons, not least of which is researcher error. In practice, it's really hard to think everything through ahead of time, and even very good researchers make mistakes. But other times it's unavoidable, simply because the researcher has ethics (e.g., see "differential attrition").

Okay. There's a sense in which almost any threat to validity can be characterized as a confound or an artifact: they're pretty vague concepts. So let's have a look at some of the most common examples…

1.13.1 History effects

History effects refer to the possibility that specific events may occur during the study itself that might influence the outcomes. For instance, something might happen in between a pre-test and a post-test. Or, in between testing participant 23 and participant 24. Alternatively, it might be that you're looking at an older study, which 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:

- 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, let's say), then you administer the drug, and then you take the same

measures afterwards. In the middle, however, because your labs are in Los Angeles, there's an earthquake, which increases the anxiety of the participants.

1.13.2 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:

- 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 goes 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

1.13.3 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; 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 at measurement at time 2 is more likely to become "bored", precisely because of the boring measurement made at time 1.

1.13.4 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.

1.13.5 Differential attrition

within the study itself, impacting both the internal and external validity of the research. This phenomenon can occur in environmental studies, where participants' commitment to certain practices might change over time, such as in multi-year studies on cover cropping.

When thinking about the effects of differential 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. Homogeneous attrition occurs when the dropout rate is roughly the same across different groups or conditions within the study. For instance, if many farmers participating in a cover cropping study discontinue the practice at similar rates, the sample may become unrepresentative. Though the internal validity (accuracy of the study's conclusions) might remain intact, the external validity (generalizability to the broader population) may suffer. In the realm of environmental science, this could mean the study's findings are less applicable to the broader community of farmers or agricultural systems.

The second type of differential attrition is **heterogeneous attrition**, in which the attrition effect is different for different groups. This is a much bigger problem: not only do you have to worry about your external validity, you also have to worry about your internal validity too. In a study on cover cropping, if certain groups, perhaps motivated by specific incentives or pressures, drop out at different rates, this can introduce a confounding variable. The comparison between the remaining participants may no longer be valid, as the groups have become fundamentally different due to the non-random dropout. This not only affects the generalizability of the findings but also the internal validity, as the study's conclusions within the sample itself may be compromised.

1.13.6 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, and 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 generalize 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.

1.13.7 Regression to the mean

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

Here's an example. Suppose we are interested in examining the effects of specific climatic conditions on the growth of a particular plant species. We identify the 20 locations with the most significant growth during a year of favorable weather conditions, such as ideal rainfall and temperature, and decide to study these areas further.

After another year, we observe that the plant growth in these 20 locations is still above average but not as exceptional as during the first year. The immediate reaction might be to conclude that a change in environmental practices or some other factor has adversely affected growth in these areas.

However, this might simply be an example of regression to the mean. Consider what is required for remarkable plant growth: optimal soil, hard work in maintenance, favorable weather conditions, and perhaps a bit of luck with other uncontrollable environmental factors.

While the soil quality and maintenance practices may be consistent from one year to the next, luck with weather is not. Certain areas that were fortunate with ideal weather conditions in the first year may not experience the same fortune in the following year. This inconsistency in luck leads to a less extreme measurement the second time.

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 artifact caused by regression to the mean (Kahneman and Tversky 1973)

1.13.8 Experimenter bias

Experimenter bias can occur when the researcher, despite conscientious efforts, inadvertently influences the results of a study. This may not involve communication with human participants as in other fields but could still manifest in several ways within environmental research.

In environmental science, the experimenter might have **expectations or preconceived notions about what the data should reveal**. For instance, a researcher studying the effects of a certain agricultural practice on soil health may unconsciously select or emphasize data that align with their expectations or theoretical commitments. This can lead to a bias in data collection or interpretation, subtly steering the results toward the expected outcome.

Experimenter bias can also arise from **methodological choices**. A researcher may design an experiment, select sites for observation, or choose measurement techniques in a way that favors a particular outcome. For example, if studying the effect of pollution on a water source, the researcher might unintentionally choose sampling times or locations that are more likely to confirm their hypotheses, overlooking variations that could provide a more comprehensive and unbiased view.

In some cases, the researcher's bias may inadvertently **influence other field or laboratory personnel** involved in the study. A team member who understands the desired outcome of the study may unconsciously alter procedures, handling of samples, or recording of data to align with the researcher's expectations.

The classic example of experimenter bias is the case study of "Clever Hans", which dates back to 1907, (Pfungst 1911; Hothersall 2004). 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 behavior 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 they did know how to count, and the horse had learned to change its behavior 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 behavior is. This provides a very good solution to the problem, but it's important to recognize that it's not quite ideal, and typically not feasible in Environmental Science. Measures to mitigate this bias include rigorous peer review, careful documentation of methods, and maintaining an awareness of potential biases.

1.13.9 Reactivity and Demand Effects

Even in environmental studies, the knowledge that a process or phenomenon is being observed can impact the behavior of the researchers or those involved in data collection. 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 the "Hawthorne Works" factory outside of Chicago (Adair 1984). A study done in the 1920s looking at the effects of lighting on worker productivity at the factory turned out to be an effect of the fact that the workers knew they were being studied, rather than the lighting. This so-called "Hawthorne effect" might influence how a field team conducts measurements if they know the specific goal of the study.

1.13.10 Placebo effects

While typically associated with clinical trials, the **placebo effect** can have an analogous impact in environmental research. For instance, a belief in the efficacy of a certain conservation practice might lead to subjective reporting or unintentional data manipulation.

1.13.11 Situation, measurement and subpopulation effects

In some respects, these terms are a catch-all term for "all other threats to external validity". The choice of location, time, measurement tools, and even who collects the data can all influence the results. Ensuring robust methodologies and considering the potential influences of these factors can help in achieving results that generalize more widely.

1.13.12 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

One final thing that I feel like I should mention. While reading what the textbooks often have to say about assessing the validity of the 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. 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 are:

- **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) some of 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, it's 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 these notes. 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 (artifactually) 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 etc. 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 has been 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 & post hoc hypothesizing**. Another way in which the authors of a study can more or less lie about what they found is by engaging in what's referred to as "data mining". As we'll discuss later in the class, 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 run 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 reanalyze your data using your new "post hoc" hypotheses. What is wrong (and, I suspect, common) is failing to acknowledge that you've done so. If you acknowledge that you did it, then other researchers are able to take your behavior into account. If you don't, then they can't. And that makes your behavior deceptive. Bad!

- **Publication bias & 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. 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 "internalize" 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, Fritz, and Scherndl (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 assumes 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.

1.14 Summary

In this chapter, we have explored essential aspects of research methodology pertinent to environmental statistics:

- **Operationalization and Measurement**: Understanding how to define and measure theoretical constructs and recognizing the distinctions between variables are foundational to research.
- **Scales of measurement and types of variables**: This section addressed the differences between discrete and continuous data, and the various scale types including nominal, ordinal, interval, and ratio scales.
- **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.
- **Terminology: 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**: Clarification of the roles that variables play in an analysis, including the distinctions between predictors and outcomes, and dependent and independent variables.

- **Research Designs**: Highlighting what constitutes an experiment within environmental research, the chapter delves into the differentiation between experimental and non-experimental research designs.
- **Validity and Threats**: A comprehensive look at whether a study measures what it aims to, understanding potential pitfalls, and recognizing the myriad ways things can go awry.

Study design is a paramount component of research methodology, with numerous textbooks and resources available to further explore these concepts. Drawing on established texts like Campbell and Stanley (1963), this chapter serves as an introduction and guide to these key topics, recognizing the intrinsic connection between statistics and study design within the field of environmental science.

1.15 Videos

1.15.1 Terms of Statistics

2 Describing Data

> Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise. —John W. Tukey

This chapter is about **descriptive statistics**. These are tools for describing data. Some things to keep in mind as we go along are:

1. There are lots of different ways to describe data
2. There is more than one "correct" way, and you get to choose the most "useful" way for the data that you are describing
3. It is possible to invent new ways of describing data, all of the ways we discuss were previously invented by other people, and they are commonly used because they are useful.
4. Describing data is necessary because there is usually too much of it, so it doesn't make any sense by itself.

2.1 This is what too many numbers looks like

Let's say you wanted to know how happy people are. So, you ask thousands of people on the street how happy they are. You let them pick any number they want from negative infinity to positive infinity. Then you record all the numbers. Now what?

Well, how about you look at the numbers and see if that helps you determine anything about how happy people are. What could the numbers look like. Perhaps something like this:

-612	284	-558	179	-666	598	333	-398	24	389
-126	-740	-84	33	-15	148	1260	241	-188	686
-161	-128	-377	-425	97	-929	-221	692	412	-674
-75	690	126	-217	-428	143	176	809	586	-143
-444	489	648	-180	810	-516	1030	307	-360	207
-100	-314	61	219	690	-656	343	-49	650	-240
-827	392	695	632	353	510	-146	442	639	156
514	-143	1112	-1428	473	78	298	-89	-361	743
-69	585	629	87	-480	-472	542	-408	-281	-44
-24	772	773	-182	1333	-109	-190	901	55	385
140	-139	131	434	-525	411	992	31	416	-39
255	114	570	-932	229	38	413	-372	805	-154
537	131	930	-119	186	844	119	439	692	-9
85	-91	932	-218	464	416	506	147	355	261
-405	-674	-122	487	-259	-69	1086	-685	-273	364
-385	-102	291	85	526	-420	-78	-292	356	-528
-591	-503	213	385	1074	-893	-568	-145	-254	272
-83	-511	-97	1172	630	-166	301	453	549	192
-863	-217	-207	1243	-30	656	-88	-399	-594	-549
406	-621	426	249	-178	1156	-95	772	335	-412
483	564	269	-309	-975	-434	689	182	-87	111
-543	452	159	1315	1109	-497	449	-766	-404	491
-643	-190	49	11	-565	-199	705	-1231	3	-66
409	-433	-459	-378	-823	660	125	-431	15	-825
606	662	-798	262	1514	-97	-211	499	-209	-16
-441	-40	645	519	396	-412	510	28	-52	-371
-1151	187	-131	548	55	-159	-351	504	-50	-539
-159	623	-12	92	1865	-1045	-230	-333	30	739
-885	253	338	-387	156	-204	-119	-333	-1002	-161
124	-191	102	1100	-578	576	76	898	-244	-32
133	101	-373	-724	1231	-695	382	-104	872	90
1216	988	-539	-204	127	-170	359	357	-1459	548
-472	405	-658	252	641	-408	-665	-361	-230	975
199	-648	-290	-430	-81	532	228	774	58	383
-180	562	501	952	767	-598	-336	-724	68	-95
406	-63	-331	-969	411	624	-192	913	-328	-24
262	65	938	395	244	713	295	-196	937	-122
1140	230	102	76	318	597	866	560	-912	378
635	-227	187	-101	-304	-409	139	979	517	-660
-1084	107	523	225	-625	311	213	-111	189	-1070
-80	-437	-17	488	96	85	42	338	-559	938
-332	479	-51	498	-108	-911	208	413	-830	-519
117	-276	564	-268	731	414	-34	1063	356	99
-362	10	859	327	1352	297	237	-233	-410	-85
-696	-775	-187	-566	214	508	416	111	10	282
-388	476	-775	-309	-233	-394	-747	460	-9	155
415	-66	559	212	-1	-571	670	539	195	-493
-226	140	-325	-312	482	518	295	-93	-663	-42
743	-992	-231	-142	-71	-779	480	50	-279	-141
350	375	-280	569	-565	-277	303	198	-69	-260

Now, what are you going to with that big pile of numbers? Look at it all day long? When you deal with data, it will deal so many numbers to you that you will be overwhelmed by them. That is why we need ways to describe the data in a more manageable fashion.

The complete description of the data is always the data itself. **Descriptive statistics** and other tools for describing data go one step further to summarize aspects of the data. Summaries are a way to compress the important bits of a thing down to a useful and manageable tidbit. It's like telling your friends why they should watch a movie: you don't replay the entire movie for them, instead you hit the highlights. Summarizing the data is just like a movie preview, only for data.

2.2 Look at the data

We already tried one way of looking at the numbers, and it wasn't useful. Let's look at some other ways of looking at the numbers, using graphs.

2.2.1 Stop, time to plot!

Let's turn all of the numbers into dots, then show them in a graph. Note, when we do this, we have not yet summarized anything about the data. Instead, we just look at all of the data in a visual format, rather than looking at the numbers.

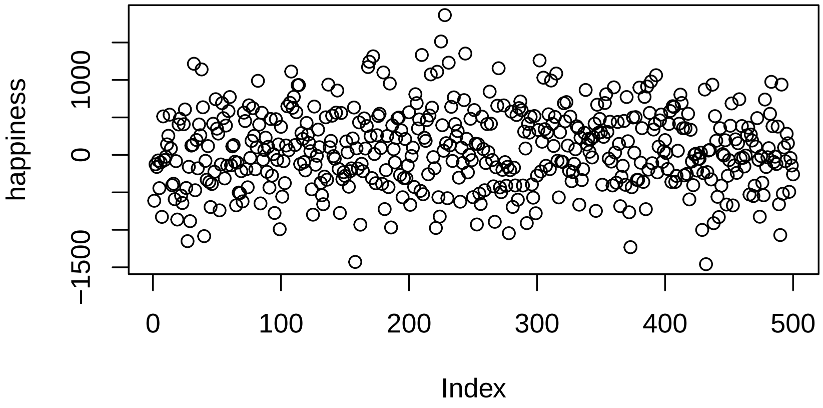


Figure 2.1: Pretend happiness ratings from 500 people

Figure **??** shows 500 measurements of happiness. The graph has two axes. The horizontal **x-axis**, going from left to right is labeled "Index". The vertical **y-axis**, going up and down, is labelled "happiness". Each dot represents one measurement of every person's happiness from our pretend study. Before we talk about what we can and cannot see about the data, it is worth mentioning that the way you plot the data will make some things easier to see and some things harder to see. So, what can we now see about the data?

There are lots of dots everywhere. It looks like there are 500 of them because the index goes to 500. It looks like some dots go as high as 1000-1500 and as low as -1500. It looks like there are more dots in the middle-ish area of the plot, sort of spread about 0.

> Take home: we can see all the numbers at once by putting them in a plot, and that is much easier and more helpful than looking at the raw numbers.

OK, so if these dots represent how happy 500 people are, what can we say about those people? First, the dots are kind of all over the place, so different people have different levels of happiness. Are there any trends? Are more people happy than unhappy, or vice-versa? It's hard to see that in the graph, so let's make a different one, called a **histogram.**

2.2.2 Histograms

Making a histogram will be our first act of officially summarizing something about the data. We will no longer look at the individual bits of data, instead we will see how the numbers group together. Let's look at Figure **??**, a histogram of the happiness data, and then explain it.

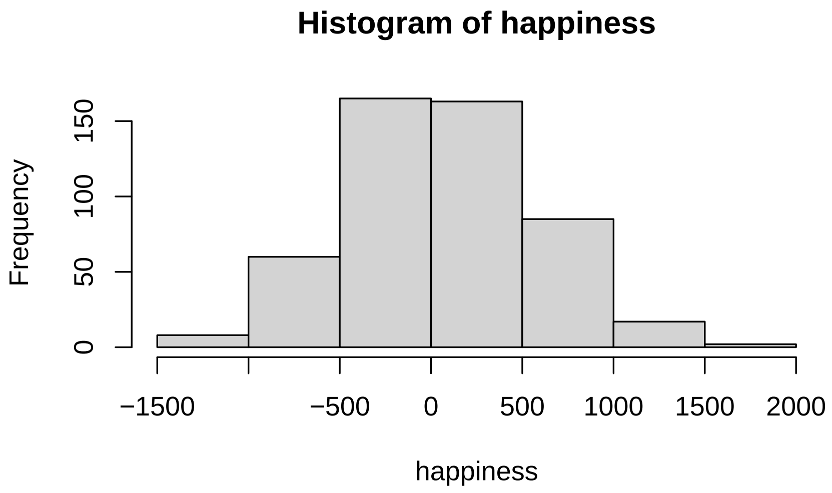


Figure 2.2: A histogram of the happiness ratings

The dots have disappeared, and now we some bars. Each bar is a summary of the dots, representing the number of dots (frequency count) inside a particular range of happiness, also called **bins**. For example, how many people gave a happiness rating between 0 and 500? The fifth bar, the one between 0 and 500 on the x-axis, tells you how many. Look how tall that bar is. How tall is it? The height is shown on the y-axis, which provides a frequency count (the number of dots or data points). It looks like around 150 people said their happiness was between 0-500.

More generally, we see there are many bins on the x-axis. We have divided the data into bins of 500. Bin #1 goes from -2000 to -1500, bin #2 goes from -1500 to -1000, and so on until the last bin. To make the histogram, we just count up the number of data points falling inside each bin, then plot those frequency counts as a function of the bins. Voila, a histogram.

What does the histogram help us see about the data? First, we can see the **shape** of data. The shape of the histogram refers to how it goes up and down. The shape tells us where the data is. For example, when the bars are low we know there isn't much data there. When the bars are high, we know there is more data there. So, where is most of the data? It looks like it's mostly in the middle two bins, between -500 and 500. We can also see the **range** of the data. This tells us the minimums and the maximums of the data. Most of the data is between -1500 and +1500, so no infinite sadness or infinite happiness in our data-set.

When you make a histogram you get to choose how wide each bar will be. For example, below are four different histograms of the very same happiness data. What changes is the width of the bins.

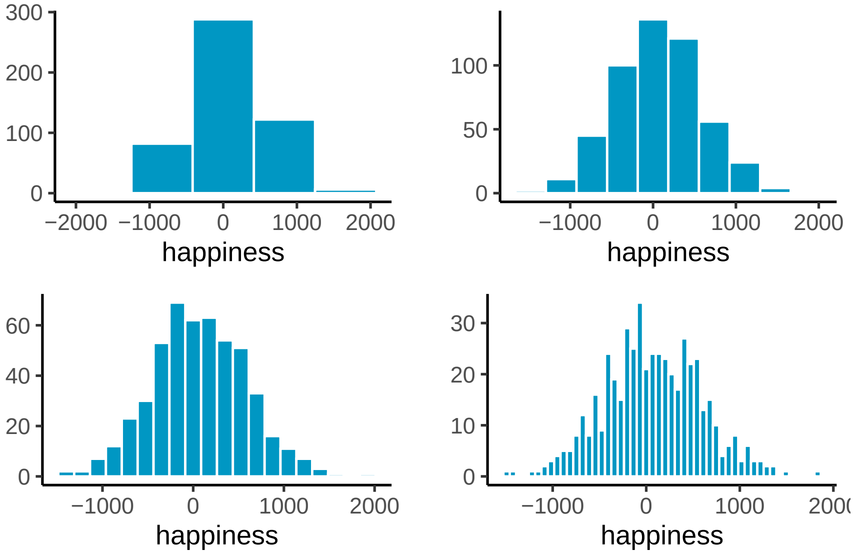


Figure 2.3: Four histograms of the same data using different bin widths

All of the histograms have roughly the same overall shape: From left to right, the bars start off small, then go up, then get small again. In other words, as the numbers get closer to zero, they start to occur more frequently. We see this general trend across all the histograms. But, some aspects of the trend fall apart when the bars get really narrow. For example, although the bars generally get taller when moving from -1000 to 0, there are some exceptions and the bars seem to fluctuate a little bit. When the bars are wider, there are less exceptions to the general trend. How wide or narrow should your histogram be? It's a Goldilocks question. Make it just right for your data.

2.3 Important Ideas: Distribution, Central Tendency, and Variance

Let's introduce three important terms we will use a lot, **distribution**, **central tendency**, and **variance**. These terms are similar to their everyday meanings (although I suspect most people don't say central tendency very often).

Distribution. When you order something from Amazon, where does it come from, and how does it get to your place? That stuff comes from one of Amazon's distribution centers. They distribute all sorts of things by spreading them around to your doorstep. "To Distribute"" is to spread something. Notice, the data in the histogram is distributed, or spread across the bins. We can also talk about a distribution as a noun. The histogram is a distribution of the frequency counts across the bins. Distributions are **very, very, very, very, very** important. They can have many different shapes. They can describe data, like in the histogram above. And as we will learn in later chapters, they can **produce** data. Many times we will be asking questions about where our data came from, and this usually means asking what kind of distribution could have created our data (more on that later.)

Central Tendency is all about sameness: What is common about some numbers? For example, is there anything similar about all of the numbers in the histogram? Yes, we can say that most of them are near 0. There is a tendency for most of the numbers to be centered near 0. Notice we are being cautious about our generalization about the numbers. We are not saying they are all 0. We are saying there is a tendency for many of them to be near zero. There are lots of ways to talk about the central tendency of some numbers. There can even be more than one kind of tendency. For example, if lots of the numbers were around -1000, and a similar large amount of numbers were grouped around 1000, we could say there was two tendencies.

Variance is all about different*ness*: What is different about some numbers?. For example, is there anything different about all of the numbers in the histogram? YES!!! The numbers are not all the same! When the numbers are not all the same, they must vary. So, the variance in the numbers refers to how the numbers are different. There are many ways to summarize the amount of variance in the numbers, and we discuss these very soon.

2.4 Measures of Central Tendency (Sameness)

We've seen that we can get a sense of data by plotting dots in a graph, and by making a histogram. These tools show us what the numbers look like, approximately how big and small they are, and how similar and different they are from another. It is good to get a feeling about the numbers in this way. But, these visual sensitudes are not very precise. In addition to summarizing numbers with graphs, we can summarize numbers using numbers (NO, please not more numbers, we promise numbers can be your friend).

2.4.1 From many numbers to one

Measures of central have one important summary goal: to reduce a pile of numbers to a single number that we can look at. We already know that looking at thousands of numbers is hopeless. Wouldn't it be nice if we could just look at one number instead? We think so. It turns out there are lots of ways to do this. Then, if your friend ever asks the frightening question, "hey, what are all these numbers like?". You can say they are like this one number right here.

But, just like in Indiana Jones and the Last Crusade (highly recommended movie), you must choose your measure of central tendency wisely.

2.4.2 Mode

The **mode** is the most frequently occurring number in your measurement. That is it. How do you find it? You have to count the number of times each number appears in your measure, then whichever one occurs the most, is the mode.

> Example: 1 1 1 2 3 4 5 6

The mode of the above set is 1, which occurs three times. Every other number only occurs once.

OK fine. What happens here:

> Example: 1 1 1 2 2 2 3 4 5 6

Hmm, now 1 and 2 both occur three times each. What do we do? We say there are two modes, and they are 1 and 2.

Why is the mode a measure of central tendency? Well, when we ask, "what are my numbers like", we can say, "most of the number are, like a 1 (or whatever the mode is)".

Is the mode a good measure of central tendency? That depends on your numbers. For example, consider these numbers

1 1 2 3 4 5 6 7 8 9

Here, the mode is 1 again, because there are two 1s, and all of the other numbers occur once. But, are most of the numbers like, a 1. No, they are mostly not 1s.

“Argh, so should I or should I not use the mode? I thought this class was supposed to tell me what to do?”. There is no telling you what to do. Every time you use a tool in statistics you have to think about what you are doing and justify why what you are doing makes sense. Sorry.

2.4.3 Median

The **median** is the exact middle of the data. After all, we are asking about central tendency, so why not go to the center of the data and see where we are. What do you mean middle of the data? Let’s look at these numbers:

1 5 4 3 6 7 9

Umm, OK. So, three is in the middle? Isn’t that kind of arbitrary. Yes. Before we can compute the median, we need to order the numbers from smallest to largest.

1 3 4 **5** 6 7 9

Now, 5 is in the middle. And, by middle we mean in the middle. There are three numbers to the left of 5, and three numbers to the right. So, five is definitely in the middle.

OK fine, but what happens when there aren’t an even number of numbers? Then the middle will be missing right? Let’s see:

1 2 3 4 5 6

There is no number between 3 and 4 in the data, the middle is empty. In this case, we compute the median by figuring out the number in between 3 and 4. So, the median would be 3.5.

Is the median a good measure of central tendency? Sure, it is often very useful. One property of the median is that it stays in the middle even when some of the other numbers get really weird. For example, consider these numbers:

1 2 3 4 4 4 **5** 6 6 6 7 7 1000

Most of these numbers are smallish, but the 1000 is a big old weird number, very different from the rest. The median is still 5, because it is in the middle of these ordered numbers. We can also see that five is pretty similar to most of the numbers (except for 1000). So, the median does a pretty good job of representing most of the numbers in the set, and it does so even if one or two of the numbers are very different from the others.

Finally, **outlier** is a term will we use to describe numbers that appear in data that are very different from the rest. 1000 is an outlier, because it lies way out there on the number line compared to the other numbers. What to do with outliers is another topic we discuss sometimes throughout this course.

2.4.4 Mean

Have you noticed this is a textbook about statistics that hasn't used a formula yet? That is about to change, but for those of you with formula anxiety, don't worry, we will do our best to explain them.

The **mean** is also called the average. And, we're guessing you might already now what the average of a bunch of numbers is? It's the sum of the numbers, divided by the number of number right? How do we express that idea in a formula? Just like this:

$Mean = \bar{X} = \frac{\sum_{i=1}^{n} x_i}{N}$

"That looks like Greek to me". Yup. The $\sum$ symbol is called **sigma**, and it stands for the operation of summing. The little "i" on the bottom, and the little "n" on the top refers to all of the numbers in the set, from the first number "i" to the last number "n". The letters are just arbitrary labels, called **variables** that we use for descriptive purposes. The x_i refers to individual numbers in the set. We sum up all of the numbers, then divide the sum by N, which is the total number of numbers. Sometimes you will see $\bar{X}$ to refer to the mean of all of the numbers.

In plain English, the formula looks like:

$mean = \frac{\text{Sum of my numbers}}{\text{Count of my numbers}}$

"Well, why didn't you just say that?". We just did.

Let's compute the mean for these five numbers:

> 3 7 9 2 6

Add em up:

> 3+7+9+2+6 = 27

Count em up:

> $i_1 = 3$, $i_2 = 7$, $i_3 = 9$, $i_4 = 2$, $i_5 = 6$; N=5, because i went from 1 to 5

Divide em:

> mean = 27 / 5 = 5.4

Or, to put the numbers in the formula, it looks like this:

$Mean = \bar{X} = \frac{\sum_{i=1}^{n} x_i}{N} = \frac{3+7+9+2+6}{5} = \frac{27}{5} = 5.4$

OK fine, that is how to compute the mean. But, like we imagined, you probably already knew that, and if you didn't that's OK, now you do. What's next?

Is the mean a good measure of central tendency? By now, you should know: it depends.

2.4.5 What does the mean mean?

It is not enough to know the formula for the mean, or to be able to use the formula to compute a mean for a set of numbers. We believe in your ability to add and divide numbers. What you really need to know is what the mean really "means". This requires that you know what the mean does, and not just how to do it. Puzzled? Let's explain.

Can you answer this question: What happens when you divide a sum of numbers by the number of numbers? What are the consequences of doing this? What is the formula doing? What kind of properties does the result give us? FYI, the answer is not that we compute the mean.

OK, so what happens when you divide any number by another number? Of course, the key word here is divide. We literally carve the number up top in the numerator into pieces. How many times do we split the top number? That depends on the bottom number in the denominator. Watch:

$\frac{12}{3} = 4$

So, we know the answer is 4. But, what is really going on here is that we are slicing and dicing up 12 aren't we. Yes, and we slicing 12 into three parts. It turns out the size of those three parts is 4. So, now we are thinking of 12 as three different pieces $12 = 4 + 4 + 4$. I know this will be obvious, but what kind of properties do our pieces have? You mean the fours? Yup. Well, obviously they are all fours. Yes. The pieces are all the same size. They are all equal. So, division equalizes the numerator by the denominator…

"Umm, I think I learned this in elementary school, what does this have to do with the mean?". The number on top of the formula for the mean is just another numerator being divided by a denominator isn't it. In this case, the numerator is a sum of all the values in your data. What if it was the sum of all of the 500 happiness ratings? The sum of all of them would just be a single number adding up all the different ratings. If we split the sum up into equal parts representing one part for each person's happiness what would we get? We would get 500 identical and equal numbers for each person. It would be like taking all of the happiness in the world, then dividing it up equally, then to be fair, giving back the same equal amount of happiness to everyone in the world. This would make some people more happy than they were before, and some people less happy right. Of course, that's because it would be equalizing the distribution of happiness for everybody. This process of equalization by dividing something into equal parts is what the **mean** does. See, it's more than just a formula. It's an idea. This is just the beginning of thinking about these kinds of ideas. We will come back to this idea about the mean, and other ideas, in later chapters.

> Pro tip: The mean is the one and only number that can take the place of every number in the data, such that when you add up all the equal parts, you get back the original sum of the data.

2.4.6 All together now

Just to remind ourselves of the mode, median, and mean, take a look at the next histogram in Figure **??**. We have overlaid the location of the mean (red), median (green), and mode (blue). For this dataset, the three measures of central tendency all give different answers. The mean is the largest because it is influenced by large numbers, even if they occur rarely. The mode and median are insensitive to large numbers that occur infrequently, so they have smaller values.

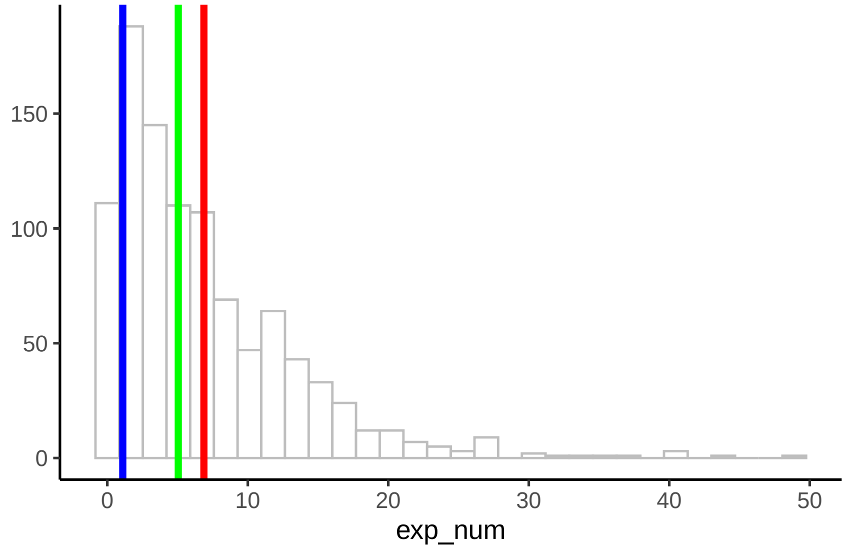


Figure 2.4: A histogram with the mean (red), the median (green), and the mode (blue)

2.5 Measures of Variation (Different*ness*)

What did you do when you wrote essays in high school about a book you read? Probably compare and contrast something right? When you summarize data, you do the same thing. Measures of central tendency give us something like comparing does, they tell us stuff about what is the same. Measures of variation give us something like contrasting does, they tell us stuff about what is different.

First, we note that whenever you see a bunch of numbers that aren't the same, you already know there are some differences. This means the numbers vary, and there is variation in the size of the numbers.

2.5.1 The Range

Consider these 10 numbers, that I already ordered from smallest to largest for you:

1 3 4 5 5 6 7 8 9 24

The numbers have variation, because they are not all the same. We can use the range to describe the width of the variation. The range refers to the **minimum** (smallest value) and **maximum** (largest value) in the set. So, the range would be 1 and 24.

The range is a good way to quickly summarize the boundaries of your data in just two numbers. By computing the range we know that none of the data is larger or smaller than the range. And, it can alert you to outliers. For example, if you are expecting your numbers to be between 1 and 7, but you find the range is 1 - 340,500, then you know you have some big numbers that shouldn't be there, and then you can try to figure out why those numbers occurred (and potentially remove them if something went wrong).

2.5.2 The Difference Scores

It would be nice to summarize the amount of different*ness* in the data. Here's why. If you thought that raw data (lots of numbers) is too big to look at, then you will be frightened to contemplate how many differences there are to look at. For example, these 10 numbers are easy to look at:

1 3 4 5 5 6 7 8 9 24

But, what about the difference between the numbers, what do those look like? We can compute the difference scores between each number, then put them in a matrix like the one below:

	1	3	4	5	5	6	7	8	9	24
1	0	2	3	4	4	5	6	7	8	23
3	-2	0	1	2	2	3	4	5	6	21
4	-3	-1	0	1	1	2	3	4	5	20
5	-4	-2	-1	0	0	1	2	3	4	19
5	-4	-2	-1	0	0	1	2	3	4	19
6	-5	-3	-2	-1	-1	0	1	2	3	18
7	-6	-4	-3	-2	-2	-1	0	1	2	17
8	-7	-5	-4	-3	-3	-2	-1	0	1	16
9	-8	-6	-5	-4	-4	-3	-2	-1	0	15
24	-23	-21	-20	-19	-19	-18	-17	-16	-15	0

We are looking at all of the possible differences between each number and every other number. So, in the top left, the difference between 1 and itself is 0. One column over to the right, the difference between 3 and 1 (3-1) is 2, etc. As you can see, this is a 10x10 matrix, which means there are 100 differences to look at. Not too bad, but if we had 500 numbers, then we would have 500*500 = 250,000 differences to look at (go for it if you like looking at that sort of thing).

Pause for a simple question. What would this matrix look like if all of the 10 numbers in our data were the same number? It should look like a bunch of 0s right? Good. In that case, we could easily see that the numbers have no variation.

But, when the numbers are different, we can see that there is a very large matrix of difference scores. How can we summarize that? How about we apply what we learned from the previous section on measures of central tendency. We have a lot of differences, so we could ask something like, what is the average difference that we have? So, we could just take all of our differences, and compute the mean difference right? What do you think would happen if we did that?

Let's try it out on these three numbers:

1 2 3

	1	2	3
1	0	1	2
2	-1	0	1
3	-2	-1	0

You might already guess what is going to happen. Let's compute the mean:

mean of difference scores $= \frac{0+1+2-1+0+1-2-1+0}{9} = \frac{0}{9} = 0$

Uh oh, we get zero for the mean of the difference scores. This will always happen whenever you take the mean of the difference scores. We can see that there are some differences between the numbers, so using 0 as the summary value for the variation in the numbers doesn't make much sense.

Furthermore, you might also notice that the matrices of difference scores are redundant. The diagonal is always zero, and numbers on one side of the diagonal are the same as the numbers on the other side, except their signs are reversed. So, that's one reason why the difference scores add up to zero.

These are little problems that can be solved by computing the **variance** and the **standard deviation**. For now, the standard deviation is a just a trick that we use to avoid getting a zero. But, later we will see it has properties that are important for other reasons.

2.5.3 The Variance

Variability, variation, variance, vary, variable, varying, variety. Confused yet? Before we describe **the variance**, we want to you be OK with how this word is used. First, don't forget the big picture. We know that variability and variation refers to the big idea of differences between numbers. We can even use the word variance in the same way. When numbers are different, they have variance.

> ℹ Note
>
> The formulas for variance and standard deviation depend on whether you think your data represents an entire population of numbers, or is sample from the population. We discuss this issue in later on. For now, we divide by N, later we discuss why you will often divide by N-1 instead.

The word **variance** also refers to a specific summary statistic, the sum of the squared deviations from the mean. Hold on what? Plain English please. The variance is the sum of the squared difference scores, where the difference scores are computed between each score and the mean. What are these scores? The scores are the numbers in the data set. Let's see the formula in English first:

$variance = \frac{\text{Sum of squared difference scores}}{\text{Number of Scores}}$

2.5.3.1 Deviations from the mean, Difference scores from the mean

We got a little bit complicated before when we computed the difference scores between all of the numbers in the data. Let's do it again, but in a more manageable way. This time, we calculate the difference between each score and the mean. The idea here is

1. We can figure out how similar our scores are by computing the mean
2. Then we can figure out how different our scores are from the mean

This could tell us, 1) something about whether our scores are really all very close to the mean (which could help us know if the mean is good representative number of the data), and 2) something about how much differences there are in the numbers.

Take a look at this table:

scores	values	mean	Difference_from_Mean
1	1	4.5	-3.5
2	6	4.5	1.5
3	4	4.5	-0.5
4	2	4.5	-2.5
5	6	4.5	1.5
6	8	4.5	3.5
Sums	27	27	0
Means	4.5	4.5	0

The first column shows we have 6 scores in the data set, and the `value` columns shows each score. The sum of the values, and the mean is presented on the last two rows. The sum and the mean were obtained by:

$\frac{1+6+4+2+6+8}{6} = \frac{27}{6} = 4.5$.

The third column `mean`, appears a bit silly. We are just listing the mean once for every score. If you think back to our discussion about the meaning of the mean, then you will remember that it equally distributes the total sum across each data point. We can see that here, if we treat each score as the mean, then every score is a 4.5. We can also see that adding up all of the means for each score gives us back 27, which is the sum of the original values. Also, we see that if we find the mean of the mean scores, we get back the mean (4.5 again).

All of the action is occurring in the fourth column, `Difference_from_Mean`. Here, we are showing the difference scores from the mean, using $X_i - \bar{X}$. In other words, we subtracted the mean from each score. So, the first score, 1, is -3.5 from the mean, the second score, 6, is +1.5 from the mean, and so on.

Now, we can look at our original scores and we can look at their differences from the mean. Notice, we don't have a matrix of raw difference scores, so it is much easier to look at out. But, we still have a problem:

We can see that there are non-zero values in the difference scores, so we know there are a differences in the data. But, when we add them all up, we still get zero, which makes it seem like there are a total of zero differences in the data…Why does this happen…and what to do about it?

2.5.3.2 The mean is the balancing point in the data

One brief pause here to point out another wonderful property of the mean. It is the balancing point in the data. If you take a pen or pencil and try to balance it on your figure so it lays flat what are you doing? You need to find the center of mass in the pen, so that half of it is on one side, and the other half is on the other side. That's how balancing works. One side = the other side.

We can think of data as having mass or weight to it. If we put our data on our bathroom scale, we could figure out how heavy it was by summing it up. If we wanted to split the data down the middle so that half of the weight was equal to the other half, then we could balance the data on top of a pin. The mean of the data tells you where to put the pin. It is the location in the data, where the numbers on the one side add up to the same sum as the numbers on the other side.

If we think this through, it means that the sum of the difference scores from the mean will always add up to zero. This is because the numbers on one side of the mean will always add up to -x (whatever the sum of those numbers is), and the numbers of the other side of the mean will always add up to +x (which will be the same value only positive). And:

$-x + x = 0$, right.

Right.

2.5.3.3 The squared deviations

Some devious someone divined a solution to the fact that differences scores from the mean always add to zero. Can you think of any solutions? For example, what could you do to the difference scores so that you could add them up, and they would weigh something useful, that is they would not be zero?

The devious solution is to square the numbers. Squaring numbers converts all the negative numbers to positive numbers. For example, $2^2 = 4$, and $-2^2 = 4$. Remember how squaring works, we multiply the number twice: $2^2 = 2 * 2 = 4$, and $-2^2 = -2 * -2 = 4$. We use the term **squared deviations** to refer to differences scores that have been squared. Deviations are things that move away from something. The difference scores move away from the mean, so we also call them **deviations**.

Let's look at our table again, but add the squared deviations.

scores	values	mean	Difference_from_Mean	Squared_Deviations
1	1	4.5	-3.5	12.25
2	6	4.5	1.5	2.25
3	4	4.5	-0.5	0.25
4	2	4.5	-2.5	6.25
5	6	4.5	1.5	2.25
6	8	4.5	3.5	12.25
Sums	27	27	0	35.5
Means	4.5	4.5	0	5.91666666666667

OK, now we have a new column called `squared_deviations`. These are just the difference scores squared. So, $-3.5^2 = 12.25$, etc. You can confirm for yourself with your cellphone calculator.

Now that all of the squared deviations are positive, we can add them up. When we do this we create something very special called the sum of squares (SS), also known as the sum of the squared deviations from the mean. We will talk at length about this SS later on in the ANOVA chapter. So, when you get there, remember that you already know what it is, just some sums of some squared deviations, nothing fancy.

2.5.3.4 Finally, the variance

Guess what, we already computed the variance. It already happened, and maybe you didn't notice. "Wait, I missed that, what happened?".

First, see if you can remember what we are trying to do here. Take a pause, and see if you can tell yourself what problem we are trying solve.

> pause

Without further ado, we are trying to get a summary of the differences in our data. There are just as many difference scores from the mean as there are data points, which can be a lot, so it would be nice to have a single number to look at, something like a mean, that would tell us about the average differences in the data.

If you look at the table, you can see we already computed the mean of the squared deviations. First, we found the sum (SS), then below that we calculated the mean = 5.916 repeating. This is **the variance**. The variance is the mean of the sum of the squared deviations:

$variance = \frac{SS}{N}$, where SS is the sum of the squared deviations, and N is the number of observations.

OK, now what. What do I do with the variance? What does this number mean? Good question. The variance is often an unhelpful number to look at. Why? Because it is not in the same scale as the original data. This is because we squared the difference scores before taking the mean. Squaring produces large numbers. For example, we see a 12.25 in there. That's a big difference, bigger than any difference between any two original values. What to do? How can we bring the numbers back down to their original unsquared size?

If you are thinking about taking the square root, that's a ding ding ding, correct answer for you. We can always unsquare anything by taking the square root. So, let's do that to 5.916. $\sqrt{5.916} = 2.4322829$.

2.5.4 The Standard Deviation

Oops, we did it again. We already computed the standard deviation, and we didn't tell you. The standard deviation is the square root of the variance…At least, it is right now, until we complicate matters for you in the next chapter.

Here is the formula for the standard deviation:

standard deviation $= \sqrt{Variance} = \sqrt{\frac{SS}{N}}$.

We could also expand this to say:

standard deviation $= \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{N}}$

Don't let those big square root signs put you off. Now, you know what they are doing there. Just bringing our measure of the variance back down to the original size of the data. Let's look at our table again:

scores	values	mean	Difference_from_Mean	Squared_Deviations
1	1	4.5	-3.5	12.25
2	6	4.5	1.5	2.25
3	4	4.5	-0.5	0.25
4	2	4.5	-2.5	6.25
5	6	4.5	1.5	2.25
6	8	4.5	3.5	12.25
Sums	27	27	0	35.5
Means	4.5	4.5	0	5.91666666666667

We measured the standard deviation as 2.4322829. Notice this number fits right in the with differences scores from the mean. All of the scores are kind of in and around + or - 2.4322829. Whereas, if we looked at the variance, 5.916 is just too big, it doesn't summarize the actual differences very well.

What does all this mean? Well, if someone told they had some number with a mean of 4.5 (like the values in our table), and a standard deviation of 2.4322829, you would get a pretty good summary of the numbers. You would know that many of the numbers are around 4.5, and you would know that not all of the numbers are 4.5. You would know that the numbers spread around 4.5. You also know that the spread isn't super huge, it's only + or - 2.4322829 on average. That's a good starting point for describing numbers.

If you had loads of numbers, you could reduce them down to the mean and the standard deviation, and still be pretty well off in terms of getting a sense of those numbers.

2.6 Using Descriptive Statistics with data

Remember, you will be learning how to compute descriptive statistics using software in the labs. Check out the lab manual exercises for descriptives to see some examples of working with real data.

2.7 Rolling your own descriptive statistics

We spent many paragraphs talking about variation in numbers, and how to use calculate the **variance** and **standard deviation** to summarize the average differences between numbers in a data set. The basic process was to 1) calculate some measure of the differences, then 2) average the differences to create a summary. We found that we couldn't average the raw difference scores, because we would always get a zero. So, we squared the differences from the mean, then averaged the squared differences differences. Finally, we square rooted our measure to bring the summary back down to the scale of the original numbers.

Perhaps you haven't heard, but there is more than one way to skin a cat, but we prefer to think of this in terms of petting cats, because some of us love cats. Jokes aside, perhaps you were also thinking that the problem of summing differences scores (so that they don't equal zero), can be solved in more than one way. Can you think of a different way, besides squaring?

2.7.1 Absolute deviations

How about just taking the absolute value of the difference scores. Remember, the absolute value converts any number to a positive value. Check out the following table:

scores	values	mean	Difference_from_Mean	Absolute_Deviations
1	1	4.5	-3.5	3.5
2	6	4.5	1.5	1.5
3	4	4.5	-0.5	0.5
4	2	4.5	-2.5	2.5
5	6	4.5	1.5	1.5
6	8	4.5	3.5	3.5
Sums	27	27	0	13
Means	4.5	4.5	0	2.16666666666667

This works pretty well too. By converting the difference scores from the mean to positive values, we can now add them up and get a non-zero value (if there are differences). Then, we can find the mean of the sum of the absolute deviations. If we were to map the terms sum of squares (SS), variance and standard deviation onto these new measures based off of the absolute deviation, how would the mapping go? For example, what value in the table corresponds to the SS? That would be the sum of absolute deviations in the last column. How about the variance and standard deviation, what do those correspond to? Remember that the variance is mean (SS/N), and the standard deviation is a square-rooted mean ($\sqrt{SS/N}$). In the table above we only have one corresponding mean, the mean of the sum of the absolute deviations. So, we have a **variance** measure that does not need to be square rooted. We might say the mean absolute deviation, is doing double-duty as a variance and a standard-deviation. Neat.

2.7.2 Other sign-inverting operations

In principle, we could create lots of different summary statistics for variance that solve the summing to zero problem. For example, we could raise every difference score to any even numbered power beyond 2 (which is the square). We could use, 4, 6, 8, 10, etc. There is an infinity of even numbers, so there is an infinity of possible variance statistics. We could also use odd numbers as powers, and then take their absolute value. Many things are possible. The important aspect to any of this is to have a reason for what you are doing, and to choose a method that works for the data-analysis problem you are trying to solve. Note also, we bring up this general issue because we want you to understand that statistics is a creative exercise. We invent things when we need them, and we use things that have already been invented when they work for the problem at hand.

2.8 Remember to look at your data

Descriptive statistics are great and we will use them a lot in the course to describe data. You may suspect that descriptive statistics also have some short-comings. This is very true. They are compressed summaries of large piles of numbers. They will almost always be unable to represent all of the numbers fairly. There are also different kinds of descriptive statistics that you could use, and it sometimes not clear which one's you should use.

Perhaps the most important thing you can do when using descriptives is to use them in combination with looking at the data in a graph form. This can help you see whether or not your descriptives are doing a good job of representing the data.

2.8.1 Anscombe's Quartet

To hit this point home, and to get you thinking about the issues we discuss in the next chapter, check this out. It's called Anscombe's Quartet, because these interesting graphs and numbers and numbers were produced by Anscombe (1973). In Figure **??** you are looking at pairs of measurements. Each graph has an X and Y axis, and each point represents two measurements. Each of the graphs looks very different, right?

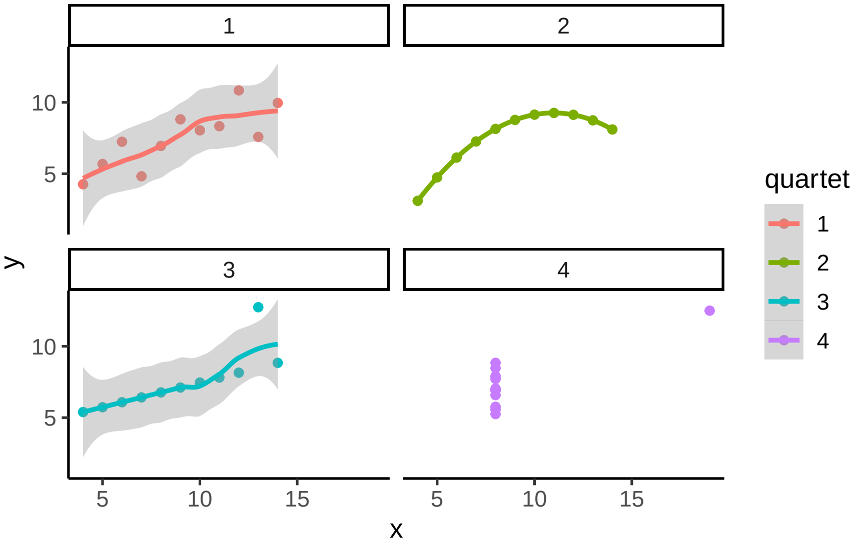


Figure 2.5: Anscombe's Quartet

Well, would you be surprised if I told that the descriptive statistics for the numbers in these graphs are exactly the same? It turns out they do have the same descriptive statistics. In the table below I present the mean and variance for the x-values in each graph, and the mean and the variance for the y-values in each graph.

quartet	mean_x	var_x	mean_y	var_y
1	9	11	7.500909	4.127269
2	9	11	7.500909	4.127629
3	9	11	7.500000	4.122620
4	9	11	7.500909	4.123249

The descriptives are all the same! Anscombe put these special numbers together to illustrate the point of graphing your numbers. If you only look at your descriptives, you don't know what patterns in the data they are hiding. If you look at the graph, then you can get a better understanding.

2.8.2 Datasaurus Dozen

If you thought that Anscombe's quartet was neat, you should take a look at the Datasaurus Dozen (Matejka and Fitzmaurice 2017). Scroll down to see the examples. You will be looking at dot plots. The dot plots show many different patterns, including dinosaurs! What's amazing is that all of the dots have very nearly the same descriptive statistics. Just another reminder to look at your data, it might look like a dinosaur!

2.9 Videos

2.9.1 Measures of center: Mode

2.9.2 Measures of center: Median and Mean

2.9.3 Standard deviation part I

2.9.4 Standard deviation part II

3 Correlation

> Correlation does not equal causation —Every Statistics and Research Methods Instructor Ever

In the last chapter we had some data. It was too much too look at and it didn't make sense. So, we talked about how to look at the data visually using plots and histograms, and we talked about how to summarize lots of numbers so we could determine their central tendencies (sameness) and variability (differentness). And, all was well with the world.

Let's not forget the big reason why we learned about descriptive statistics. The big reason is that we are interested in getting answers to questions using data.

If you are looking for a big theme to think about while you take this course, the theme is: how do we ask and answer questions using data?

For every section in this book, you should be connecting your inner monologue to this question, and asking yourself: How does what I am learning about help me answer questions with data? Advance warning: we know it is easy to forget this stuff when we dive into the details, and we will try to throw you a rope to help you out along the way…remember, we're trying to answer questions with data.

We started Chapter two with some fake data on human happiness, remember? We imagined that we asked a bunch of people to tell us how happy they were, then we looked at the numbers they gave us. Let's continue with this imaginary thought experiment.

What do you get when you ask people to use a number to describe how happy they are? A bunch of numbers. What kind of questions can you ask about those numbers? Well, you can look at the numbers and estimate their general properties as we already did. We would expect those numbers tell us some things we already know. There are different people, and different people are different amounts of happy. You've probably met some of those of really happy people, and really unhappy people, and you yourself probably have some amount of happiness. "Great, thanks Captain Obvious".

Before moving on, you should also be skeptical of what the numbers might mean. For example, if you force people to give a number between 0-100 to rate their happiness, does this number truly reflect how happy that person is? Can a person know how happy they are? Does the question format bias how they give their answer? Is happiness even a real thing? These are all good questions about the **validity** of the construct (happiness itself) and the measure (numbers) you are using to quantify it. For now, though, we will side-step those very important questions, and assume that, happiness is a thing, and our measure of happiness measures something about how happy people are.

OK then, after we have measured some happiness, I bet you can think of some more pressing questions. For example, what causes happiness to go up or down. If you knew the causes of happiness what could you do? How about increase your own happiness; or, help people who are unhappy; or, better appreciate why Eeyore from Winnie the Pooh is unhappy; or, present valid scientific arguments that argue against incorrect claims about what causes happiness. A causal theory and understanding of happiness could be used for all of those things. How can we get there?

Imagine you were an alien observer. You arrived on earth and heard about this thing called happiness that people have. You want to know what causes happiness. You also discover that planet earth has lots of other things. Which of those things, you wonder, cause happiness? How would your alien-self get started on this big question.

As a person who has happiness, you might already have some hunches about what causes changes in happiness. For example things like: weather, friends, music, money, education, drugs, books, movies, beliefs, personality, color of your shoes, eyebrow length, number of cat's you see per day, frequency of subway delay, a lifetime supply of chocolate, et cetera et cetera (as Willy Wonka would say), might all contribute to happiness in someway. There could be many different causes of happiness.

3.1 If something caused something else to change, what would that look like?

Before we go around determining the causes of happiness, we should prepare ourselves with some analytical tools so that we could identify what causation looks like. If we don't prepare ourselves for what we might find, then we won't know how to interpret our own data. Instead, we need to anticipate what the data could look like. Specifically, we need to know what data would look like when one thing does not cause another thing, and what data would look like when one thing does cause another thing. This chapter does some of this preparation. Fair warning: we will find out some tricky things. For example, we can find patterns that look like one thing is causing another, even when that one thing DOES NOT CAUSE the other thing. Hang in there.

3.1.1 Charlie and the Chocolate factory

Let's imagine that a person's supply of chocolate has a causal influence on their level of happiness. Let's further imagine that, like Charlie, the more chocolate you have the more happy you will be, and the less chocolate you have, the less happy you will be. Finally, because we suspect happiness is caused by lots of other things in a person's life, we anticipate that the relationship between chocolate supply and happiness won't be perfect. What do these assumptions mean for how the data should look?

Our first step is to collect some imaginary data from 100 people. We walk around and ask the first 100 people we meet to answer two questions:

1. how much chocolate do you have, and
2. how happy are you.

For convenience, both the scales will go from 0 to 100. For the chocolate scale, 0 means no chocolate, 100 means lifetime supply of chocolate. Any other number is somewhere in between. For the happiness scale, 0 means no happiness, 100 means all of the happiness, and in between means some amount in between.

Here is some sample data from the first 10 imaginary subjects.

subject	chocolate	happiness
1	1	1
2	2	2
3	2	2
4	2	3
5	4	3
6	5	4
7	5	6
8	7	8
9	6	5
10	7	5

We asked each subject two questions so there are two scores for each subject, one for their chocolate supply, and one for their level of happiness. You might already notice some relationships between amount of chocolate and level of happiness in the table. To make those relationships even more clear, let's plot all of the data in a graph.

3.1.2 Scatter plots

When you have two measurements worth of data, you can always turn them into dots and plot them in a scatter plot. A scatter plot has a horizontal x-axis, and a vertical y-axis. You get to choose which measurement goes on which axis. Let's put chocolate supply on the x-axis, and happiness level on the y-axis. Figure **??** shows 100 dots for each subject.

You might be wondering, why are there only 100 dots for the data. Didn't we collect 100 measures for chocolate, and 100 measures for happiness, shouldn't there be 200 dots? Nope. Each dot is for one subject, there are 100 subjects, so there are 100 dots.

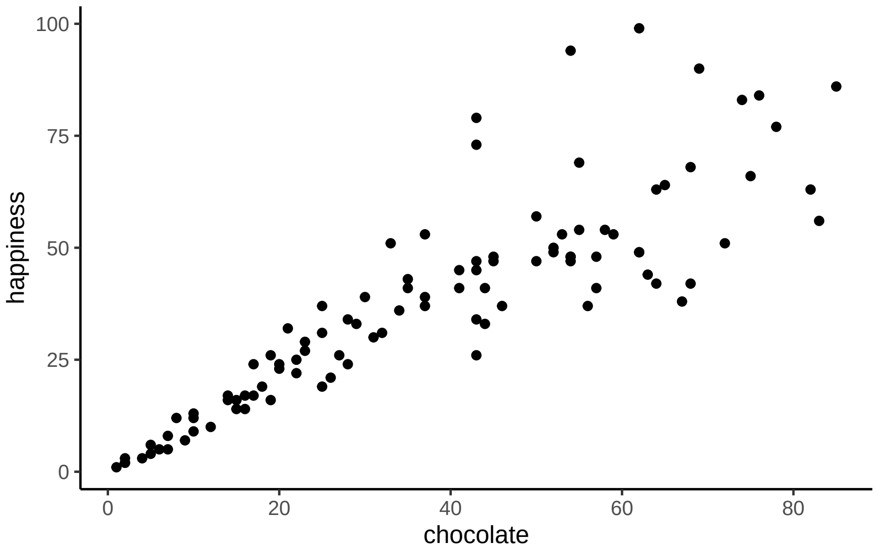


Figure 3.1: Imaginary data showing a positive correlation between amount of chocolate and amount happiness

What do the dots mean? Each dot has two coordinates, an x-coordinate for chocolate, and a y-coordinate for happiness. The first dot, all the way on the bottom left is the first subject in the table, who had close to 0 chocolate and close to zero happiness. You can look at any dot, then draw a straight line down to the x-axis: that will tell you how much chocolate that subject has. You can draw a straight line left to the y-axis: that will tell you how much happiness the subject has.

Now that we are looking at the scatter plot, we can see many things. The dots are scattered around a bit aren't they, hence **scatter plot**. Even when the dot's don't scatter, they're still called scatter plots, perhaps because those pesky dots in real life have so much scatter all the time. More important, the dots show a relationship between chocolate supply and happiness. Happiness is lower for people with smaller supplies of chocolate, and higher for people with larger supplies of chocolate. It looks like the more chocolate you have the happier you will be, and vice-versa. This kind of relationship is called a **positive correlation**.

3.1.3 Positive, Negative, and No-Correlation

Seeing as we are in the business of imagining data, let's imagine some more. We've already imagined what data would look like if larger chocolate supplies increase happiness. We'll show that again in a bit. What do you imagine the scatter plot would look like if the relationship was reversed, and larger chocolate supplies decreased happiness. Or, what do you imagine the scatter plot would look like if there was no relationship, and the amount of chocolate that you have doesn't do anything to your happiness. We invite your imagination to look at Figure **??**:

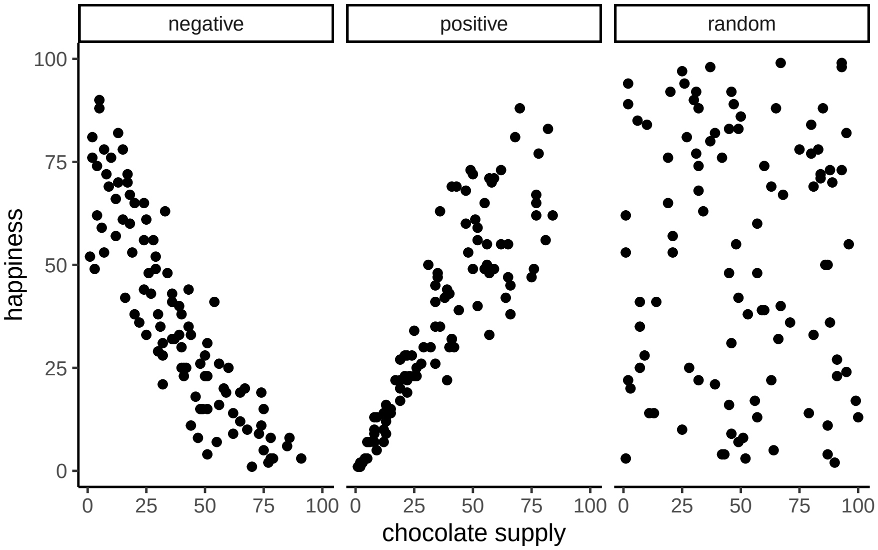


Figure 3.2: Three scatterplots showing negative, positive, and zero correlation

The first panel shows a **negative correlation**. Happiness goes down as chocolate supply increases. Negative correlation occurs when one thing goes up and the other thing goes down; or, when more of X is less of Y, and vice-versa. The second panel shows a **positive correlation**. Happiness goes up as chocolate as chocolate supply increases. Positive correlation occurs when both things go up together, and go down together: more of X is more of Y, and vice-versa. The third panel shows **no correlation**. Here, there doesn't appear to be any obvious relationship between chocolate supply and happiness. The dots are scattered all over the place, the truest of the scatter plots.

> **i** Note
>
> We are wading into the idea that measures of two things can be related, or correlated with one another. It is possible for the relationships to be more complicated than just going up, or going down. For example, we could have a relationship that where the dots go up for the first half of X, and then go down for the second half.

Zero correlation occurs when one thing is not related in any way to another things: changes in X do not relate to any changes in Y, and vice-versa.

3.2 Pearson's r

"So you've examined your scatter plots and now you might be wondering how to quantify what you see. We've already covered how to generate descriptive statistics for individual variables—think of single measures like happiness levels or chocolate consumption, summarized through means, variances, and so on. But what if you want to capture the relationship between two such variables in a single descriptive statistic? Is that even possible? Karl Pearson to the rescue.

ℹ Note

The stories about the invention of various statistics are very interesting, you can read more about them in the book, "The Lady Tasting Tea" (Salsburg 2001)

There's a statistic for that, and Karl Pearson invented it. Everyone now calls it, "Pearson's r". We will find out later that Karl Pearson was a big-wig editor at Biometrika in the 1930s. He took a hating to another big-wig statistician, Sir Ronald Fisher (who we learn about later), and they had some statistics fights. Even in the stats world, not everyone plays nice in the sandbox.

How does Pearson's r work? Let's look again at the first 10 subjects in our fake experiment:

subject	chocolate	happiness
1	1	1
2	2	2
3	2	2
4	2	3
5	4	3
6	5	4
7	5	6
8	7	8
9	6	5
10	7	5
Sums	41	39
Means	4.1	3.9

What could we do to these numbers to produce a single summary value that represents the relationship between the chocolate supply and happiness?

3.2.1 The idea of co-variance

"Oh please no, don't use the word variance again". Yes, we're doing it, we're going to use the word variance again, and again, until it starts making sense. Remember what variance means about some numbers. It means the numbers have some change in them, they are not all the same, some of them are big, some are small. We can see that there is variance in chocolate supply across the 10 subjects. We can see that there is variance in happiness across the 10 subjects. We also saw in the scatter plot, that happiness increases as chocolate supply increases; which is a positive relationship, a positive correlation. What does this have to do with variance? Well, it means there is a relationship between the variance in chocolate supply, and the variance in happiness levels. The two measures vary together don't they? When we have two measures that vary together, they are like a happy couple who share their variance. This is what co-variance refers to, the idea that the pattern of varying numbers in one measure is shared by the pattern of varying numbers in another measure.

Co-variance is **very, very, very, very** important. I suspect that the word co-variance is initially confusing, especially if you are not yet fully comfortable with the meaning of variance for a single measure. Nevertheless, we must proceed and use the idea of co-variance over and over again to firmly implant it into your statistical mind (we already said, but redundancy works, it's a thing).

> Pro tip: Three-legged race is a metaphor for co-variance. Two people tie one leg to each other, then try to walk. It works when they co-vary their legs together (positive relationship). They can also co-vary in an unhelpful way, when one person tries to move forward exactly when the other person tries to move backward. This is still co-variance (negative relationship). Funny random walking happens when there is no co-variance. This means one person does whatever they want, and so does the other person. There is a lot of variance, but the variance is shared randomly, so it's just a bunch of legs moving around accomplishing nothing.
>
> Pro tip #2: Successfully playing paddy-cake occurs when two people coordinate their actions so they have positively shared co-variance.

3.3 Turning the numbers into a measure of co-variance

"OK, so if you are saying that co-variance is just another word for correlation or relationship between two measures, I'm good with that. I suppose we would need some way to measure that." Correct, back to our table…notice anything new?

subject	chocolate	happiness	Chocolate_X_Happiness
1	1	1	1
2	2	2	4
3	2	2	4
4	2	3	6
5	4	3	12
6	5	4	20
7	5	6	30
8	7	8	56
9	6	5	30
10	7	5	35
Sums	41	39	198
Means	4.1	3.9	19.8

We've added a new column called `Chocolate_X_Happiness`, which translates to Chocolate scores multiplied by Happiness scores. Each row in the new column, is the product, or multiplication of the chocolate and happiness score for that row. Yes, but why would we do this?

Last chapter we took you back to Elementary school and had you think about division. Now it's time to do the same thing with multiplication. We assume you know how that works. One number times another, means taking the first number, and adding it as many times as the second says to do,

$2 * 2 = 2 + 2 = 4$

$2 * 6 = 2 + 2 + 2 + 2 + 2 + 2 = 12$, or $6 + 6 = 12$, same thing.

Yes, you know all that. But, can you bend multiplication to your will, and make it do your bidding when need to solve a problem like summarizing co-variance? Multiplication is the droid you are looking for.

We know how to multiple numbers, and all we have to next is think about the consequences of multiplying sets of numbers together. For example, what happens when you multiply two small numbers together, compared to multiplying two big numbers together? The first product should be smaller than the second product right? How about things like multiplying a small number by a big number? Those products should be in between right?.

Then next step is to think about how the products of two measures sum together, depending on how they line up. Let's look at another table:

scores	X	Y	A	B	XY	AB
1	1	1	1	10	1	10
2	2	2	2	9	4	18
3	3	3	3	8	9	24
4	4	4	4	7	16	28
5	5	5	5	6	25	30
6	6	6	6	5	36	30
7	7	7	7	4	49	28
8	8	8	8	3	64	24
9	9	9	9	2	81	18
10	10	10	10	1	100	10
Sums	55	55	55	55	385	220
Means	5.5	5.5	5.5	5.5	38.5	22

Look at the X and Y column. The scores for X and Y perfectly co-vary. When X is 1, Y is 1; when X is 2, Y is 2, etc. They are perfectly aligned. The scores for A and B also perfectly co-vary, just in the opposite manner. When A is 1, B is 10; when A is 2, B is 9, etc. B is a reversed copy of A.

Now, look at the column XY. These are the products we get when we multiply the values of X across with the values of Y. Also, look at the column AB. These are the products we get when we multiply the values of A across with the values of B. So far so good.

Now, look at the `Sums` for the XY and AB columns. Not the same. The sum of the XY products is 385, and the sum of the AB products is 220. For this specific set of data, the numbers 385 and 220 are very important. They represent the biggest possible sum of products (385), and the smallest possible sum of products (220). There is no way of re-ordering the numbers 1 to 10, say for X, and the numbers 1 to 10 for Y, that would ever produce larger or smaller numbers. Don't believe me? Check this out:

Figure **??** shows 1000 computer simulations. I convinced my computer to randomly order the numbers 1 to 10 for X, and randomly order the numbers 1 to 10 for Y. Then, I multiplied X and Y, and added the products together. I did this 1000 times. The dots show the sum of the products for each simulation. The two black lines show the maximum possible sum (385), and the minimum possible sum (220), for this set of numbers. Notice, how all of the dots are in between the maximum and minimum possible values. Told you so.

"OK fine, you told me so…So what, who cares?". We've been looking for a way to summarize the co-variance between two measures right? Well, for these numbers, we have found one, haven't we. It's the sum of the products. We know that when the sum of the products is 385, we have found a perfect, positive correlation. We know, that when the sum of the products is 220, we have found a perfect negative correlation. What about the numbers in between. What could we conclude about the correlation if we found the sum of the products to be 350. Well, it's going to be positive, because it's close to 385, and that's perfectly positive. If the sum of the products was 240, that's going to be negative, because it's close to the perfectly

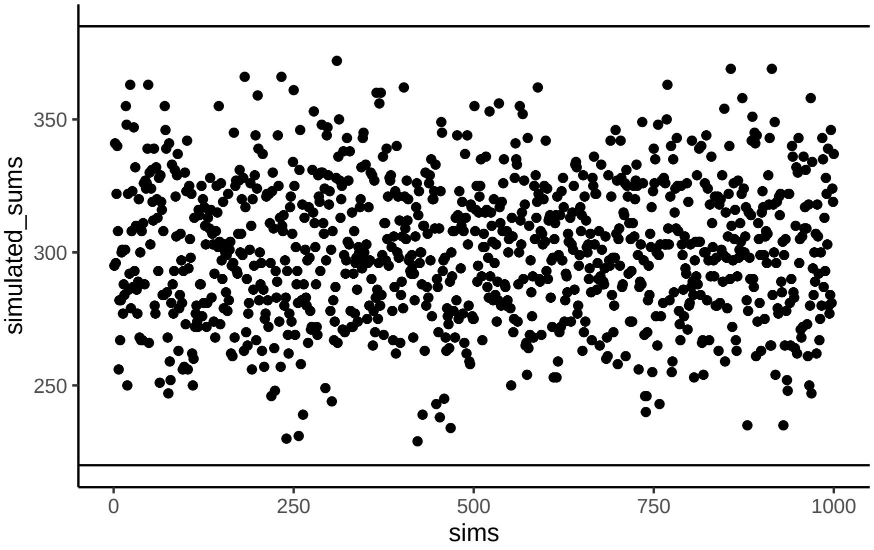


Figure 3.3: Simulated sums of products showing the kinds of values than can be produced by randomly ordering the numbers in X and Y.

negatively correlating 220. What about no correlation? Well, that's going to be in the middle between 220 and 385 right.

We have just come up with a data-specific summary measure for the correlation between the numbers 1 to 10 in X, and the numbers 1 to 10 in Y, it's the sum of the products. We know the maximum (385) and minimum values (220), so we can now interpret any product sum for this kind of data with respect to that scale.

> Pro tip: When the correlation between two measures increases in the positive direction, the sum of their products increases to its maximum possible value. This is because the bigger numbers in X will tend to line up with the bigger numbers in Y, creating the biggest possible sum of products. When the correlation between two measures increases in the negative direction, the sum of their products decreases to its minimum possible value. This is because the bigger numbers in X will tend to line up with the smaller numbers in Y, creating the smallest possible sum of products. When there is no correlation, the big numbers in X will be randomly lined up with the big and small numbers in Y, making the sum of the products, somewhere in the middle.

3.3.1 Co-variance, the measure

We took some time to see what happens when you multiply sets of numbers together. We found that $big * big = bigger$ and $small * small =$ still small, and $big * small =$ in the middle.

The purpose of this was to give you some conceptual idea of how the co-variance between two measures is reflected in the sum of their products. We did something very straightforward. We just multiplied X with Y, and looked at how the product sums get big and small, as X and Y co-vary in different ways.

Now, we can get a little bit more formal. In statistics, **co-variance** is not just the straight multiplication of values in X and Y. Instead, it's the multiplication of the deviations in X from the mean of X, and the deviation in Y from the mean of Y. Remember those difference scores from the mean we talked about last chapter? They're coming back to haunt you know, but in a good way like Casper the friendly ghost.

Let's see what this look like in a table:

subject	chocolate	happiness	C_d	H_d	Cd_x_Hd
1	1	1	-3.1	-2.9	8.99
2	2	2	-2.1	-1.9	3.99
3	2	2	-2.1	-1.9	3.99
4	2	3	-2.1	-0.9	1.89
5	4	3	-0.1	-0.9	0.09
6	5	4	0.9	0.1	0.09
7	5	6	0.9	2.1	1.89
8	7	8	2.9	4.1	11.89
9	6	5	1.9	1.1	2.09
10	7	5	2.9	1.1	3.19
Sums	41	39	0	0	38
Means	4.1	3.9	0	0	3.81

We have computed the deviations from the mean for the chocolate scores (column `C_d`), and the deviations from the mean for the happiness scores (column `H_d`). Then, we multiplied them together (last column). Finally, you can see the mean of the products listed in the bottom right corner of the table, the official **the covariance**.

The formula for the co-variance is:

$cov(X,Y) = \frac{\sum_i^n (x_i - \bar{X})(y_i - \bar{Y})}{N}$

OK, so now we have a formal single number to calculate the relationship between two variables. This is great, it's what we've been looking for. However, there is a problem. Remember when we learned how to compute just the plain old **variance**. We looked at that number, and we didn't know what to make of it. It was squared, it wasn't in the same scale as the original data. So, we square rooted the **variance** to produce the **standard deviation**, which gave us a more interpretable number in the range of our data. The **co-variance** has a similar problem. When you calculate the co-variance as we just did, we don't know immediately know its scale. Is a 3 big? is a 6 big? is a 100 big? How big or small is this thing?

From our prelude discussion on the idea of co-variance, we learned the sum of products between two measures ranges between a maximum and minimum value. The same is true of the co-variance. For a given set of data, there is a maximum possible positive value for the co-variance (which occurs when there is perfect positive correlation). And, there is a minimum possible negative value for the co-variance (which occurs when there is a perfect negative correlation). When there is zero co-variation, guess what happens. Zeroes. So, at the very least, when we look at a co-variation statistic, we can see what direction it points, positive or negative. But, we don't know how big or small it is compared to the maximum or minimum possible value, so we don't know the relative size, which means we can't say how strong the correlation is. What to do?

3.3.2 Pearson's r we there yet

Yes, we are here now. Wouldn't it be nice if we could force our measure of co-variation to be between -1 and +1?

-1 would be the minimum possible value for a perfect negative correlation. +1 would be the maximum possible value for a perfect positive correlation. 0 would mean no correlation. Everything in between 0 and -1 would be increasingly large negative correlations. Everything between 0 and +1 would be increasingly large positive correlations. It would be a fantastic, sensible, easy to interpret system. If only we could force the co-variation number to be between -1 and 1. Fortunately, for us, this episode is brought to you by Pearson's r, which does precisely this wonderful thing.

Let's take a look at a formula for Pearson's r:

$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{cov(X,Y)}{SD_X SD_Y}$

We see the symbol σ here, that's more Greek for you. σ is often used as a symbol for the standard deviation (SD). If we read out the formula in English, we see that r is the co-variance of X and Y, divided by the product of the standard deviation of X and the standard deviation of Y. Why are we dividing the co-variance by the product of the standard deviations. This operation has the effect of **normalizing** the co-variance into the range -1 to 1.

> Note
>
> But, we will fill this part in as soon as we can…promissory note to explain the magic. FYI, it's not magic. Brief explanation here is that dividing each measure by its standard deviation ensures that the values in each measure are in the same range as one another.

For now, we will call this mathematical magic. It works, but we don't have space to tell you why it works right now.

It's worth saying that there are loads of different formulas for computing Pearson's r. You can find them by Googling them. We will probably include more of them here, when we get around to it. However, they all give you the same answer. And, they are all not as pretty as each other. Some of them might even look scary. In other statistics textbook you will often find formulas that are easier to use for calculation purposes. For example, if you only had a pen and paper, you might use one or another formula because it helps you compute the answer faster by hand. To be honest, we are not very interested in teaching you how to plug numbers into formulas. We give one lesson on that here: Put the numbers into the letters, then compute the answer. Sorry to be snarky. Nowadays you have a computer that you should use for this kind of stuff. So, we are more interested in teaching you what the calculations mean, rather than how to do them. Of course, every week we are showing you how to do the calculations in lab with computers, because that is important too.

Does Pearson's r really stay between -1 and 1 no matter what? It's true, take a look at the following simulation. Here I randomly ordered the numbers 1 to 10 for an X measure, and did the same for a Y measure. Then, I computed Pearson's r, and repeated this process 1000 times. As you can see from Figure **??** all of the dots are between -1 and 1. Neat huh.

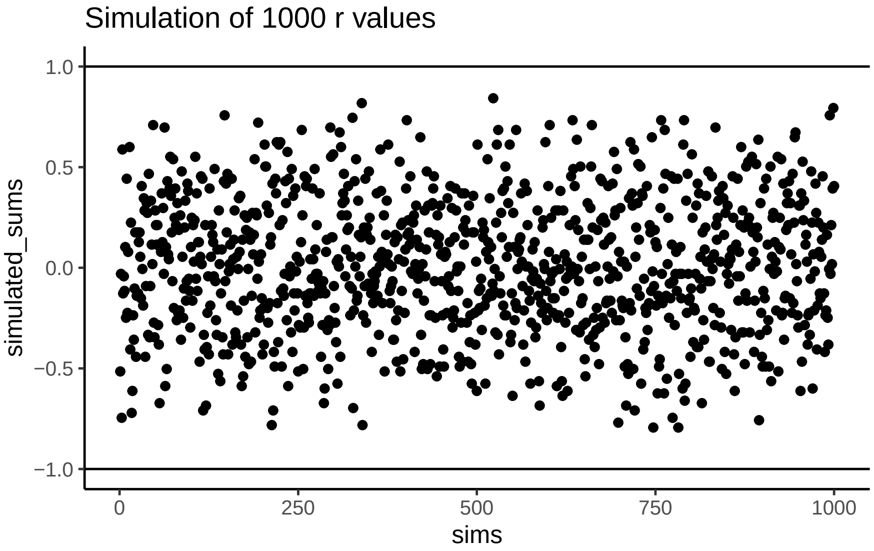


Figure 3.4: A simulation of of correlations. Each dot represents the r-value for the correlation between an X and Y variable that each contain the numbers 1 to 10 in random orders. The figure ilustrates that many r-values can be obtained by this random process

3.4 Examples with Data

In the lab for correlation you will be shown how to compute correlations in real data-sets using software. To give you a brief preview, let's look at some data from the world happiness report

(2018).

This report measured various attitudes across people from different countries. For example, one question asked about how much freedom people thought they had to make life choices. Another question asked how confident people were in their national government. **?@fig-3hrsdata** is a scatterplot showing the relationship between these two measures. Each dot represents means for different countries.

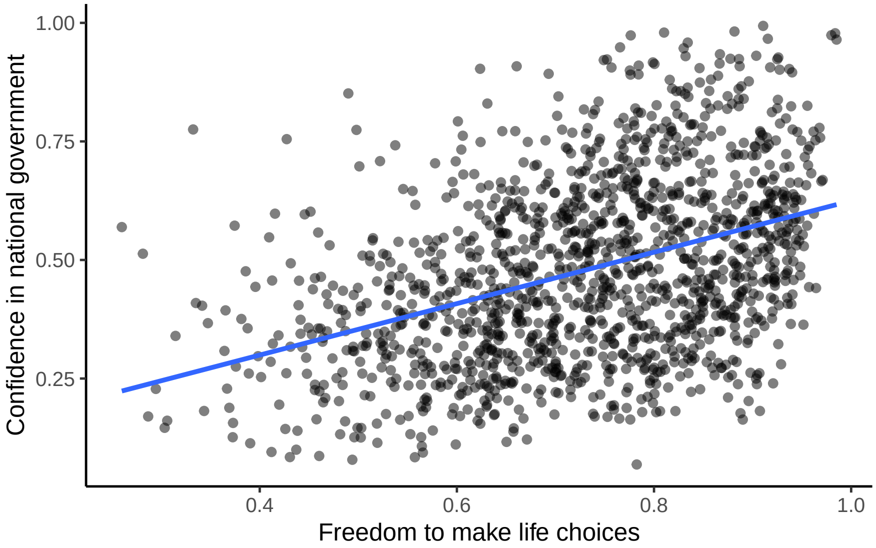


Figure 3.5: Relationship between freedom to make life choices and confidence in national government. Data from the world happiness report for 2018

We put a blue line on the scatterplot to summarize the positive relationship. It appears that as "freedom to make life choices goes up", so to does confidence in national government. It's a positive correlation.

The actual correlation, as measured by Pearson's r is:

```
#> [1] 0.4080963
```

You will do a lot more of this kind of thing in the lab. Looking at the graph you might start to wonder: Does freedom to make life choices cause changes how confident people are in their national government? Our does it work the other way? Does being confident in your national government give you a greater sense of freedom to make life choices? Or, is this just a random relationship that doesn't mean anything? All good questions. These data do not provide the answers, they just suggest a possible relationship.

3.5 Regression: A mini intro

We're going to spend the next little bit adding one more thing to our understanding of correlation. It's called **linear regression**. It sounds scary, and it really is. You'll find out much later in your Statistics education that everything we will be soon be talking about can be thought of as a special case of regression. But, we don't want to scare you off, so right now we just introduce the basic concepts.

First, let's look at a linear regression. This way we can see what we're trying to learn about. Figure **??** shows the same scatter plots as before with something new: lines!

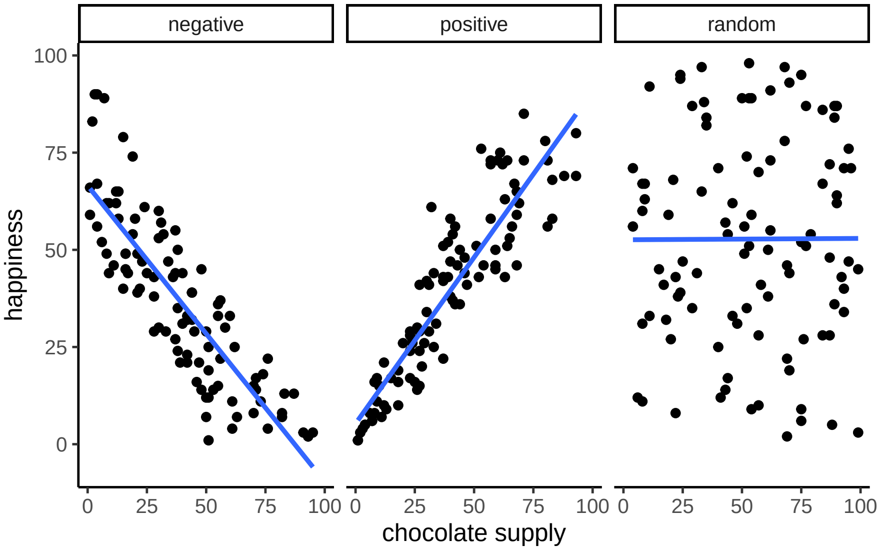


Figure 3.6: Three scatterplots showing negative, positive, and a random correlation (where the r-value is expected to be 0), along with the best fit regression line

3.5.1 The best fit line

Notice anything about these blue lines? Hopefully you can see, at least for the first two panels, that they go straight through the data, just like a kebab skewer. We call these lines **best fit** lines, because according to our definition (soon we promise) there are no other lines that you could draw that would do a better job of going straight throw the data.

One big idea here is that we are using the line as a kind of mean to describe the relationship between the two variables. When we only have one variable, that variable exists on a single dimension, it's 1D. So, it is appropriate that we only have one number, like the mean, to describe it's central tendency. When we have two variables, and plot them together, we now have a two-dimensional space. So, for two dimensions we could use a bigger thing that is 2d, like a line, to summarize the central tendency of the relationship between the two variables.

What do we want out of our line? Well, if you had a pencil, and a printout of the data, you could draw all sorts of straight lines any way you wanted. Your lines wouldn't even have to go through the data, or they could slant through the data with all sorts of angles. Would all of those lines be very good a describing the general pattern of the dots? Most of them would not. The best lines would go through the data following the general shape of the dots. Of the best lines, however, which one is the best? How can we find out, and what do we mean by that? In short, the best fit line is the one that has the least error.

ℹ Note

R code for plotting residuals thanks to Simon Jackson's blog post: https://drsimonj.svbtle.com/visualising-residuals

Check out this next plot, it shows a line through some dots. But, it also shows some teeny tiny lines. These lines drop down from each dot, and they land on the line. Each of these little lines is called a **residual**. They show you how far off the line is for different dots. It's measure of error, it shows us just how wrong the line is. After all, it's pretty obvious that not all of the dots are on the line. This means the line does not actually represent all of the dots. The line is wrong. But, the best fit line is the least wrong of all the wrong lines.

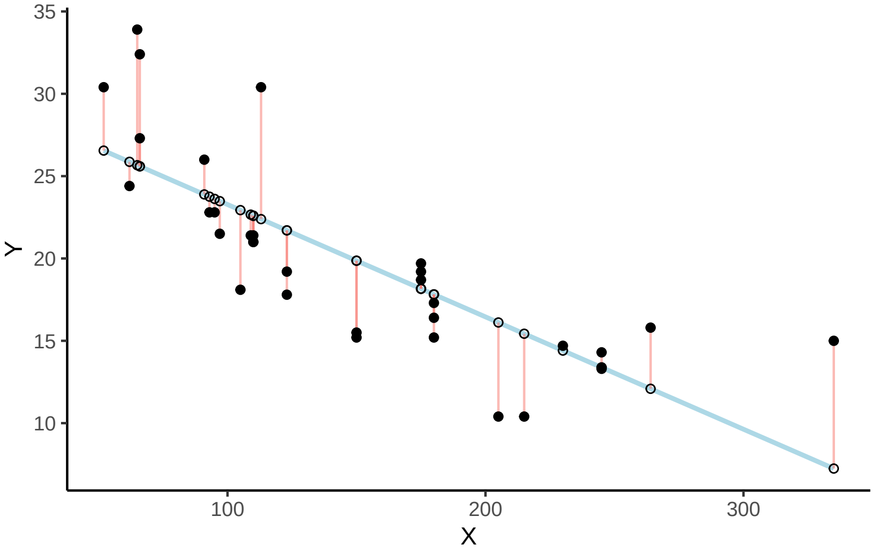


Figure 3.7: Black dots represent data points. The blue line is the best fit regression line. The white dots are repesent the predicted location of each black dot. The red lines show the error between each black dot and the regression line. The blue line is the best fit line because it minimizes the error shown by the red lines.

There's a lot going on in Figure **??**. First, we are looking at a scatter plot of two variables, an X and Y variable. Each of the black dots are the actual values from these variables. You can see there is a negative correlation here, as X increases, Y tends to decrease. We drew a regression line through the data, that's the blue line. There's these little white dots too. This is where the line thinks the black dots should be. The red lines are the important residuals we've been talking about. Each black dot has a red line that drops straight down, or straight up from the location of the black dot, and lands directly on the line. We can already see that many of the dots are not on the line, so we already know the line is "off" by some amount for each dot. The red line just makes it easier to see exactly how off the line is.

The important thing that is happening here, is that the the blue line is drawn is such a way, that it minimizes the total length of the red lines. For example, if we wanted to know how wrong this line was, we could simply gather up all the red lines, measure how long they are, and then add all the wrongness together. This would give us the total amount of wrongness. We usually call this the error. In fact, we've already talked about this idea before when we discussed standard deviation. What we will actually be doing with the red lines, is computing the sum of the squared deviations from the line. That sum is the total amount of error. Now, this blue line here minimizes the sum of the squared deviations. Any other line would produce a larger total error.

?@fig-3regressionGIF is an animation to see this in action. The animations compares the best fit line in blue, to some other possible lines in black. The black line moves up and down. The red lines show the error between the black line and the data points. As the black line moves toward the best fit line, the total error, depicted visually by the grey area shrinks to it's minimum value. The total error expands as the black line moves away from the best fit line.

Whenever the black line does not overlap with the blue line, it is worse than the best fit line. The blue regression line is like Goldilocks, it's just right, and it's in the middle.

Figure **??** shows how the sum of squared deviations (the sum of the squared lengths of the red lines) behaves as we move the line up and down. What's going on here is that we are computing a measure of the total error as the black line moves through the best fit line. This represents the sum of the squared deviations. In other words, we square the length of each red line from the above animation, then we add up all of the squared red lines, and get the total error (the total sum of the squared deviations). The graph below shows what the total error looks like as the black line approaches then moves away from the best fit line. Notice, the dots in this graph start high on the left side, then they swoop down to a minimum at the bottom middle of the graph. When they reach their minimum point, we have found a line that minimizes the total error. This is the best fit regression line.

OK, so we haven't talked about the y-intercept yet. But, what this graph shows us is how the total error behaves as we move the line up and down. The y-intercept here is the thing we change that makes our line move up and down. As you can see the dots go up when we move the line down from 0 to -5, and the dots go up when we move the line up from 0 to +5. The

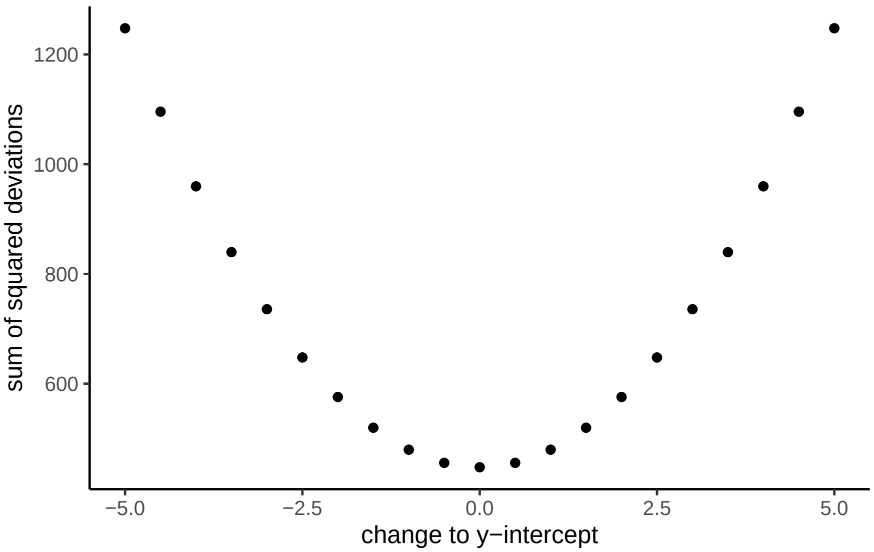


Figure 3.8: A plot of the sum of the squared deviations for different lines moving up and down, through the best fit line. The best fit line occurs at the position that minimizes the sum of the sqaured deviations.

best line, that minimizes the error occurs right in the middle, when we don't move the blue regression line at all.

3.5.2 Lines

OK, fine you say. So, there is one magic line that will go through the middle of the scatter plot and minimize the sum of the squared deviations. How do I find this magic line? We'll show you. But, to be completely honest, you'll almost never do it the way we'll show you here. Instead, it's much easier to use software and make your computer do it for. You'll learn how to that in the labs.

Before we show you how to find the regression line, it's worth refreshing your memory about how lines work, especially in 2 dimensions. Remember this?

$y = ax + b$, or also $y = mx + b$ (sometimes a or m is used for the slope)

This is the formula for a line. Another way of writing it is:

$y = slope * x +$ y-intercept

The slope is the slant of the line, and the y-intercept is where the line crosses the y-axis. Let's look at the lines in Figure **??**.

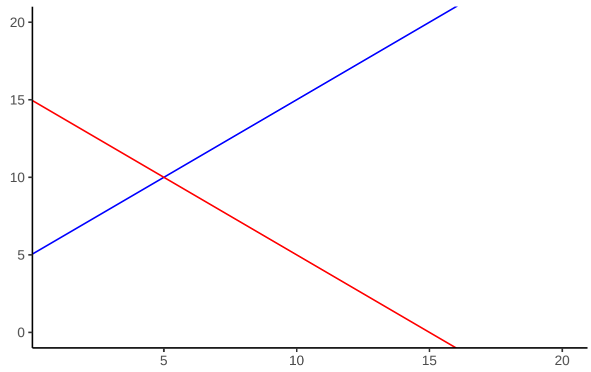


Figure 3.9: Two different lines with different y-intercepts (where the line crosses the y-axis), and different slopes. A positive slope makes the line go up from left to right. A negative slope makes the line go down from left to right.

The formula for the blue line is $y = 1 * x + 5$. Let's talk about that. When x = 0, where is the blue line on the y-axis? It's at five. That happens because 1 times 0 is 0, and then we just have the five left over. How about when x = 5? In that case y =10. You just need the plug in the numbers to the formula, like this:

$y = 1 * x + 5$ $y = 1 * 5 + 5 = 5 + 5 = 10$

The point of the formula is to tell you where y will be, for any number of x. The slope of the line tells you whether the line is going to go up or down, as you move from the left to the right. The blue line has a positive slope of one, so it goes up as x goes up. How much does it go up? It goes up by one for everyone one of x! If we made the slope a 2, it would be much steeper, and go up faster. The red line has a negative slope, so it slants down. This means y goes down, as x goes up. When there is no slant, and we want to make a perfectly flat line, we set the slope to 0. This means that y doesn't go anywhere as x gets bigger and smaller.

That's lines.

3.5.3 Computing the best fit line

If you have a scatter plot showing the locations of scores from two variables, the real question is how can you find the slope and the y-intercept for the best fit line? What are you going to do? Draw millions of lines, add up the residuals, and then see which one was best? That would take forever. Fortunately, there are computers, and when you don't have one around, there's also some handy formulas.

Note

It's worth pointing out just how much computers have changed everything. Before computers everyone had to do these calculations by hand, such a chore! Aside from the deeper mathematical ideas in the formulas, many of them were made for convenience, to speed up hand calculations, because there were no computers. Now that we have computers, the hand calculations are often just an exercise in algebra. Perhaps they build character. You decide.

We'll show you the formulas. And, work through one example by hand. It's the worst, we know. By the way, you should feel sorry for me as I do this entire thing by hand for you.

Here are two formulas we can use to calculate the slope and the intercept, straight from the data. We won't go into why these formulas do what they do. These ones are for "easy" calculation.

$intercept = b = \frac{\sum y \sum x^2 - \sum x \sum xy}{n \sum x^2 - (\sum x)^2}$

$slope = m = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$

In these formulas, the x and the y refer to the individual scores. Here's a table showing you how everything fits together.

scores	x	y	x_squared	y_squared	xy
1	1	2	1	4	2
2	4	5	16	25	20
3	3	1	9	1	3
4	6	8	36	64	48
5	5	6	25	36	30
6	7	8	49	64	56
7	8	9	64	81	72
Sums	34	39	200	275	231

We see 7 sets of scores for the x and y variable. We calculated x^2 by squaring each value of x, and putting it in a column. We calculated y^2 by squaring each value of y, and putting it in a column. Then we calculated xy, by multiplying each x score with each y score, and put that in a column. Then we added all the columns up, and put the sums at the bottom. These are all the number we need for the formulas to find the best fit line. Here's what the formulas look like when we put numbers in them:

$intercept = b = \frac{\sum y \sum x^2 - \sum x \sum xy}{n \sum x^2 - (\sum x)^2} = \frac{39 * 200 - 34 * 231}{7 * 200 - 34^2} = -.221$

$slope = m = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{7 * 231 - 34 * 39}{7 * 275 - 34^2} = 1.19$

Great, now we can check our work, let's plot the scores in a scatter plot and draw a line through it with slope = 1.19, and a y-intercept of -.221. As shown in Figure **??**, the line should go through the middle of the dots.

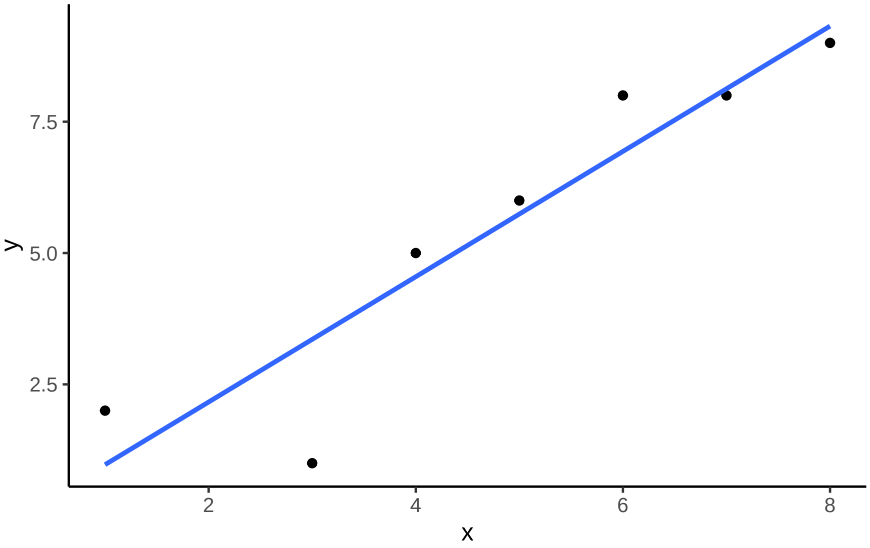


Figure 3.10: An example regression line with confidence bands going through a few data points in a scatterplot

3.6 Interpreting Correlations

What does the presence or the absence of a correlation between two measures mean? How should correlations be interpreted? What kind of inferences can be drawn from correlations? These are all very good questions. A first piece of advice is to use caution when interpreting correlations. Here's why.

3.6.1 Correlation does not equal causation

Perhaps you have heard that correlation does not equal causation. Why not? There are lots of reasons why not. However, before listing some of the reasons let's start with a case where we would expect a causal connection between two measurements. Consider, buying a snake plant for your home. Snake plants are supposed to be easy to take care of because you can mostly ignore them.

Like most plants, snake plants need some water to stay alive. However, they also need just the right amount of water. Imagine an experiment where 1000 snake plants were grown in a house.

Each snake plant is given a different amount of water per day, from zero teaspoons of water per day to 1000 teaspoons of water per day. We will assume that water is part of the causal process that allows snake plants to grow. The amount of water given to each snake plant per day can also be one of our measures. Imagine further that every week the experimenter measures snake plant growth, which will be the second measurement. Now, can you imagine for yourself what a scatter plot of weekly snake plant growth by tablespoons of water would look like?

3.6.1.1 Even when there is causation, there might not be obvious correlation

The first plant given no water at all would have a very hard time and eventually die. It should have the least amount of weekly growth. How about the plants given only a few teaspoons of water per day. This could be just enough water to keep the plants alive, so they will grow a little bit but not a lot. If you are imagining a scatter plot, with each dot being a snake plant, then you should imagine some dots starting in the bottom left hand corner (no water & no plant growth), moving up and to the right (a bit of water, and a bit of growth). As we look at snake plants getting more and more water, we should see more and more plant growth, right? "Sure, but only up to a point". Correct, there should be a trend for a positive correlation with increasing plant growth as amount of water per day increases. But, what happens when you give snake plants too much water? From personal experience, they die. So, at some point, the dots in the scatter plot will start moving back down again. Snake plants that get way too much water will not grow very well.

The imaginary scatter plot you should be envisioning could have an upside U shape. Going from left to right, the dot's go up, they reach a maximum, then they go down again reaching a minimum. Computing Pearson's r for data like this can give you r values close to zero. The scatter plot could look something like Figure **??**.

Granted this looks more like an inverted V, than an inverted U, but you get the picture right? There is clearly a relationship between watering and snake plant growth. But, the correlation isn't in one direction. As a result, when we compute the correlation in terms of Pearson's r, we get a value suggesting no relationship.

```
#> [1] -0.004432389
```

What this really means is there is no linear relationship that can be described by a single straight line. When we need lines or curves going in more than one direction, we have a nonlinear relationship.

This example illustrates some conundrums in interpreting correlations. We already know that water is needed for plants to grow, so we are rightly expecting there to be a relationship between our measure of amount of water and plant growth. If we look at the first half of the data we see a positive correlation, if we look at the last half of the data we see a negative

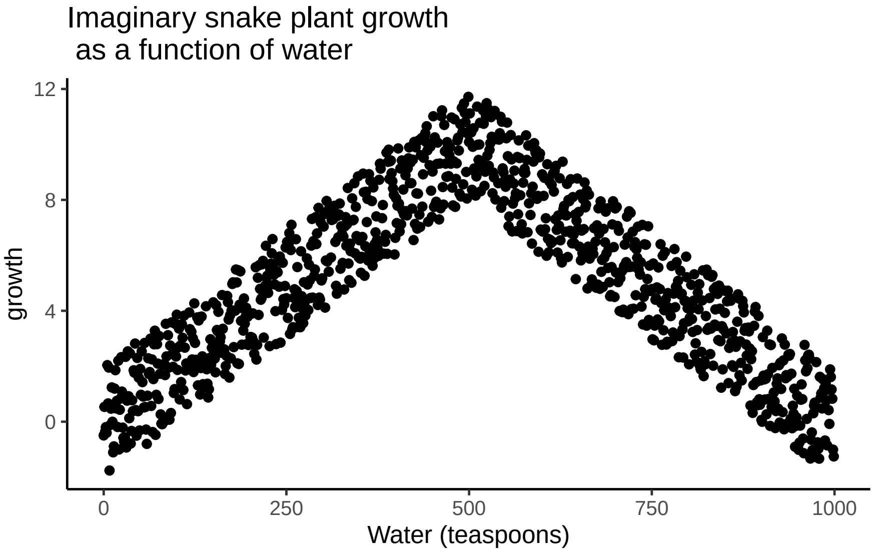


Figure 3.11: Illustration of a possible relationship between amount of water and snake plant growth. Growth goes up with water, but eventually goes back down as too much water makes snake plants die.

correlation, and if we look at all of the data we see no correlation. Yikes. So, even when there is a causal connection between two measures, we won't necessarily obtain clear evidence of the connection just by computing a correlation coefficient.

> Pro Tip: This is one reason why plotting your data is so important. If you see an upside U shape pattern, then a correlation analysis is probably not the best analysis for your data.

3.6.1.2 Confounding variable, or Third variable problem

Anybody can correlate any two things that can be quantified and measured. For example, we could find a hundred people, ask them all sorts of questions like:

1. how happy are you
2. how old are you
3. how tall are you
4. how much money do you make per year
5. how long are your eyelashes
6. how many books have you read in your life
7. how loud is your inner voice

Let's say we found a positive correlation between yearly salary and happiness. Note, we could have just as easily computed the same correlation between happiness and yearly salary. If we found a correlation, would you be willing to infer that yearly salary causes happiness? Perhaps it does play a small part. But, something like happiness probably has a lot of contributing causes. Money could directly cause some people to be happy. But, more likely, money buys people access to all sorts of things, and some of those things might contribute happiness. These "other" things are called **third** variables. For example, perhaps people living in nicer places in more expensive houses are more happy than people in worse places in cheaper houses. In this scenario, money isn't causing happiness, it's the places and houses that money buys. But, even is this were true, people can still be more or less happy in lots of different situations.

The lesson here is that a correlation can occur between two measures because of a third variable that is not directly measured. So, just because we find a correlation, does not mean we can conclude anything about a causal connection between two measurements.

3.6.2 Correlation and Random chance

Another very important aspect of correlations is the fact that they can be produced by random chance. This means that you can find a positive or negative correlation between two measures, even when they have absolutely nothing to do with one another. You might have hoped to find zero correlation when two measures are totally unrelated to each other. Although this certainly happens, unrelated measures can accidentally produce **spurious** correlations, just by chance alone.

Let's demonstrate how correlations can occur by chance when there is no causal connection between two measures. Imagine two participants. One is at the North pole with a lottery machine full of balls with numbers from 1 to 10. The other is at the south pole with a different lottery machine full of balls with numbers from 1 to 10. There are an endless supply of balls in the machine, so every number could be picked for any ball. Each participant randomly chooses 10 balls, then records the number on the ball. In this situation we will assume that there is no possible way that balls chosen by the first participant could causally influence the balls chosen by the second participant. They are on the other side of the world. We should assume that the balls will be chosen by chance alone.

Here is what the numbers on each ball could look like for each participant:

Ball	North_pole	South_pole
1	5	2
2	8	5
3	4	9
4	5	7
5	2	10
6	9	3
7	6	8
8	2	7
9	3	7
10	10	3

In this one case, if we computed Pearson's r, we would find that r = -0.6954454. But, we already know that this value does not tell us anything about the relationship between the balls chosen in the north and south pole. We know that relationship should be completely random, because that is how we set up the game.

The better question here is to ask what can random chance do? For example, if we ran our game over and over again thousands of times, each time choosing new balls, and each time computing the correlation, what would we find?First, we will find fluctuation. The r value will sometimes be positive, sometimes be negative, sometimes be big and sometimes be small. Second, we will see what the fluctuation looks like. This will give us a window into the kinds of correlations that chance alone can produce. Let's see what happens.

3.6.2.1 Monte-carlo simulation of random correlations

It is possible to use a computer to simulate our game as many times as we want. This process is often termed **monte-carlo simulation**.

Below is a script written for the programming language R. We won't go into the details of the code here. However, let's briefly explain what is going on. Notice, the part that says `for(sim in 1:1000)`. This creates a loop that repeats our game 1000 times. Inside the loop there are variables named `North_pole` and `South_pole`. During each simulation, we sample 10 random numbers (between 1 to 10) into each variable. These random numbers stand for the numbers that would have been on the balls from the lottery machine. Once we have 10 random numbers for each, we then compute the correlation using `cor(North_pole,South_pole)`. Then, we save the correlation value and move on to the next simulation. At the end, we will have 1000 individual Pearson r values.

```
simulated_correlations <- length(0)
for(sim in 1:1000){
  North_pole <- runif(10,1,10)
  South_pole <- runif(10,1,10)
  simulated_correlations[sim] <- cor(North_pole,South_pole)
}

sim_df <- data.frame(sims=1:1000,simulated_correlations)

ggplot(sim_df, aes(x = sims, y = simulated_correlations))+
  geom_point()+
  theme_classic()+
  geom_hline(yintercept = -1)+
  geom_hline(yintercept = 1)+
  ggtitle("Simulation of 1000 r values")
```

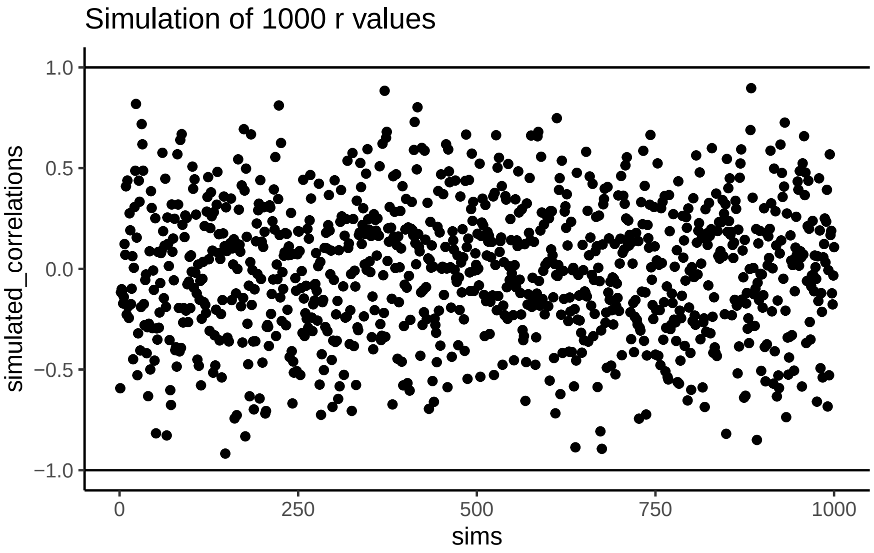


Figure 3.12: Another figure showing a range of r-values that can be obtained by chance.

Figure **??** shows the 1000 Pearson r values from the simulation. Does the figure below look familiar to you? We have already conducted a similar kind of simulation before. Each dot in the scatter plot shows the Pearson r for each simulation from 1 to 1000. As you can see the dots are all over of the place, in between the range -1 to 1. The important lesson here is that random chance produced all of these correlations. This means we can find "correlations" in the data that are completely meaningless, and do not reflect any causal relationship between one measure and another.

Let's illustrate the idea of finding "random" correlations one more time, with a little movie. This time, we will show you a scatter plot of the random values sampled for the balls chosen from the North and South pole. If there is no relationship we should see dots going everywhere. If there happens to be a positive relationship (purely by chance), we should see the dots going from the bottom left to the top right. If there happens to be a negative relationship (purely by chance), we should see the dots going from the top left down to the bottom right.

On more thing to prepare you for the movie. There are three scatter plots below in Figure **??**, showing negative, positive, and zero correlations between two variables. You've already seen this graph before. We are just reminding you that the blue lines are helpful for seeing the correlation.Negative correlations occur when a line goes down from the top left to bottom right. Positive correlations occur when a line goes up from the bottom left to the top right. Zero correlations occur when the line is flat (doesn't go up or down).

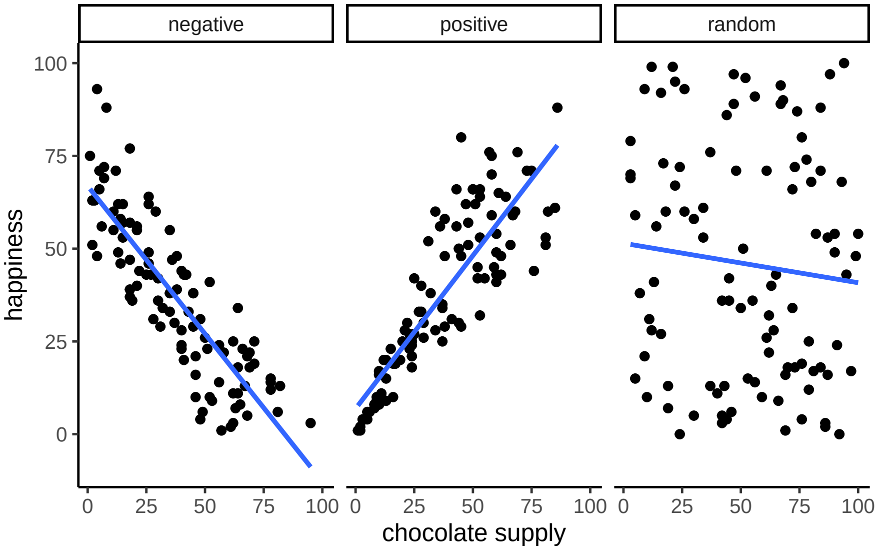


Figure 3.13: A reminder of what positive, negative, and zero correlation looks like

OK, now we are ready for the movie. **?@fig-3randcor10gif** shows the process of sampling two sets of numbers randomly, one for the X variable, and one for the Y variable. Each time we sample 10 numbers for each, plot them, then draw a line through them. Remember, these numbers are all completely random, so we should expect, on average that there should be no correlation between the numbers. However, this is not what happens. You can the line going all over the place. Sometimes we find a negative correlation (line goes down), sometimes we see a positive correlation (line goes up), and sometimes it looks like zero correlation (line is more flat).

You might be thinking this is kind of disturbing. If we know that there should be no correlation between two random variables, how come we are finding correlations? This is a big problem right? I mean, if someone showed me a correlation between two things, and then claimed one thing was related to another, how could know I if it was true. After all, it could be chance! Chance can do that too.

Fortunately, all is not lost. We can look at our simulated data in another way, using a histogram. Remember, just before the movie, we simulated 1000 different correlations using random numbers. By, putting all of those r values into a histogram, we can get a better sense of how chance behaves. We can see what kind of correlations chance is likely or unlikely to produce. Figure **??** is a histogram of the simulated r values.

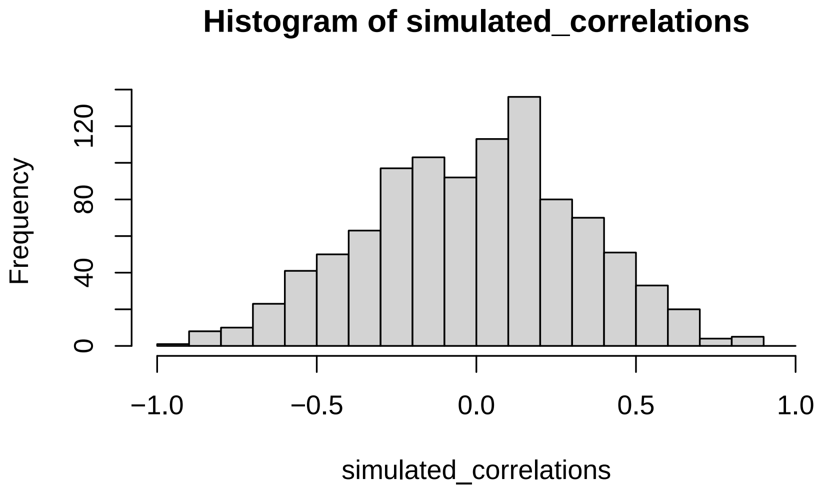


Figure 3.14: A histogram showing the frequency distribution of r-values for completely random values between an X and Y variable (sample-size=10). A rull range of r-values can be obtained by chance alone. Larger r-values are less common than smaller r-values

Notice that this histogram is not flat. Most of the simulated r values are close to zero. Notice, also that the bars get smaller as you move away from zero in the positive or negative direction. The general take home here is that chance can produce a wide range of correlations. However, not all correlations happen very often. For example, the bars for -1 and 1 are very small. Chance does not produce nearly perfect correlations very often. The bars around -.5 and .5 are smaller than the bars around zero, as medium correlations do not occur as often as small correlations by chance alone.

You can think of this histogram as the window of chance. It shows what chance often does, and what it often does not do. If you found a correlation under these very same circumstances (e.g., measured the correlation between two sets of 10 random numbers), then you could consult this window. What should you ask the window? How about, could my observed correlation (the one that you found in your data) have come from this window. Let's say you found a correlation of $r = .1$. Could a .1 have come from the histogram? Well, look at the histogram around where the .1 mark on the x-axis is. Is there a big bar there? If so, this means that chance produces this value fairly often. You might be comfortable with the inference: Yes, this

.1 could have been produced by chance, because it is well inside the window of chance. How about $r = .5$? The bar is much smaller here, you might think, "well, I can see that chance does produce .5 some times, so chance could have produced my .5. Did it? Maybe, maybe not, not sure". Here, your confidence in a strong inference about the role of chance might start getting a bit shakier.

How about an $r = .95$?. You might see that the bar for .95 is very very small, perhaps too small to see. What does this tell you? It tells you that chance does not produce .95 very often, hardly if at all, pretty much never. So, if you found a .95 in your data, what would you infer? Perhaps you would be comfortable inferring that chance did not produce your .95, after .95 is mostly outside the window of chance.

3.6.2.2 Increasing sample-size decreases opportunity for spurious correlation

Before moving on, let's do one more thing with correlations. In our pretend lottery game, each participant only sampled 10 balls each. We found that this could lead to a range of correlations between the numbers randomly drawn from either sides of the pole. Indeed, we even found some correlations that were medium to large in size. If you were a researcher who found such correlations, you might be tempted to believe there was a relationship between your measurements. However, we know in our little game, that those correlations would be spurious, just a product of random sampling.

The good news is that, as a researcher, you get to make the rules of the game. You get to determine how chance can play. This is all a little bit metaphorical, so let's make it concrete.

We will see what happens in four different scenarios. First, we will repeat what we already did. Each participant will draw 10 balls, then we compute the correlation, and do this over 1000 times and look at a histogram. Second, we will change the game so each participant draws 50 balls each, and then repeat our simulation. Third, and fourth, we will change the game so each participant draws 100 balls each, and then 1000 balls each, and repeat etc.

Figure **??** shows four different histograms of the Pearson r values in each of the different scenarios. Each scenario involves a different sample-size, from, 10, 50, 100 to 1000.

By inspecting the four histograms you should notice a clear pattern. The width or range of each histogram shrinks as the sample-size increases. What is going on here? Well, we already know that we can think of these histograms as windows of chance. They tell us which r values occur fairly often, which do not. When our sample-size is 10, lots of different r values happen. That histogram is very flat and spread out. However, as the sample-size increases, we see that the window of chance gets pulled in. For example, by the time we get to 1000 balls each, almost all of the Pearson r values are very close to 0.

One take home here, is that increasing sample-size narrows the window of chance. So, for example, if you ran a study involving 1000 samples of two measures, and you found a correlation of .5, then you can clearly see in the bottom right histogram that .5 does not occur very often

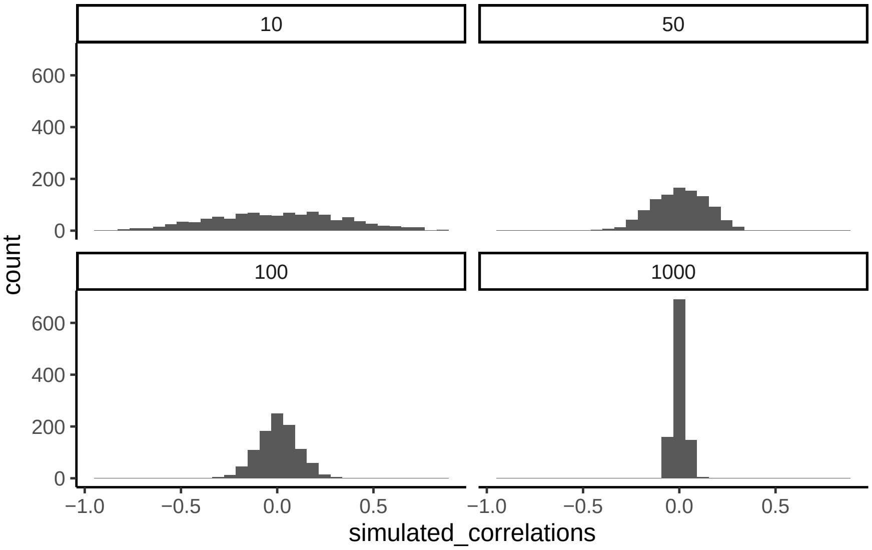


Figure 3.15: Four histograms showing the frequency distributions of r-values between completely random X and Y variables as a function of sample-size. The width of the distributions shrink as sample-size increases. Smaller sample-sizes are more likely to produce a wider range of r-values by chance. Larger sample-sizes always produce a narrow range of small r-values

by chance alone. In fact, there is no bar, because it didn't happen even once in the simulation. As a result, when you have a large sample size like n = 1000, you might be more confident that your observed correlation (say of .5) was not a spurious correlation. If chance is not producing your result, then something else is.

Finally, notice how your confidence about whether or not chance is mucking about with your results depends on your sample size. If you only obtained 10 samples per measurement, and found $r = .5$, you should not be as confident that your correlation reflects a real relationship. Instead, you can see that r's of .5 happen fairly often by chance alone.

> Pro tip: when you run an experiment you get to decide how many samples you will collect, which means you can choose to narrow the window of chance. Then, if you find a relationship in the data you can be more confident that your finding is real, and not just something that happened by chance.

3.6.3 Some more movies

Let's ingrain these idea with some more movies. When our sample-size is small (N is small), sampling error can cause all sort "patterns" in the data. This makes it possible, and indeed common, for "correlations" to occur between two sets of numbers. When we increase the sample-size, sampling error is reduced, making it less possible for "correlations" to occur just by chance alone. When N is large, chance has less of an opportunity to operate.

3.6.3.1 Watching how correlation behaves when there is no correlation

Below we randomly sample numbers for two variables, plot them, and show the correlation using a line. There are four panels, each showing the number of observations in the samples, from 10, 50, 100, to 1000 in each sample.

Remember, because we are randomly sampling numbers, there should be no relationship between the X and Y variables. But, as we have been discussing, because of chance, we can sometimes observe a correlation (due to chance). The important thing to watch is how the line behaves across the four panels in **?@fig-3corRandfour**. The line twirls around in all directions when the sample size is 10. It is also moves around quite a bit when the sample size is 50 or 100. It still moves a bit when the sample size is 1000, but much less. In all cases we expect that the line should be flat, but every time we take new samples, sometimes the line shows us pseudo patterns.

Which line should you trust? Well, hopefully you can see that the line for 1000 samples is the most stable. It tends to be very flat every time, and it does not depend so much on the particular sample. The line with 10 observations per sample goes all over the place. The take home here, is that if someone told you that they found a correlation, you should want to know how many observations they hand in their sample. If they only had 10 observations, how could you trust the claim that there was a correlation? You can't!!! Not now that you know samples that are that small can do all sorts of things by chance alone. If instead, you found out the sample was very large, then you might trust that finding a little bit more. For example, in the above movie you can see that when there are 1000 samples, we never see a strong or weak correlation; the line is always flat. This is because chance almost never produces strong correlations when the sample size is very large.

In the above example, we sampled numbers random numbers from a uniform distribution. Many examples of real-world data will come from a normal or approximately normal distribution. We can repeat the above, but sample random numbers from the same normal distribution. There will still be zero actual correlation between the X and Y variables, because everything is sampled randomly. **?@fig-3normCorfour** shows the same behavior. The computed correlation for small sample-sizes fluctuate wildly, and large sample sizes do not.

OK, so what do things look like when there actually is a correlation between variables?

3.6.3.2 Watching correlations behave when there really is a correlation

Sometimes there really are correlations between two variables that are not caused by chance. **?@fig-3realcorFour** shows a movie of four scatter plots. Each shows the correlation between two variables. Again, we change the sample-size in steps of 10, 50 100, and 1000. The data have been programmed to contain a real positive correlation. So, we should expect that the line will be going up from the bottom left to the top right. However, there is still variability in the data. So this time, sampling error due to chance will fuzz the correlation. We know it is there, but sometimes chance will cause the correlation to be eliminated.

Notice that in the top left panel (sample-size = 10), the line is twirling around much more than the other panels. Every new set of samples produces different correlations. Sometimes, the line even goes flat or downward. However, as we increase sample-size, we can see that the line doesn't change very much, it is always going up showing a positive correlation.

The main takeaway here is that even when there is a positive correlation between two things, you might not be able to see it if your sample size is small. For example, you might get unlucky with the one sample that you measured. Your sample could show a negative correlation, even when the actual correlation is positive! Unfortunately, in the real world we usually only have the sample that we collected, so we always have to wonder if we got lucky or unlucky. Fortunately, if you want to remove luck, all you need to do is collect larger samples. Then you will be much more likely to observe the real pattern, rather the pattern that can be introduced by chance.

3.7 Summary

In this section we have talked about correlation, and started to build some intuitions about **inferential statistics**, which is the major topic of the remaining chapters. For now, the main ideas are:

1. We can measure relationships in data using things like correlation
2. The correlations we measure can be produced by numerous things, so they are hard to to interpret
3. Correlations can be produced by chance, so have the potential to be completely meaningless.
4. However, we can create a model of exactly what chance can do. The model tells us whether chance is more or less likely to produce correlations of different sizes
5. We can use the chance model to help us make decisions about our own data. We can compare the correlation we found in our data to the model, then ask whether or not chance could have or was likely to have produced our results.

4 Probability, Sampling, and Estimation

Sections 4.1 & 4.9 - Adapted text by Danielle Navarro Section 4.10 - 4.11 & 4.13 - Mix of Matthew Crump & Danielle Navarro Section 4.12 - 4.13 - Adapted text by Danielle Navarro, all sections modified by Mallory Barnes.

> I have studied many languages-French, Spanish and a little Italian, but no one told me that Statistics was a foreign language. —Charmaine J. Forde

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 summarize a data set. To a lot of people, this is all there is to statistics: it's about calculating averages, collecting all the numbers, 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 tools **that let 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 voters at random, and 230 (23%) of those claimed that they intended to vote for the party. For the 2010 Federal election, the Australian Electoral Commission reported 4,610,795 enrolled voters in New South Whales; 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 primary vote is somewhere between 230/4610795 (about 0.005%) and 4610025/4610795 (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 do so. 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, our tools for making statistical inferences are 1) built on top of **probability theory**, and 2) require an understanding of how samples behave when you take them from distributions (defined by probability theory…). So, this chapter has two main parts. A brief introduction to probability theory, and an introduction to sampling from distributions.

4.1 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 two six sided dice, how likely is it that I'll roll two sixes?
- How likely is it that five cards drawn from a perfectly shuffled deck will all be hearts?
- What are the chances that I'll win the lottery?

Notice that all of these questions have something in common. In each case the "truth of the world" is known, and my question relates to the "what kind of events" will happen. In the first question I **know** that the coin is fair, so there's a 50% chance that any individual coin flip will come up heads. In the second question, I **know** that the chance of rolling a 6 on a single die is 1 in 6. In the third question I **know** that the deck is shuffled properly. And in the fourth question, I **know** that the lottery follows specific rules. You get the idea. The critical point is that probabilistic questions start with a known *model* of the world, and we use that model to do some calculations.

The underlying model can be quite simple. For instance, in the coin flipping example, we can write down the model like this: $P(\text{heads}) = 0.5$ which you can read as "the probability of heads is 0.5".

As we'll see later, in the same way that percentages are numbers that range from 0% to 100%, probabilities are just numbers that range from 0 to 1. When using this probability model to answer the first question, I don't actually know exactly what's going to happen. Maybe I'll get 10 heads, like the question says. But maybe I'll get three heads. That's the key thing: in probability theory, the **model** is known, but the **data** are not.

So that's probability. What about statistics? Statistical questions work the other way around. In statistics, we know the truth about the world. All we have is the data, and it is from the data that we want to **learn** the truth about the world. Statistical questions tend to look more like these:

- If my friend flips a coin 10 times and gets 10 heads, are they playing a trick on me?
- If five cards off the top of the deck are all hearts, how likely is it that the deck was shuffled?
- If the lottery commissioner's spouse wins the lottery, how likely is it that the lottery was rigged?

This time around, the only thing we have are data. What I **know** is that I saw my friend flip the coin 10 times and it came up heads every time. And what I want to *infer* is whether or not I should conclude that what I just saw was actually a fair coin being flipped 10 times in a row, or whether I should suspect that my friend is playing a trick on me. The data I have look like this:

```
H H H H H H H H H H
```

and what I'm trying to do is work out which "model of the world" I should put my trust in. If the coin is fair, then the model I should adopt is one that says that the probability of heads is 0.5; that is, $P(\text{heads}) = 0.5$. If the coin is not fair, then I should conclude that the probability of heads is **not** 0.5, which we would write as $P(\text{heads}) \neq 0.5$. In other words, the statistical inference problem is to figure out which of these probability models is right. Clearly, the statistical question isn't the same as the probability question, but they're deeply connected to one another. Because of this, a good introduction to statistical theory will start with a discussion of what probability is and how it works.

4.2 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. If you had to explain "probability" to a five year old, you could do a pretty good job. But if you've ever had that experience in real life, you might walk away from the conversation feeling like you didn't quite get it right, and that (like many everyday concepts) it turns out that you don't **really** know what it's all about.

So I'll have a go at it. Let's suppose I want to bet on a soccer game between two teams of robots, **Arduino Arsenal** and **C Milan**. After thinking about it, I decide that there is an 80% probability that **Arduino Arsenal** winning. What do I mean by that? Here are three possibilities…

- They're robot teams, so I can make them play over and over again, and if I did that, **Arduino Arsenal** would win 8 out of every 10 games on average.
- For any given game, I would only agree that betting on this game is only "fair" if a $1 bet on **C Milan** gives a $5 payoff (i.e. I get my $1 back plus a $4 reward for being correct), as would a $4 bet on **Arduino Arsenal** (i.e., my $4 bet plus a $1 reward).
- My subjective "belief" or "confidence" in an **Arduino Arsenal** victory is four times as strong as my belief in a **C Milan** victory.

Each of these seems sensible. However they're not identical, and not every statistician would endorse all of them. The reason is that there are different statistical ideologies (yes, really!) and depending on which one you subscribe to, you might say that some of those statements are meaningless or irrelevant. In this section, I give a brief introduction the two main approaches that exist in the literature. These are by no means the only approaches, but they're the two big ones.

4.2.1 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 N_H/N every time. Here's what I'd get (I did literally flip coins to produce this!):

number of flips	1	2	3	4	5	6	7	8	9	10
number of heads	0	1	2	3	4	4	4	5	6	7
proportion	.00	.50	.67	.75	.80	.67	.57	.63	.67	.70

number of flips	11	12	13	14	15	16	17	18	19	20
number of heads	8	8	9	10	10	10	10	10	10	11
proportion	.73	.67	.69	.71	.67	.63	.59	.56	.53	.55

Notice that at the start of the sequence, the **proportion** of heads fluctuates wildly, starting at .00 and rising as high as .80. Later on, one gets the impression that it dampens out a bit, with more and more of the values actually being pretty close to the "right" answer of .50. This is the frequentist definition of probability in a nutshell: flip a fair coin over and over again, and as N grows large (approaches infinity, denoted $N \to \infty$), the proportion of heads will converge to 50%. There are some subtle technicalities that the mathematicians care about, but qualitatively speaking, that's how the frequentists define probability. Unfortunately, I don't have an infinite number of coins, or the infinite patience required to flip a coin an infinite number of times. However, I do have a computer, and computers excel at mindless repetitive tasks. So I asked my computer to simulate flipping a coin 1000 times, and then drew a picture of what happens to the proportion N_H/N as N increases. Actually, I did it four times, just to make sure it wasn't a fluke. The results are shown in Figure **??**. 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. Second, 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. 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.

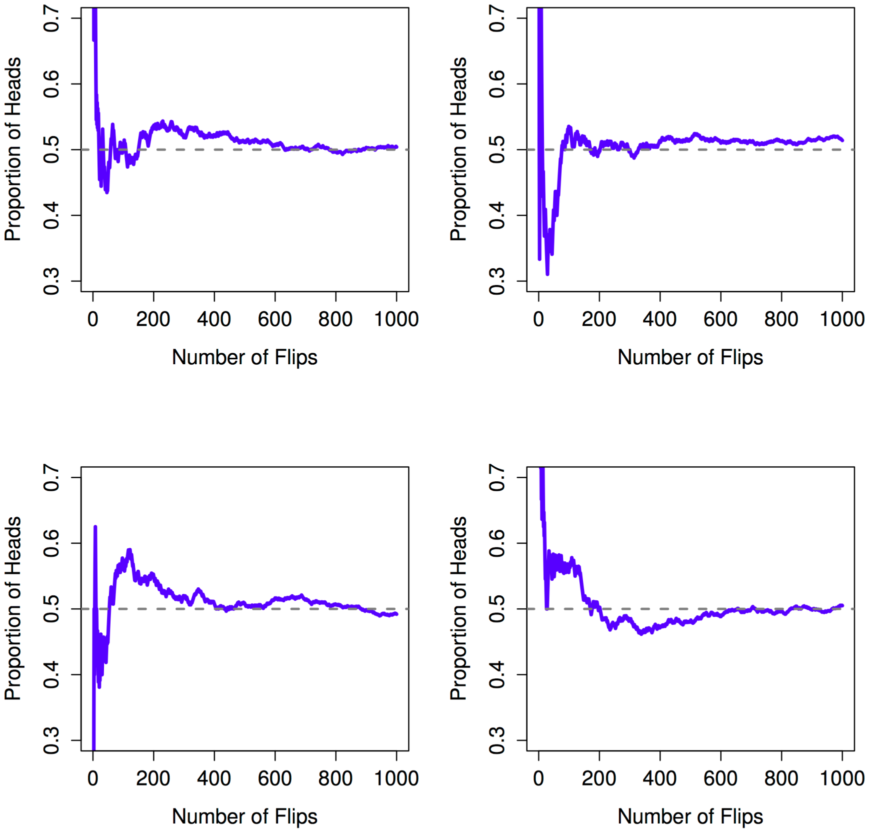


Figure 4.1: An illustration of how frequentist probability works. If you flip a fair coin over and over again, the proportion of heads that you've seen eventually settles down, and converges to the true probability of 0.5. Each panel shows four different simulated experiments: in each case, we pretend we flipped a coin 1000 times, and kept track of the proportion of flips that were heads as we went along. Although none of these sequences actually ended up with an exact value of .5, if we'd extended the experiment for an infinite number of coin flips they would have.

More seriously, the frequentist definition has a narrow scope. There are lots of things out there that human beings are happy to assign probability to in everyday language, but cannot (even in theory) be mapped onto a hypothetical sequence of events. For instance, if a meteorologist comes on TV and says, “the probability of rain in Adelaide on 2 November 2048 is 60%” we humans are happy to accept this. But it’s not clear how to define this in frequentist terms. There’s only one city of Adelaide, and only 2 November 2048. There’s no infinite sequence of events here, just a once-off thing. Frequentist probability genuinely **forbids** us from making probability statements about a single event. From the frequentist perspective, it will either rain tomorrow or it will not; there is no “probability” that attaches to a single non-repeatable event. Now, it should be said that there are some very clever tricks that frequentists can use to get around this. One possibility is that what the meteorologist means is something like this: “There is a category of days for which I predict a 60% chance of rain; if we look only across those days for which I make this prediction, then on 60% of those days it will actually rain”. It’s very weird and counter intuitive to think of it this way, but you do see frequentists do this sometimes.

4.2.2 The Bayesian view

The **Bayesian view** of probability is often called the subjectivist view, and it is a minority view among statisticians, but one that has been steadily gaining traction for the last several decades. There are many flavors of Bayesianism, making hard to say exactly what “the” Bayesian view is. The most common way of thinking about subjective probability is to define the probability of an event as the **degree of belief** that an intelligent and rational agent assigns to that truth of that event. From that perspective, probabilities don’t exist in the world, but rather in the thoughts and assumptions of people and other intelligent beings. However, in order for this approach to work, we need some way of operationalising “degree of belief”. One way that you can do this is to formalize 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: if it rains tomorrow, then I win $5, but if it doesn’t rain then I lose $5. Clearly, from my perspective, this is a pretty good bet. On the other hand, if I think that the probability of rain is only 40%, then it’s a bad bet to take. Thus, we can operationalize the notion of a “subjective probability” in terms of what bets I’m willing to accept.

What are the advantages and disadvantages to the Bayesian approach? The main advantage is that it allows you to assign probabilities to any event you want to. You don’t need to be limited to those events that are repeatable. The main disadvantage (to many people) is that we can’t be purely objective – specifying a probability requires us to specify an entity that has the relevant degree of belief. This entity might be a human, an alien, a robot, or even a statistician, but there has to be an intelligent agent out there that believes in things. To many people this is uncomfortable: it seems to make probability arbitrary. While the Bayesian approach does require that the agent in question be rational (i.e., obey the rules of probability), it does allow everyone to have their own beliefs; I can believe the coin is fair and you don't have to, even though we're both rational. The frequentist view doesn't allow any two observers to attribute different probabilities to the same event: when that happens, then at least one of them must be wrong. The Bayesian view does not prevent this from occurring. Two observers with different background knowledge can legitimately hold different beliefs about the same event. In short, where the frequentist view is sometimes considered to be too narrow (forbids lots of things that that we want to assign probabilities to), the Bayesian view is sometimes thought to be too broad (allows too many differences between observers).

4.2.3 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 do? Would some of these statements be meaningless to a frequentist or a Bayesian? If you've understood the two perspectives, you should have some sense of how to answer those questions.

Okay, assuming you understand the different, you might be wondering which of them is **right**? Honestly, I don't know that there is a right answer. As far as I can tell there's nothing mathematically incorrect about the way frequentists think about sequences of events, and there's nothing mathematically incorrect about the way that Bayesians define the beliefs of a rational agent. In fact, when you dig down into the details, Bayesians and frequentists actually agree about a lot of things. Many frequentist methods lead to decisions that Bayesians agree a rational agent would make. Many Bayesian methods have very good frequentist properties.

For the most part, I'm a pragmatist so I'll use any statistical method that I trust. As it turns out, that makes me prefer Bayesian methods, for reasons I'll explain towards the end of the book, but I'm not fundamentally opposed to frequentist methods. Not everyone is quite so relaxed. For instance, consider Sir Ronald Fisher, one of the towering figures of 20th century statistics and a vehement opponent to all things Bayesian, whose paper on the mathematical foundations of statistics referred to Bayesian probability as "an impenetrable jungle [that] arrests progress towards precision of statistical concepts" Fisher (1922, 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, 114). The history of statistics, as you might gather, is not devoid of entertainment.

4.3 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 pants.

4.3.1 Introducing probability distributions

One of the disturbing truths about my life is that I only own 5 pairs of pants: 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 do: that's why they call me Mister Imaginative. Now, on any given day, I pick out exactly one of pair of pants to wear. Not even I'm so stupid as to try to wear two pairs of pants, and thanks to years of training I never go outside without wearing pants anymore. If I were to describe this situation using the language of probability theory, I would refer to each pair of pants (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 pants), then the outcome will be one and only one of these events. Like I said, these days I always wear exactly one pair of pants, so my pants 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 pants in probabilistic terms. Sad.

Okay, now that we have a sample space (a wardrobe), which is built from lots of possible elementary events (pants), 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 pants). On the other hand, if $P(X) = 1$ it means that event X is certain to occur (i.e., I always wear those pants). For probability values in the middle, it means that I sometimes wear those pants. For instance, if $P(X) = 0.5$ it means that I wear those pants half of the time.

At this point, we're almost done. The last thing we need to recognize is that "something always happens". Every time I put on pants, I really do end up wearing pants (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, this is an example of a probability distribution

Which pants?	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	$P(X_5) = .1$

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 to visualize this distribution, as shown in Figure **??**. 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 pants. Everyone wins!

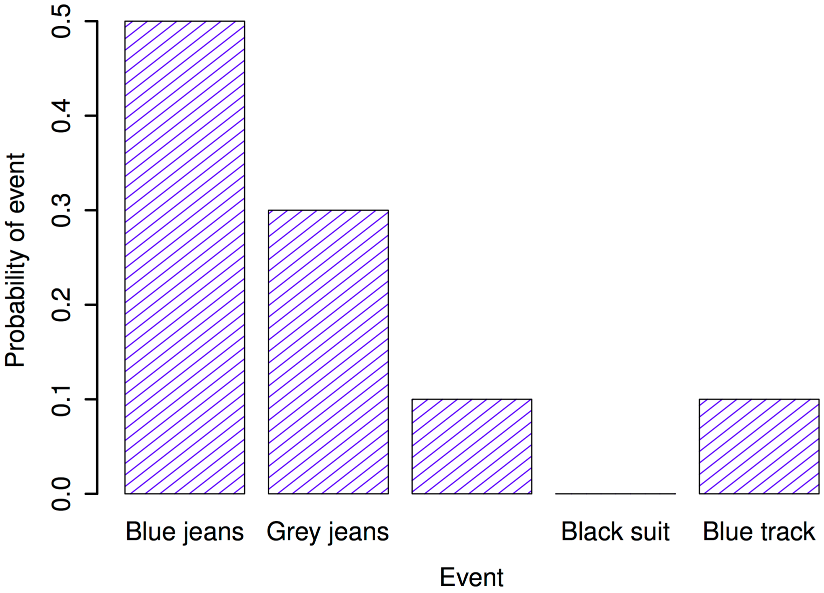


Figure 4.2: A visual depiction of the pants probability distribution. There are five elementary events, corresponding to the five pairs of pants that I own. Each event has some probability of occurring: this probability is a number between 0 to 1. The sum of these probabilities is 1.

The only other thing that I need to point out is that probability theory 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 pants example, it's perfectly legitimate to refer to the probability that I wear jeans. In this scenario, the "Dan wears jeans" event 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 (X_1, X_2, X_3). 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 $P(E)$ is: we just add everything up. In this particular case

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

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.

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 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 4.4: Some basic rules that probabilities must satisfy. You don't really need to know these rules in order to understand the analyses that we'll talk about later in the book, but they are important if you want to understand probability theory a bit more deeply.

	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	$P(A \cap B)$	$=$	$P(A\|B) P(B)$

Now that we have the ability to "define" non-elementary events in terms of elementary ones, we can actually use this to construct (or, if you want to be all mathematicallish, "derive") some of the other rules of probability. These rules are listed above, and while I'm pretty confident that very few of my readers actually care about how these rules are constructed, I'm going to show you anyway: even though it's boring and you'll probably never have a lot of use for these derivations, if you read through it once or twice and try to see how it works, you'll find that probability starts to feel a bit less mysterious, and with any luck a lot less daunting. So here goes. Firstly, in order to construct the rules I'm going to need a sample space X that consists of a bunch of elementary events x, and two non-elementary events, which I'll call A and B. Let's say:

$$\begin{array}{rcl} X &=& (x_1, x_2, x_3, x_4, x_5) \\ A &=& (x_1, x_2, x_3) \\ B &=& (x_3, x_4) \end{array}$$

To make this a bit more concrete, let's suppose that we're still talking about the pants distribution. If so, A corresponds to the event "jeans", and B corresponds to the event "black":

$$\begin{array}{rcl} \text{“jeans”} & = & (\text{“blue jeans”}, \text{“grey jeans”}, \text{“black jeans”}) \\ \text{“black”} & = & (\text{“black jeans”}, \text{“black suit”}) \end{array}$$

So now let's start checking the rules that I've listed in the table.

In the first line, the table says that

$$P(\neg A) = 1 - P(A)$$

and what it **means** is that the probability of "not A" is equal to 1 minus the probability of A. A moment's thought (and a tedious example) make it obvious why this must be true. If A corresponds to the even that I wear jeans (i.e., one of x_1 or x_2 or x_3 happens), then the only meaningful definition of "not A" (which is mathematically denoted as $\neg A$) is to say that $\neg A$ consists of **all** elementary events that don't belong to A. In the case of the pants distribution it means that $\neg A = (x_4, x_5)$, or, to say it in English: "not jeans" consists of all pairs of pants that aren't jeans (i.e., the black suit and the blue tracksuit). Consequently, every single elementary event belongs to either A or $\neg A$, but not both. Okay, so now let's rearrange our statement above:

$$P(\neg A) + P(A) = 1$$

which is a trite way of saying either I do wear jeans or I don't wear jeans: the probability of "not jeans" plus the probability of "jeans" is 1. Mathematically:

$$\begin{array}{rcl} P(\neg A) & = & P(x_4) + P(x_5) \\ P(A) & = & P(x_1) + P(x_2) + P(x_3) \end{array}$$

so therefore

$$\begin{array}{rcl} P(\neg A) + P(A) & = & P(x_1) + P(x_2) + P(x_3) + P(x_4) + P(x_5) \\ & = & \sum_{x \in X} P(x) \\ & = & 1 \end{array}$$

Excellent. It all seems to work.

Wow, I can hear you saying. That's a lot of xs to tell me the freaking obvious. And you're right: this **is** freaking obvious. The whole **point** of probability theory to to formalize and mathematize a few very basic common sense intuitions. So let's carry this line of thought forward a bit further. In the last section I defined an event corresponding to **not** A, which I denoted $\neg A$. Let's now define two new events that correspond to important everyday concepts: A **and** B, and A **or** B. To be precise:

English statement:	Mathematical notation:
"A and B" both happen	$A \cap B$
at least one of "A or B" happens	$A \cup B$

Since A and B are both defined in terms of our elementary events (the xs) we're going to need to try to describe $A \cap B$ and $A \cup B$ in terms of our elementary events too. Can we do this? Yes we can The only way that both A and B can occur is if the elementary event that we observe turns out to belong to both A and B. Thus "$A \cap B$" includes only those elementary events that belong to both A and B...

$$\begin{aligned} A &= (x_1, x_2, x_3) \\ B &= (x_3, x_4) \\ A \cap B &= (x_3) \end{aligned}$$

So, um, the only way that I can wear "jeans" (x_1, x_2, x_3) and "black pants" (x_3, x_4) is if I wear "black jeans" (x_3). Another victory for the bloody obvious.

At this point, you're not going to be at all shocked by the definition of $A \cup B$, though you're probably going to be extremely bored by it. The only way that I can wear "jeans" or "black pants" is if the elementary pants that I actually do wear belongs to A or to B, or to both. So...

$$\begin{aligned} A &= (x_1, x_2, x_3) \\ B &= (x_3, x_4) \\ A \cup B &= (x_1, x_2, x_3, x_4) \end{aligned}$$

Oh yeah baby. Mathematics at its finest.

So, we've defined what we mean by $A \cap B$ and $A \cup B$. Now let's assign probabilities to these events. More specifically, let's start by verifying the rule that claims that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Using our definitions earlier, we know that $A \cup B = (x_1, x_2, x_3, x_4)$, so

$$P(A \cup B) = P(x_1) + P(x_2) + P(x_3) + P(x_4)$$

and making similar use of the fact that we know what elementary events belong to A, B and $A \cap B$....

$$\begin{aligned} P(A) &= P(x_1) + P(x_2) + P(x_3) \\ P(B) &= P(x_3) + P(x_4) \\ P(A \cap B) &= P(x_3) \end{aligned}$$

and therefore

$$\begin{aligned} P(A) + P(B) - P(A \cap B) &= P(x_1) + P(x_2) + P(x_3) + P(x_3) + P(x_4) - P(x_3) \\ &= P(x_1) + P(x_2) + P(x_3) + P(x_4) \\ &= P(A \cup B) \end{aligned}$$

Done.

The next concept we need to define is the notion of "B given A", which is typically written $B|A$. Here's what I mean: suppose that I get up one morning, and put on a pair of pants. An elementary event x has occurred. Suppose further I yell out to my wife (who is in the other room, and so cannot see my pants) "I'm wearing jeans today!". Assuming that she believes that I'm telling the truth, she knows that A is true. **Given** that she knows that A has happened, what is the **conditional probability** that B is also true? Well, let's think about what she knows. Here are the facts:

- **The non-jeans events are impossible**. If A is true, then we know that the only possible elementary events that could have occurred are x_1, x_2 and x_3 (i.e.,the jeans). The non-jeans events x_4 and x_5 are now impossible, and must be assigned probability zero. In other words, our **sample space** has been restricted to the jeans events. But it's still the case that the probabilities of these these events **must** sum to 1: we know for sure that I'm wearing jeans.
- **She's learned nothing about which jeans I'm wearing**. Before I made my announcement that I was wearing jeans, she already knew that I was five times as likely to be wearing blue jeans ($P(x_1) = 0.5$) than to be wearing black jeans ($P(x_3) = 0.1$). My announcement doesn't change this... I said **nothing** about what color my jeans were, so it must remain the case that $P(x_1)/P(x_3)$ stays the same, at a value of 5.

There's only one way to satisfy these constraints: set the impossible events to have zero probability (i.e., $P(x|A) = 0$ if x is not in A), and then divide the probabilities of all the others by $P(A)$. In this case, since $P(A) = 0.9$, we divide by 0.9. This gives:

which pants?	elementary event	old prob, $P(x)$	new prob, $P(x\|A)$
blue jeans	x_1	0.5	0.556
grey jeans	x_2	0.3	0.333
black jeans	x_3	0.1	0.111
black suit	x_4	0	0
blue tracksuit	x_5	0.1	0

In mathematical terms, we say that

$$P(x|A) = \frac{P(x)}{P(A)}$$

if $x \in A$, and $P(x|A) = 0$ otherwise. And therefore...

$$\begin{aligned} P(B|A) &= P(x_3|A) + P(x_4|A) \\ &= \frac{P(x_3)}{P(A)} + 0 \\ &= \frac{P(x_3)}{P(A)} \end{aligned}$$

Now, recalling that $A \cap B = (x_3)$, we can write this as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

and if we multiply both sides by $P(A)$ we obtain:

$$P(A \cap B) = P(B|A)P(A)$$

which is the third rule that we had listed in the table.

4.4 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") 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.

4.4.1 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 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. 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

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

and we're interested in the special case where $X = 4$, $\theta = .167$ and $N = 20$. There's only one more piece of notation I want to refer to before moving on to discuss the solution to the problem. If I want to say that X is generated randomly from a binomial distribution with parameters θ and N, the notation I would use is as follows:

$$X \sim \text{Binomial}(\theta, N)$$

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. To that end, Figure Figure **??** 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 "pants probability" plot I drew in Figure **??**. 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 times 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.

4.4.2 Working with the binomial distribution in R

R has a function called `dbinom` that calculates binomial probabilities for us. The main arguments to the function are

- `x` This is a number, or vector of numbers, specifying the outcomes whose probability you're trying to calculate.
- `size` This is a number telling R the size of the experiment.
- `prob` This is the success probability for any one trial in the experiment.

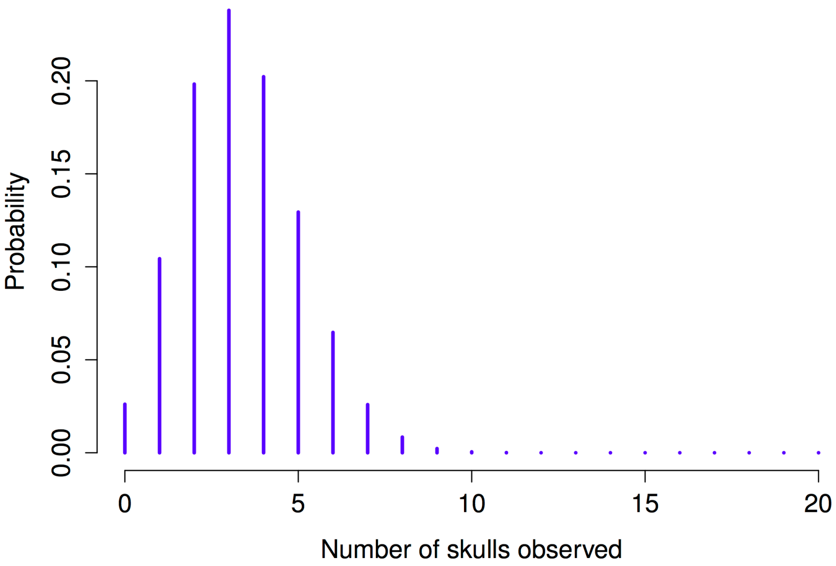


Figure 4.3: The binomial distribution with size parameter of N =20 and an underlying success probability of 1/6. Each vertical bar depicts the probability of one specific outcome (i.e., one possible value of X). Because this is a probability distribution, each of the probabilities must be a number between 0 and 1, and the heights of the bars must sum to 1 as well.

So, in order to calculate the probability of getting skulls, from an experiment of trials, in which the probability of getting a skull on any one trial is … well, the command I would use is simply this:

```
dbinom( x = 4, size = 20, prob = 1/6 )
#> [1] 0.2022036
```

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 $\theta = 1/2$. 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 **??** *a* 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 **??** *b*. The distribution stays roughly in the middle, but there's a bit more variability in the possible outcomes.

At this point, I should probably explain the name of the `dbinom` function. Obviously, the "binom" part comes from the fact that we're working with the binomial distribution, but the "d" prefix is probably a bit of a mystery. In this section I'll give a partial explanation: specifically, I'll explain why there is a prefix. As for why it's a "d" specifically, you'll have to wait until the next section. What's going on here is that R actually provides **four** functions in relation to the binomial distribution. These four functions are `dbinom`, `pbinom`, `rbinom` and `qbinom`, and each one calculates a different quantity of interest. Not only that, R does the same thing for **every** probability distribution that it implements. No matter what distribution you're talking about, there's a `d` function, a `p` function, `r` a function and a `q` function.

Let's have a look at what all four functions do. Firstly, all four versions of the function require you to specify the `size` and `prob` arguments: no matter what you're trying to get R to calculate, it needs to know what the parameters are. However, they differ in terms of what the other argument is, and what the output is. So let's look at them one at a time.

- The `d` form we've already seen: you specify a particular outcome `x`, and the output is the probability of obtaining exactly that outcome. (the "d" is short for *density*, but ignore that for now).
- The `p` form calculates the *cumulative probability*. You specify a particular quantile `q` , and it tells you the probability of obtaining an outcome **smaller than or equal to** `q`.
- The `q` form calculates the *quantiles* of the distribution. You specify a probability value `p`, and it gives you the corresponding percentile. That is, the value of the variable for which there's a probability `p` of obtaining an outcome lower than that value.

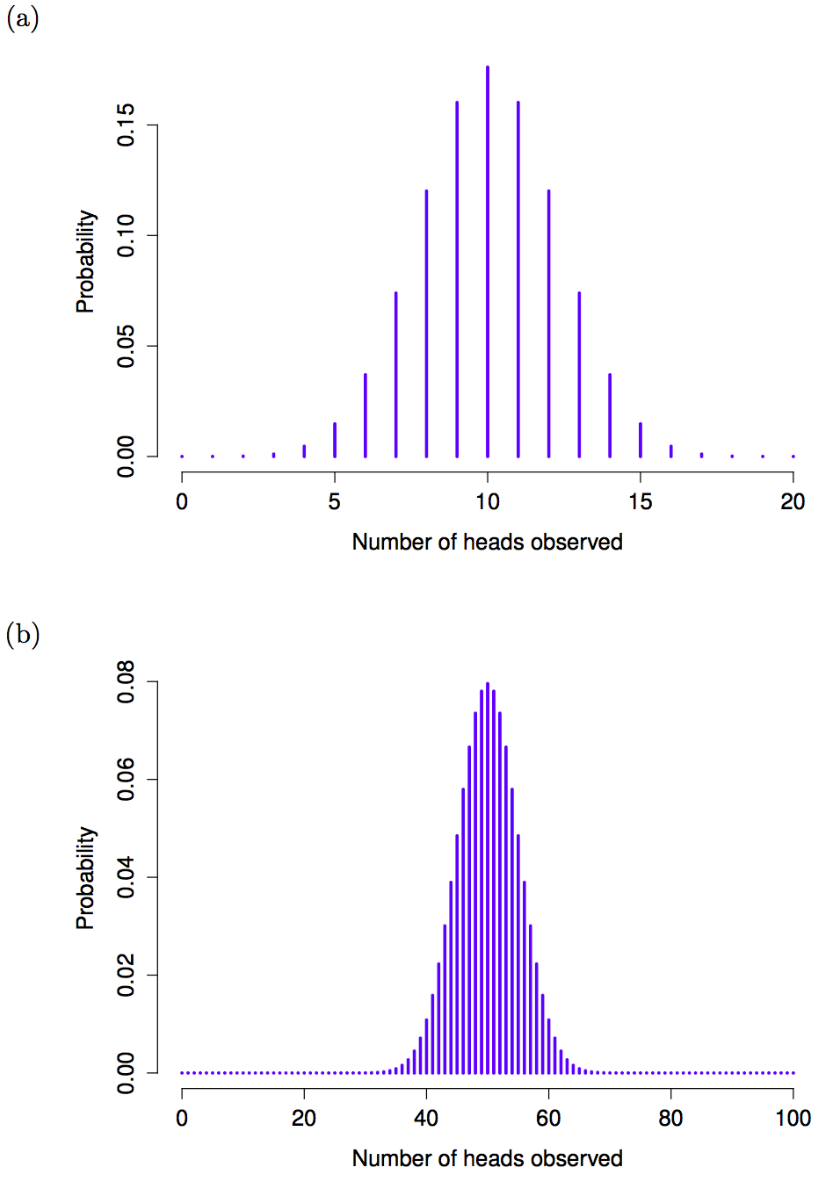


Figure 4.4: Two binomial distributions, involving a scenario in which I'm flipping a fair coin, so the underlying success probability is 1/2. In panel (a), we assume I'm flipping the coin N = 20 times. In panel (b) we assume that the coin is flipped N = 100 times.

- The `r` form is a *random number generator*: specifically, it generates `n` random outcomes from the distribution.

This is a little abstract, so let's look at some concrete examples. Again, we've already covered `dbinom` so let's focus on the other three versions. We'll start with `pbinom`, and we'll go back to the skull-dice example. Again, I'm rolling 20 dice, and each die has a 1 in 6 chance of coming up skulls. Suppose, however, that I want to know the probability of rolling 4 **or fewer** skulls. If I wanted to, I could use the `dbinom` function to calculate the exact probability of rolling 0 skulls, 1 skull, 2 skulls, 3 skulls and 4 skulls and then add these up, but there's a faster way. Instead, I can calculate this using the `pbinom` function. Here's the command:

```
pbinom( q= 4, size = 20, prob = 1/6)
#> [1] 0.7687492
```

In other words, there is a 76.9% chance that I will roll 4 or fewer skulls. Or, to put it another way, R is telling us that a value of 4 is actually the 76.9th percentile of this binomial distribution.

Next, let's consider the `qbinom` function. Let's say I want to calculate the 75th percentile of the binomial distribution. If we're sticking with our skulls example, I would use the following command to do this:

```
qbinom( p = 0.75, size = 20, prob = 1/6 )
#> [1] 4
```

Hm. There's something odd going on here. Let's think this through. What the `qbinom` function appears to be telling us is that the 75th percentile of the binomial distribution is 4, even though we saw from the function that 4 is **actually** the 76.9th percentile. And it's definitely the `pbinom` function that is correct. I promise. The weirdness here comes from the fact that our binomial distribution doesn't really **have** a 75th percentile. Not really. Why not? Well, there's a 56.7% chance of rolling 3 or fewer skulls (you can type `pbinom(3, 20, 1/6)` to confirm this if you want), and a 76.9% chance of rolling 4 or fewer skulls. So there's a sense in which the 75th percentile should lie "in between" 3 and 4 skulls. But that makes no sense at all! You can't roll 20 dice and get 3.9 of them come up skulls. This issue can be handled in different ways: you could report an in between value (or **interpolated** value, to use the technical name) like 3.9, you could round down (to 3) or you could round up (to 4).

The `qbinom` function rounds upwards: if you ask for a percentile that doesn't actually exist (like the 75th in this example), R finds the smallest value for which the the percentile rank is **at least** what you asked for. In this case, since the "true" 75th percentile (whatever that would mean) lies somewhere between 3 and 4 skulls, R Rounds up and gives you an answer of 4. This subtlety is tedious, I admit, but thankfully it's only an issue for discrete distributions like the binomial. The other distributions that I'll talk about (normal, t, χ^2 and F) are all continuous, and so R can always return an exact quantile whenever you ask for it.

Finally, we have the random number generator. To use the `rbinom` function, you specify how many times R should "simulate" the experiment using the `n` argument, and it will generate random outcomes from the binomial distribution. So, for instance, suppose I were to repeat my die rolling experiment 100 times. I could get R to simulate the results of these experiments by using the following command:

```
rbinom( n = 100, size = 20, prob = 1/6 )
#>   [1] 3 2 6 4 3 3 3 2 4 2 0 4 6 0 4 3 5 7 3 2 3 5 6 5 1 6 3 5 3 6 4 3 6 3 2 2 0
#>  [38] 2 3 4 2 5 1 5 4 4 5 2 1 3 0 3 4 0 5 3 1 2 5 7 3 4 6 1 2 2 6 6 6 2 4 3 3 6
#>  [75] 5 4 2 1 3 2 6 2 2 4 7 2 3 4 4 3 3 2 5 4 3 5 3 5 4 8
```

As you can see, these numbers are pretty much what you'd expect given the distribution shown in Figure **??** . Most of the time I roll somewhere between 1 to 5 skulls. There are a lot of subtleties associated with random number generation using a computer, but for the purposes of this book we don't need to worry too much about them.

4.5 The normal distribution

While the binomial distribution is conceptually the simplest distribution to understand, it's not the most important one. That particular honor goes to the *normal distribution*, which is 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:

$$X \sim \text{Normal}(\mu, \sigma)$$

Of course, that's just notation. It doesn't tell us anything interesting about the normal distribution itself. The mathematical formula for the normal distribution is:

The formula is 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 to much. Instead, we look at how R can be used to work with normal distributions. The R functions for the normal distribution are *dnorm()*, *pnorm()*, *qnorm()* and *rnorm()*. However, they behave in pretty much exactly the same way as the corresponding functions for the binomial distribution, so there's not a lot that you need to know. The only thing that I should point out is that the argument names for the parameters are *mean* and *sd*. In pretty much every other respect, there's nothing else to add.

Instead of focusing on the maths, 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 **??**, which plots a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. You can see where the name "bell curve"

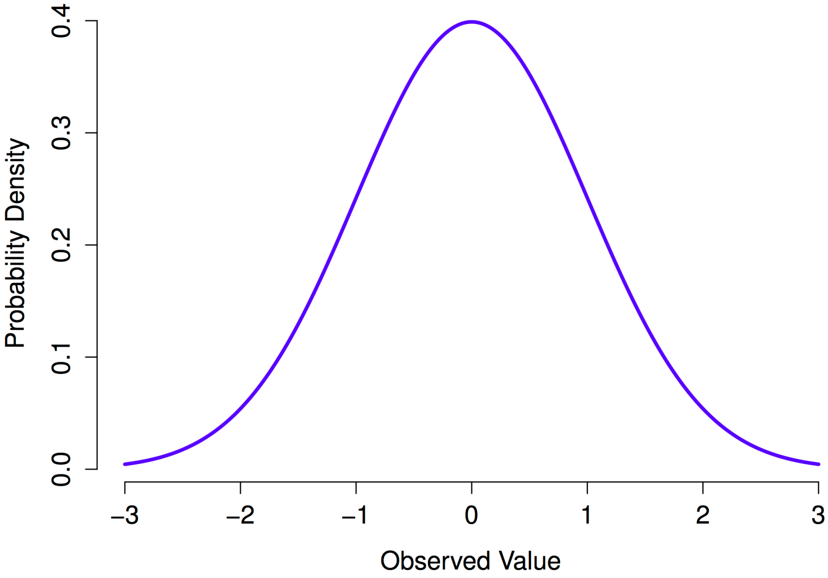


Figure 4.5: The normal distribution with mean = 0 and standard deviation = 1. The x-axis corresponds to the value of some variable, and the y-axis tells us something about how likely we are to observe that value. However, notice that the y-axis is labelled Probability Density and not Probability. There 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 isn't 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!).

$$\underline{\text{Normal}}$$

$$p(X \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$$

Figure 4.6: Formula for the normal distribution

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 Figure **??** 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.

With this in mind, let's see if we can 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. One parameter we can change is the mean. This will shift the distribution to the right or left. The animation in **?@fig-4normalMeanShift** shows a normal distribution with mean = 0, moving up and down from mean = 0 to mean = 5. Note, when you change the mean the whole shape of the distribution does not change, it just shifts from left to right. In the animation the normal distribution bounces up and down a little, but that's just a quirk of the animation (plus it looks fun that way).

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. The animation in **?@fig-4normalSDShift** shows what happens when you start with a small standard deviation (sd = 0.5), and move to larger and larger standard deviation (up to sd = 5). As you can see, the distribution spreads out and becomes wider as the standard deviation increases.

Notice 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.

4.5.1 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: 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 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. Maybe you're wondering why R uses the "d" prefix for functions like *dnorm().* And maybe, just maybe, you've been playing around with the *dnorm()* function, and you accidentally typed in a command like this:

```
dnorm( x = 1, mean = 1, sd = 0.1 )
#> [1] 3.989423
```

And if you've done the last part, you're probably very confused. I've asked R to calculate the probability that *x = 1*, for a normally distributed variable with *mean = 1* and standard deviation *sd = 0.1*; and it tells me that the probability is 3.99. But, as we discussed earlier, probabilities *can't* be larger than 1. So either I've made a mistake, or that's not a probability.

As it turns out, the second answer is correct. What we've calculated 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 something. 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 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".

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 I haven't actually explained what the *dnorm()* function actually calculates. Equivalently, 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.

4.6 Other useful distributions

There are many other useful distributions, these include the `t` distribution, the `F` distribution, and the chi squared distribution. We will soon discover more about the `t` and `F` distributions when we discuss t-tests and ANOVAs in later chapters.

4.7 Summary of Probability

We've talked 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 talking about some of the more important probability distributions that statisticians work with. We talked about things like this:

- Probability theory versus statistics
- Frequentist versus Bayesian views of probability
- Basics of probability theory
- Binomial distribution, normal distribution

As you'd expect, this 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.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 statistics classes 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. Many academics 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/frequentist distinction.

4.8 Samples, populations and sampling

Remember, the role of descriptive statistics is to concisely summarize what we **do** know. In contrast, the purpose of inferential statistics is to “learn what we do not know from what we do”. 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 we’ll talk about sampling theory first because estimation theory doesn’t make sense until you understand sampling. So, this chapter divides into sampling theory, and how to make use of sampling theory to discuss how statisticians think about estimation. We have already done lots of sampling, so you are already familiar with some of the big ideas.

Sampling theory plays a huge role in specifying the assumptions upon which your statistical inferences rely. And in order to talk about “making inferences” the way statisticians think about it, we need to be a bit more explicit about what it is that we’re drawing inferences **from** (the sample) and what it is that we’re drawing inferences **about** (the population).

In almost every situation of interest, what we have available to us as researchers is a **sample** of data. We might have run experiment with some number of participants; a polling company might have phoned some number of people to ask questions about voting intentions; etc. Regardless: the data set available to us is finite, and incomplete. We can’t possibly get every person in the world to do our experiment; a polling company doesn’t have the time or the money to ring up every voter in the country etc. In our earlier discussion of descriptive statistics, this sample was the only thing we were interested in. Our only goal was to find ways of describing, summarizing and graphing that sample. This is about to change.

4.8.1 Defining a population

A sample is a concrete thing. You can open up a data file, and there’s the data from your sample. A **population**, on the other hand, is a more abstract idea. It refers to the set of all possible people, or all possible observations, that you want to draw conclusions about, and is generally **much** bigger than the sample. In an ideal world, the researcher would begin the study with a clear idea of what the population of interest is, since the process of designing a study and testing hypotheses about the data that it produces does depend on the population about which you want to make statements. However, that doesn't always happen in practice: usually the researcher has a fairly vague idea of what the population is and designs the study as best he/she can on that basis.

Sometimes it's easy to state the population of interest. For instance, in the "polling company" example, the population consisted of all voters enrolled at the a time of the study – millions of people. The sample was a set of 1000 people who all belong to that population. In most situations the situation is much less simple. In a typical a psychological experiment, determining the population of interest is a bit more complicated. Suppose I run an experiment using 100 undergraduate students as my participants. My goal, as a cognitive scientist, is to try to learn something about how the mind works. So, which of the following would count as "the population":

- All of the undergraduate psychology students at the University of Adelaide?
- Undergraduate psychology students in general, anywhere in the world?
- Australians currently living?
- Australians of similar ages to my sample?
- Anyone currently alive?
- Any human being, past, present or future?
- Any biological organism with a sufficient degree of intelligence operating in a terrestrial environment?
- Any intelligent being?

Each of these defines a real group of mind-possessing entities, all of which might be of interest to me as a cognitive scientist, and it's not at all clear which one ought to be the true population of interest.

4.8.2 Simple random samples

Irrespective of how we 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, imagine 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 colors, black and white.

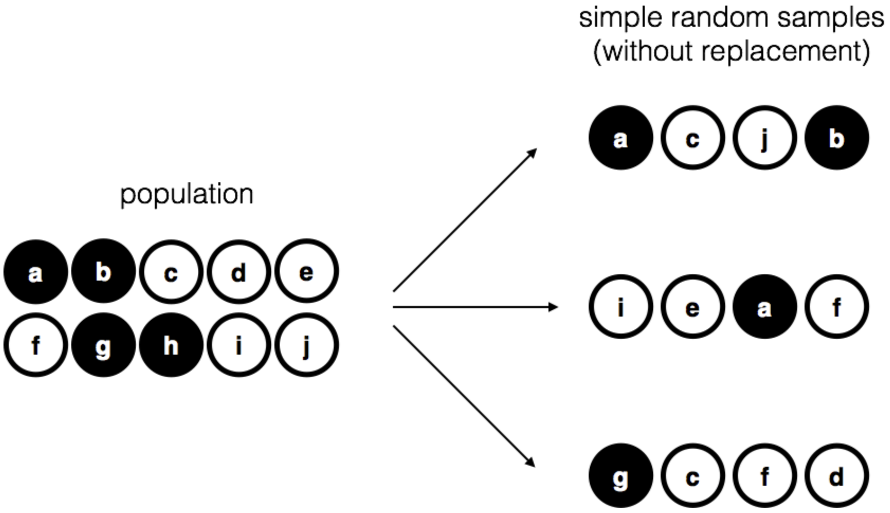


Figure 4.7: Simple random sampling without replacement from a finite population

This set of chips is the population of interest, and it is depicted graphically on the left of Figure **??**.

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 **??**. Each time you get different results, but the procedure is identical in each case. The fact that the same procedure can lead to different results each time, we refer to it as a **random** process. However, because we shook the bag before pulling any chips out, it seems reasonable to think that every chip has the same chance of being selected. A procedure in which every member of the population has the same chance of being selected is called a **simple random sample**. The fact that we did **not** put the chips back in the bag after pulling them out means that you can't observe the same thing twice, and in such cases the observations are said to have been sampled **without replacement**.

To help make sure you understand the importance of the sampling procedure, consider an alternative way in which the experiment could have been run. Suppose that my 5-year old son had opened the bag, and decided to pull out four black chips without putting any of them back in the bag. This **biased** sampling scheme is depicted in Figure **??**.

Now consider the evidentiary value of seeing 4 black chips and 0 white chips. Clearly, it depends on the sampling scheme, does it not? If you know that the sampling scheme is biased to select only black chips, then a sample that consists of only black chips doesn't tell you very

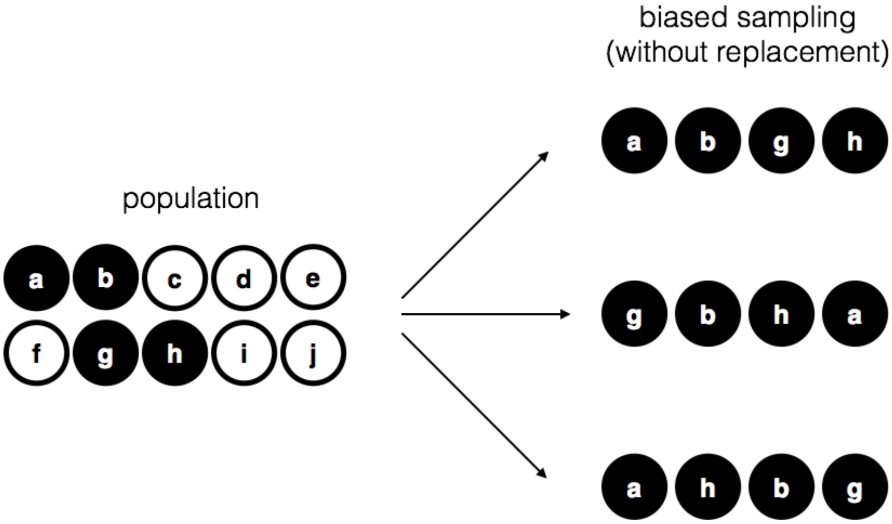


Figure 4.8: Biased sampling without replacement from a finite populations.

much about the population! For this reason, statisticians really like it when a data set can be considered a simple random sample, because it makes the data analysis **much** easier.

A third procedure is worth mentioning. This time around we close our eyes, shake the bag, and pull out a chip. This time, however, we record the observation and then put the chip back in the bag. Again we close our eyes, shake the bag, and pull out a chip. We then repeat this procedure until we have 4 chips. Data sets generated in this way are still simple random samples, but because we put the chips back in the bag immediately after drawing them it is referred to as a sample **with replacement**. The difference between this situation and the first one is that it is possible to observe the same population member multiple times, as illustrated in Figure **??**.

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.

4.8.3 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

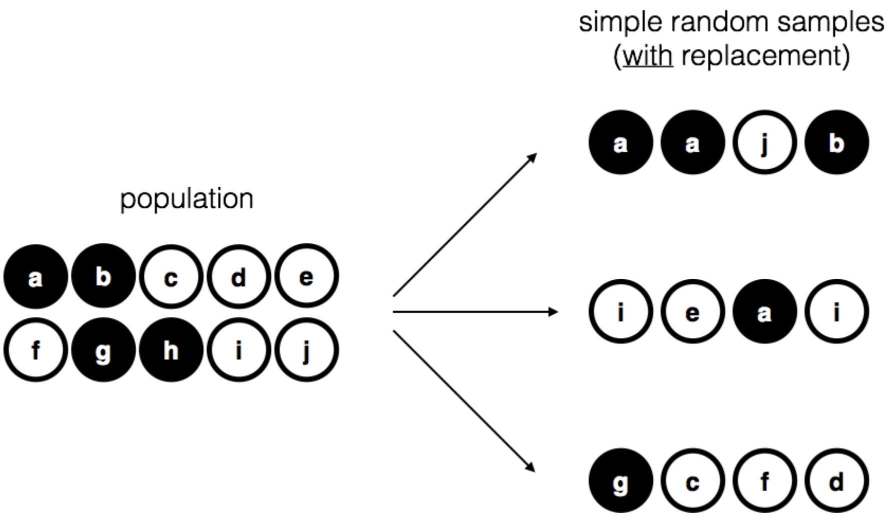


Figure 4.9: Simple random sampling with replacement from a finite population.

sampling of the undergraduate psychology students at Adelaide university, even though this is by far the narrowest population that I might want to generalize 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 that 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 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: if you weren't careful you might end up outing people who don't want to be outed (very, very bad form), and even if you don't make that mistake it can still be intrusive to use people's social networks to study them. It's certainly very hard to get people's informed consent **before** contacting them, yet in many cases the simple act of contacting them and saying "hey we want to study you" can be hurtful. Social networks are complex things, and just because you can use them to get data doesn't always mean you should.

- **Convenience sampling** is more or less what it sounds like. The samples are chosen in a way that is convenient to the researcher, and not selected at random from the population of interest. Snowball sampling is one type of convenience sampling, but there are many others. A common example in psychology are studies that rely on undergraduate psychology students. These samples are generally non-random in two respects: firstly, reliance on undergraduate psychology students automatically means that your data are restricted to a single sub-population. Secondly, the students usually get to pick which studies they participate in, so the sample is a self selected subset of psychology students not a randomly selected subset. In real life, most studies are convenience samples of one form or another. This is sometimes a severe limitation, but not always.

4.8.4 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 **??** and Figure **??**. However, it's not quite as bad as it sounds. Some types of biased samples are entirely unproblematic. For instance, when using a stratified sampling technique you actually **know** what the bias is because you created it deliberately, often to **increase** the effectiveness of your study, and there are statistical techniques that you can use to adjust for the biases you've introduced (not covered in this book!). So in those situations it's not a problem.

More generally though, it's important to remember that random sampling is a means to an end, not the end in itself. Let's assume you've relied on a convenience sample, and as such you can assume it's biased. A bias in your sampling method is only a problem if it causes you to draw the wrong conclusions. When viewed from that perspective, I'd argue that we don't need the sample to be randomly generated in **every** respect: we only need it to be random with respect to the psychologically-relevant phenomenon of interest. Suppose I'm doing a study looking at working memory capacity. In study 1, I actually have the ability to sample randomly from all human beings currently alive, with one exception: I can only sample people born on a Monday. In study 2, I am able to sample randomly from the Australian population. I want to generalize 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, industrialized country with a very well-developed education system. People growing up in that system will have had life experiences much more similar to the experiences of the people who designed the tests for working memory capacity. This shared experience might easily translate into similar beliefs about how to "take a test", a shared assumption about how psychological experimentation works, and so on. These things might actually matter. For instance, "test taking" style might have taught the Australian participants how to direct their attention exclusively on fairly abstract test materials relative to people that haven't grown up in a similar environment; leading to a misleading picture of what working memory capacity is.

There are two points hidden in this discussion. Firstly, when designing your own studies, it's important to think about what population you care about, and try hard to sample in a way that is appropriate to that population. In practice, you're usually forced to put up with a "sample of convenience" (e.g., psychology lecturers sample psychology students because that's the least expensive way to collect data, and our coffers aren't exactly overflowing with gold), but if so you should at least spend some time thinking about what the dangers of this practice might be.

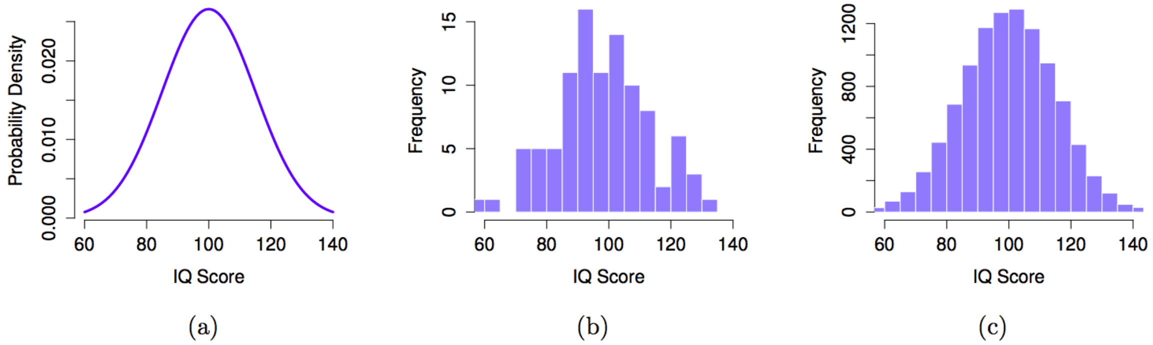
Secondly, if you're going to criticize someone else's study because they've used a sample of convenience rather than laboriously sampling randomly from the entire human population, at least have the courtesy to offer a specific theory as to **how** this might have distorted the results. Remember, everyone in science is aware of this issue, and does what they can to alleviate it. Merely pointing out that "the study only included people from group BLAH" is entirely unhelpful, and borders on being insulting to the researchers, who are aware of the issue. They just don't happen to be in possession of the infinite supply of time and money required to construct the perfect sample. In short, if you want to offer a responsible critique of the sampling process, then be **helpful**. Rehashing the blindingly obvious truisms that I've been rambling on about in this section isn't helpful.

4.8.5 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 operationalize our abstract theoretical ideas in terms of concrete measurements, statisticians operationalize the concept of a "population" in terms of mathematical objects that they know how to work with. You've already come across these objects they're called probability distributions (remember, the place where data comes from).

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 **??** *a*.



(a)

(b)

(c)

Figure 4.10: 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.

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.

Now suppose we collect some data. We select 100 people at random and administer an IQ test, giving a simple random sample from the population. The sample would consist of a collection of numbers like this:

```
106 101 98 80 74 ... 107 72 100
```

Each of these IQ scores is sampled from a normal distribution with mean 100 and standard deviation 15. So if I plot a histogram of the sample, I get something like the one shown in Figure **??** *b*. As you can see, the histogram is **roughly** the right shape, but it's a very crude approximation to the true population distribution shown in Figure **??** *a*. The mean of the sample is fairly close to the population mean 100 but not identical. In this case, it turns out that the people in the 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 the 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 we'll talk about how you can estimate population parameters using your sample statistics and how to work out how confident you are in your estimates but before we get to that there's a few more ideas in sampling theory that you need to know about.

4.9 The law of large numbers

We just looked at the results of one fictitious IQ experiment with a sample size of $N = 100$. The results were somewhat encouraging: the true population mean is 100, and the sample mean of 98.5 is a pretty reasonable approximation to it. In many scientific studies that level of precision is perfectly acceptable, but in other situations you need to be a lot more precise. If we want our sample statistics to be much closer to the population parameters, what can we do about it?

The obvious answer is to collect more data. Suppose that we ran a much larger experiment, this time measuring the IQ's of 10,000 people. We can simulate the results of this experiment using R, using the **rnorm()** function, which generates random numbers sampled from a normal distribution. For an experiment with a sample size of **n = 10000**, and a population with **mean = 100** and **sd = 15**, R produces our fake IQ data using these commands:

```
IQ <- rnorm(n=10000, mean=100, sd=15) #generate IQ scores
IQ <- round(IQ) # make round numbers
```

Cool, we just generated 10,000 fake IQ scores. Where did they go? Well, they went into the variable IQ on my computer. You can do the same on your computer too by copying the above code. 10,000 numbers is too many numbers to look at. We can look at the first 100 like this:

```
print(IQ[1:100])
#>   [1] 110  79 108 122  97  81 105 123  86 110  83 107 100  99  75  97 111 114
#>  [19] 115 110  98  98  88  96  68  87 101  98 111 133 102  91  95 115 107  67
#>  [37] 102 117  85  78 123  99  91 110 112  95  91 102 108 116  85  79  95 115
#>  [55] 101  96  75 112 104  92 106 122 114  87 100  99 120  88  88  92 111 114
#>  [73]  99 109  99 126 123 127 100 101  99  96  86  88 104 100 113  68 106  98
#>  [91] 132 110  94 108  77 116 106  85  92 112
```

We can compute the mean IQ using the command **mean(IQ)** and the standard deviation using the command **sd(IQ)**, and draw a histogram using **hist()**. The histogram of this much larger sample is shown in Figure @ref(fig:IQdist)c. Even a moment's inspections makes clear that the larger sample is a much better approximation to the true population distribution than the smaller one. This is reflected in the sample statistics: the mean IQ for the larger sample turns out to be 99.9, and the standard deviation is 15.1. These values are now very close to the true population.

I feel a bit silly saying this, but the thing I want you to take away from this is that large samples generally give you better information. I feel silly saying it because it's so bloody obvious that it shouldn't need to be said. In fact, it's such an obvious point that when Jacob Bernoulli – one of the founders of probability theory – formalized 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** (see Stigler, 1986, p65).

Okay, so the passage comes across as a bit condescending (not to mention sexist), but his main point is correct: it really does feel obvious that more data will give you better answers. The question is, why is this so? Not surprisingly, this intuition that we all share turns out to be correct, and statisticians refer to it as the **law of large numbers**. The law of large numbers is a mathematical law that applies to many different sample statistics, but the simplest way to think about it is as a law about averages. The sample mean is the most obvious example of a statistic that relies on averaging (because that's what the mean is... an average), so let's look at that. **When applied to the sample mean, what the law of large numbers states is that as the sample gets larger, the sample mean tends to get closer to the true population mean.** Or, to say it a little bit more precisely, as the sample size "approaches" infinity (written as $N \rightarrow \infty$) the sample mean approaches the population mean ($\bar{X} \rightarrow \mu$).

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.

4.10 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, 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 behavior of the sample mean when it is calculated from a more modest data set!

4.10.1 Sampling distribution of the sample means

"Oh no, what is the sample distribution of the sample means? Is that even allowed in English?". Yes, unfortunately, this is allowed. The **sampling distribution of the sample means** is the next most important thing you will need to understand. IT IS SO IMPORTANT THAT IT IS NECESSARY TO USE ALL CAPS. It is only confusing at first because it's long and uses sampling and sample in the same phrase.

Don't worry, we've been prepping you for this. You know what a distribution is right? It's where numbers comes from. It makes some numbers occur more or less frequently, or the same as other numbers. You know what a sample is right? It's the numbers we take from a distribution. So, what could the sampling distribution of the sample means refer to?

First, what do you think the sample means refers to? Well, if you took a sample of numbers, you would have a bunch of numbers…then, you could compute the mean of those numbers. The sample mean is the mean of the numbers in the sample. That is all. So, what is this distribution you speak of? Well, what if you took a bunch of samples, put one here, put one there, put some other ones other places. You have a lot of different samples of numbers. You could compute the mean for each them. Then you would have a bunch of means. What do those means look like? Well, if you put them in a histogram, you could find out. If you did that, you would be looking at (roughly) a distribution, AKA **the sampling distribution of the sample means**.

"I'm following along sort of, why would I want to do this instead of watching Netflix…". Because, the sampling distribution of the sample means gives you another window into chance. A very useful one that you can control, just like your remote control, by pressing the right design buttons.

4.10.2 Seeing the pieces

To make a sampling distribution of the sample means, we just need the following:

1. A distribution to take numbers from
2. A bunch of different samples from the distribution
3. The means of each of the samples
4. Get all of the sample means, and plot them in a histogram

Question for yourself: What do you think the sampling distribution of the sample means will look like? Will it tend to look the shape of the distribution that the samples came from? Or not? Good question, think about it.

Let's do those four things. We will sample numbers from the uniform distribution. Figure **??** shows the uniform distribution for sampling the set of integers from 1 to 10:

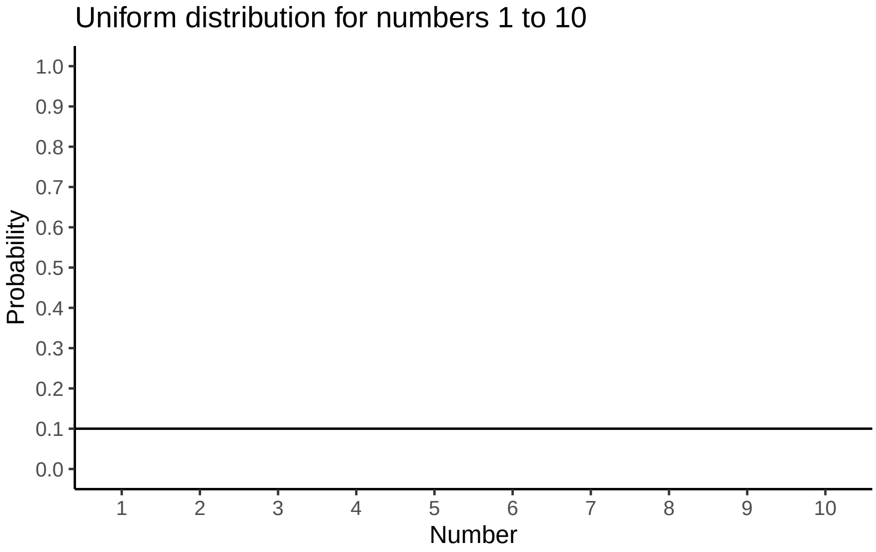


Figure 4.11: A uniform distribution illustrating the probabilites of sampling the numbers 1 to 10. In a uniform distribution, all numbers have an equal probability of being sampled, so the line is flat indicating all numbers have the same probability

?@fig-4sample20unif animates the process of taking a bunch of samples from the uniform distribution. We will set our sample-size to 20. It's easier to see how the sample mean behaves in a movie. Each histogram shows a new sample. The red line shows where the mean of the sample is. The samples are all very different from each other, but the red line doesn't move around very much, it always stays near the middle. However, the red line does move around a little bit, and this variance is what we call the sampling distribution of the sample mean.

OK, what have we got here? We have an animation of 10 different samples. Each sample has 20 observations and these are summarized in each of histograms that show up in the animation. Each histogram has a red line. The red line shows you where the mean of each sample is located. So, we have found the sample means for the 10 different samples from a uniform distribution.

First question. Are the sample means all the same? The answer is no. They are all kind of similar to each other though, they are all around five plus or minus a few numbers. This is interesting. Although all of our samples look pretty different from one another, the means of our samples look more similar than different.

Second question. What should we do with the means of our samples? Well, how about we collect them them all, and then plot a histogram of them. This would allow us to see what the distribution of the sample means looks like. The next histogram is just this. Except, rather than taking 10 samples, we will take 10,000 samples. For each of them we will compute the means. So, we will have 10,000 means. Figure **??** shows the histogram of the sample means:

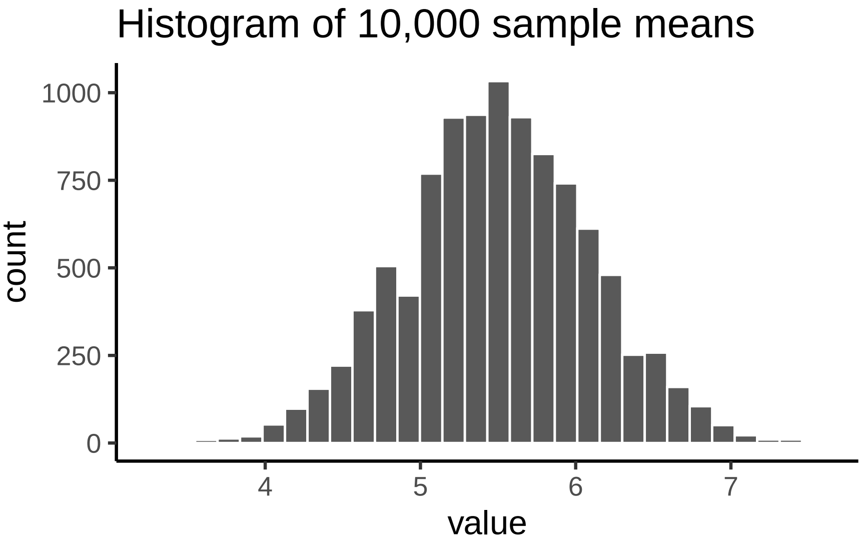


Figure 4.12: A histogram showing the sample means for 10,000 samples, each size 20, from the uniform distribution of numbers from 1 to 10. The expected mean is 5.5, and the histogram is centered on 5.5. The mean of each sample is not always 5.5 because of sampling error or chance

“Wait what? This doesn’t look right. I thought we were taking samples from a uniform distribution. Uniform distributions are flat. THIS DOES NOT LOOK LIKE A FLAT DISTRIBTUION, WHAT IS GOING ON, AAAAAGGGHH.”. We feel your pain.

Remember, we are looking at the distribution of sample means. It is indeed true that the distribution of sample means does not look the same as the distribution we took the samples from. Our distribution of sample means goes up and down. In fact, this will almost always be the case for distributions of sample means. This fact is called the **central limit theorem**, which we talk about later.

For now, let's talk about about what's happening. Remember, we have been sampling numbers between the range 1 to 10. We are supposed to get each number with roughly equal frequency, because we are sampling from a uniform distribution. So, let's say we took a sample of 10 numbers, and happened to get one of each from 1 to 10.

```
1 2 3 4 5 6 7 8 9 10
```

What is the mean of those numbers? Well, its 1+2+3+4+5+6+7+8+9+10 = 55 / 10 = 5.5. Imagine if we took a bigger sample, say of 20 numbers, and again we got exactly 2 of each number. What would the mean be? It would be (1+2+3+4+5+6+7+8+9+10)*2 = 110 / 20 = 5.5. Still 5.5. You can see here, that the mean value of our uniform distribution is 5.5. Now that we know this, we might expect that most of our samples will have a mean near this number. We already know that every sample won't be perfect, and it won't have exactly an equal amount of every number. So, we will expect the mean of our samples to vary a little bit. The histogram that we made shows the variation. Not surprisingly, the numbers vary around the value 5.5.

4.10.3 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 you sampled some numbers from an experiment you wrote down the largest number in the experiment. Doing this over and over again would give you a very different sampling distribution, namely the **sampling distribution of the maximum**. You could calculate the smallest number, or the mode, or the median, of the variance, or the standard deviation, or anything else from your sample. Then, you could repeat many times, and produce the sampling distribution of those statistics. Neat!

Just for fun here are some different sampling distributions for different statistics. We will take a normal distribution with mean = 100, and standard deviation =20. Then, we'll take lots of samples with n = 50 (50 observations per sample). We'll save all of the sample statistics, then plot their histograms in Figure **??**. Let's do it:

We just computed 4 different sampling distributions, for the mean, standard deviation, maximum value, and the median. If you just look quickly at these histograms you might think they all basically look the same. Hold up now. It's very important to look at the x-axes. They are different. For example, the sample mean goes from about 90 to 110, whereas the standard deviation goes from 15 to 25.

These sampling distributions are super important, and worth thinking about. What should you think about? Well, here's a clue. These distributions are telling you what to expect from

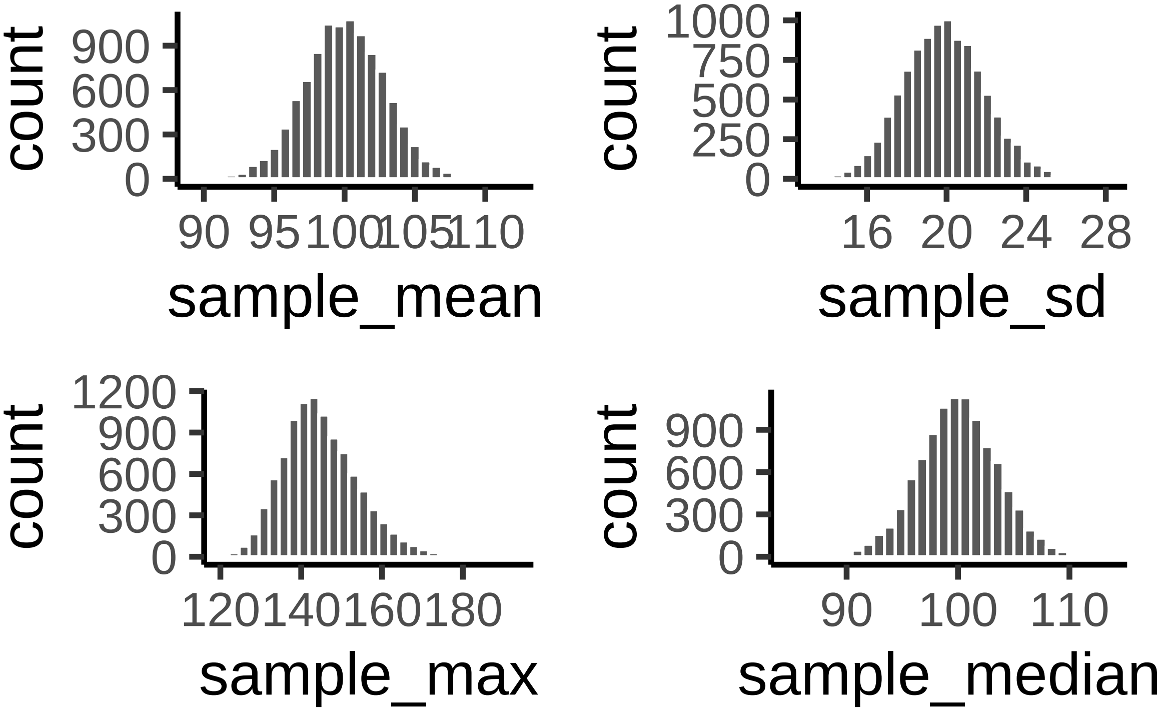


Figure 4.13: Each panel shows a histogram of a different sampling statistic

your sample. Critically, they are telling you what you should expect from a sample, when you take one from the specific distribution that we used (normal distribution with mean =100 and SD = 20). What have we learned. We've learned a tonne. We've learned that we can expect our sample to have a mean somewhere between 90 and 108ish. Notice, the sample means are never more extreme. We've learned that our sample will usually have some variance, and that the the standard deviation will be somewhere between 15 and 25 (never much more extreme than that). We can see that sometime we get some big numbers, say between 120 and 180, but not much bigger than that. And, we can see that the median is pretty similar to the mean. If you ever took a sample of 50 numbers, and your descriptive statistics were inside these windows, then perhaps they came from this kind of normal distribution. If your sample statistics are very different, then your sample probably did not come this distribution. By using simulation, we can find out what samples look like when they come from distributions, and we can use this information to make inferences about whether our sample came from particular distributions.

4.11 The central limit theorem

OK, so now you've seen lots of sampling distributions, and you know what the sampling distribution of the mean is. Here, we'll focus on **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 (you've already seen it bounce around): 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.

Let's give ourselves a nice movie to see everything in action. We're going to sample numbers from a normal distribution. **?@fig-4samplingmean** has four panels, each panel represents a different sample size (n), including sample-sizes of 10, 50, 100, and 1000. The red line shows the shape of the normal distribution. The grey bars show a histogram of each of the samples that we take. The red line shows the mean of an individual sample (the middle of the grey bars). As you can see, the red line moves around a lot, especially when the sample size is small (10).

The new bits are the blue bars and the blue lines. The blue bars represent the sampling distribution of the sample mean. For example, in the panel for sample-size 10, we see a bunch of blue bars. This is a histogram of 10 sample means, taken from 10 samples of size 10. In the 50 panel, we see a histogram of 50 sample means, taken from 50 samples of size 50, and so on. The blue line in each panel is the mean of the sample means ("aaagh, it's a mean of means", yes it is).

What should you notice? Notice that the range of the blue bars shrinks as sample size increases. The sampling distribution of the mean is quite wide when the sample-size is 10, it narrows as sample-size increases to 50 and 100, and it's just one bar, right in the middle when sample-size goes to 1000. What we are seeing is that the mean of the sampling distribution approaches the mean of the population as sample-size increases.

So, the sampling distribution of the mean is another distribution, and it has some variance. It varies more when sample-size is small, and varies less when sample-size is large. 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 movie, as the sample size N increases, the SEM decreases.

Okay, so that's one part of the story. However, there's something we've been glossing over a little bit. We've seen it already, but it's worth looking at it one more time. Here's the thing: **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. This is the central limit theorem.

To see the central limit theorem in action, we are going to look at some histograms of sample means from different kinds of distributions. It is very important to recognize that you are looking at distributions of sample means, not distributions of individual samples.

Here we go, Figure **??** shows sampling from a normal distribution. The red line is the normal distribution where each sample is drawn from. The mean for each sample of numbers is computed, and the distribution of sample means is shown by the blue bars. Note that the shape of red line and the blue bars are similar, they both look like a normal distribution.

Let's do it again. This time we will sample from a flat uniform distribution shown by the red line. However, Figure **??** shows the distribution of sample means represented by the blue bars is not flat, it looks like a normal distribution.

One more time with an exponential distribution (shown in red) where smaller numbers are more likely to be sampled than larger numbers. Even though way more of the numbers in a given sample will be smaller than larger, according to Figure **??** the sampling distribution of the mean does not look the red line. Instead, the sampling distribution of the mean looks like a bell-shaped normal curve. This is the central limit theorem in action.

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

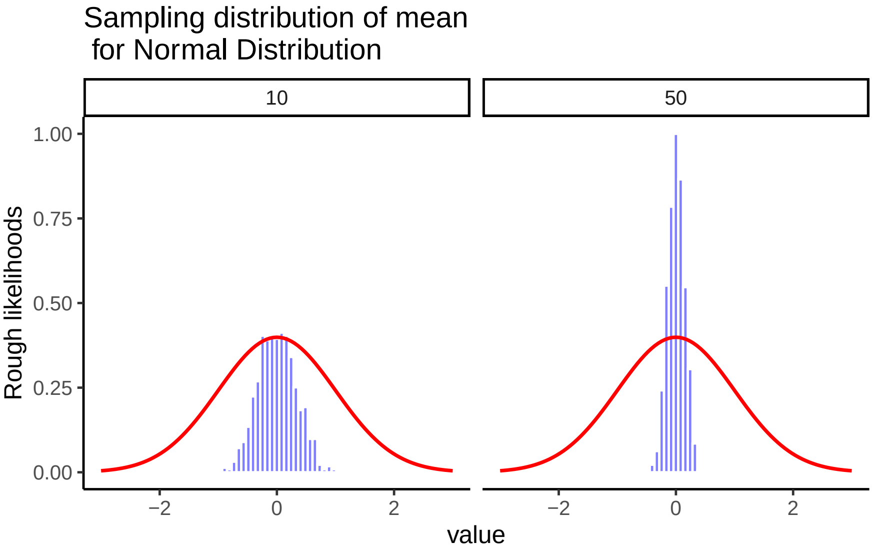


Figure 4.14: Comparison of two normal distributions, and histograms for the sampling distribution of the mean for different samples-sizes. The range of sampling distribution of the mean shrinks as sample-size increases

- The shape of the sampling distribution becomes normal as the sample size increases

As it happens, not only are all of these statements true, there is a very famous theorem in statistics that proves all three of them, known as the **central limit theorem**. Among other things, the central limit theorem tells us that if the population distribution has mean μ and standard deviation σ, then the sampling distribution of the mean also has mean μ, and the standard error of the mean is

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

Because we divide the population standard deviation σ by the square root of the sample size N, the SEM gets smaller as the sample size increases. It also tells us that the shape of the sampling distribution becomes normal.

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.

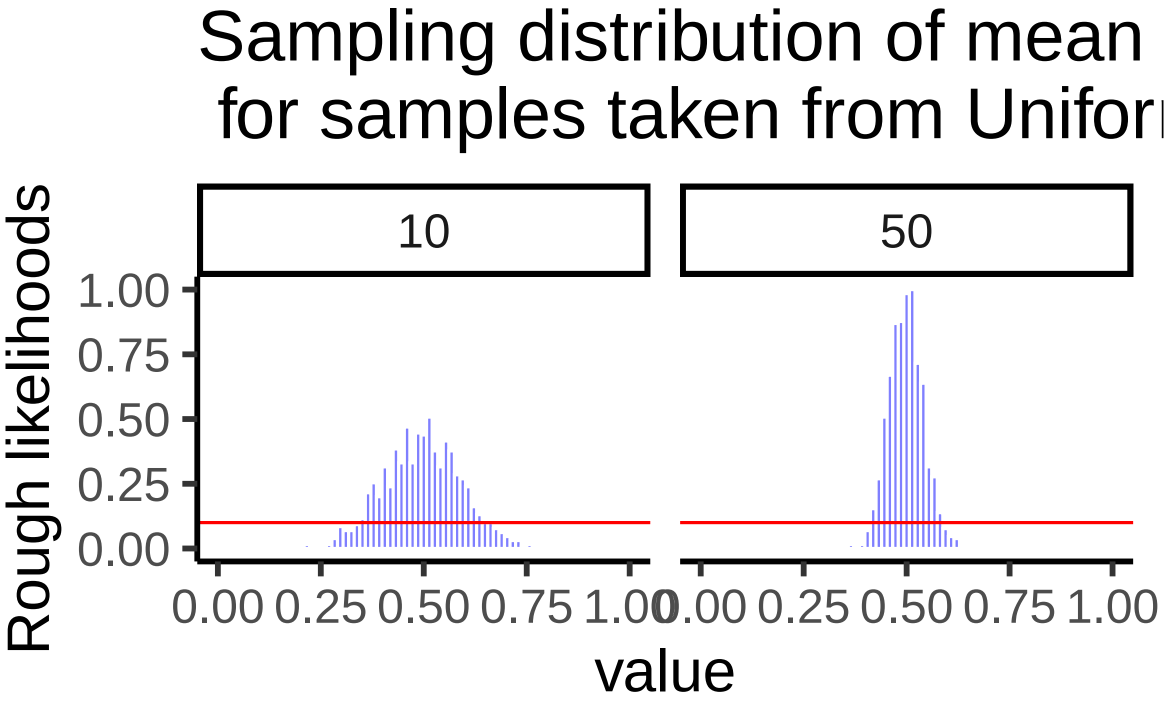


Figure 4.15: Illustration that the shape of the sampling distribution of the mean is normal, even when the samples come from a non-normal (uniform in this case) distribution

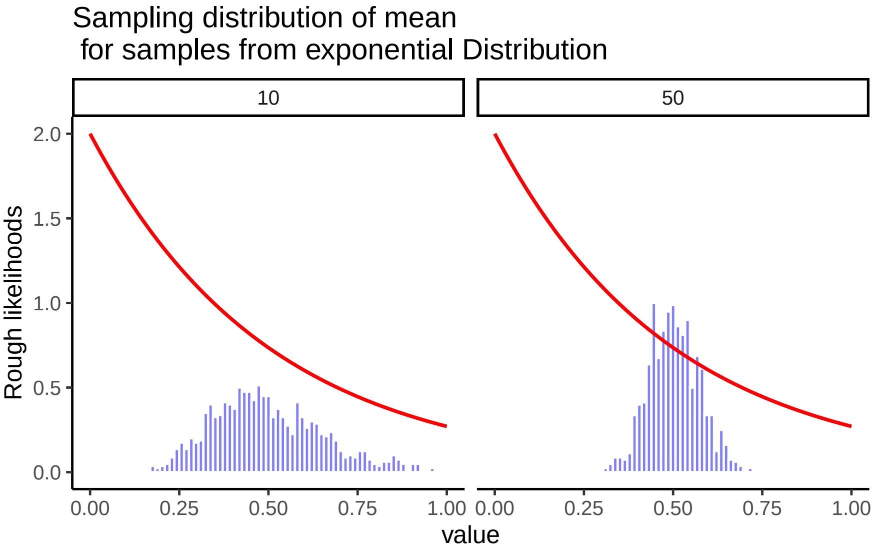


Figure 4.16: Illustration that the shape of the sampling distribution of the mean is normal, even when the samples come from an exponential distribution

4.12 z-scores

We are now in a position to combine some of things we've been talking about in this chapter, and introduce you to a new tool, **z-scores**. It turns out we won't use **z-scores** very much in this textbook. However, you can't take a class on statistics and not learn about **z-scores**.

We are going to look at a normal distribution in Figure **??**, and draw lines through the distribution at 0, +/- 1, +/-2, and +/- 3 standard deviations from the mean:

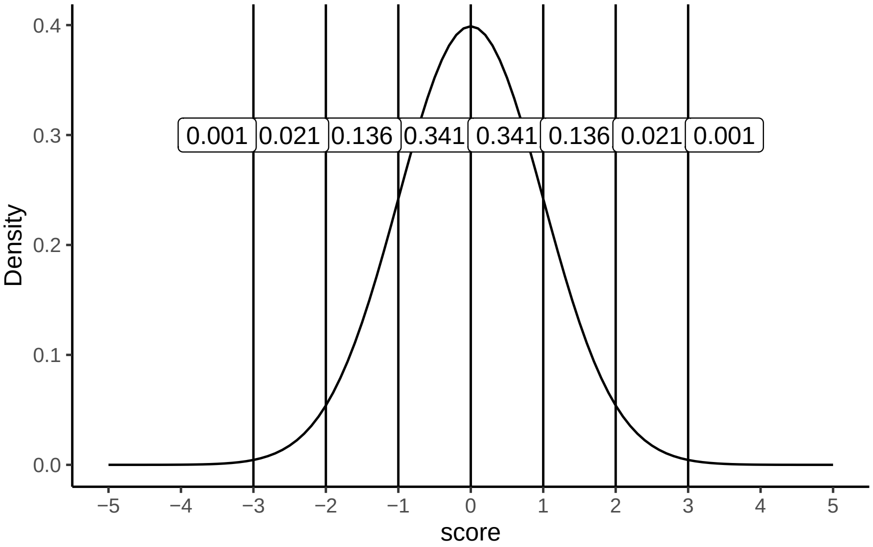


Figure 4.17: A normal distribution. Each line represents a standard deviation from the mean. The labels show the proportions of scores that fall between each bar.

The figure shows a normal distribution with mean = 0, and standard deviation = 1. We've drawn lines at each of the standard deviations: -3, -2, -1, 0, 1, 2, and 3. We also show some numbers in the labels, in between each line. These numbers are proportions. For example, we see the proportion is .341 for scores that fall between the range 0 and 1. Scores between 0 and 1 occur 34.1% of the time. Scores in between -1 and 1, occur 68.2% of the time, that's more than half of the scores. Scores between 1 and occur about 13.6% of the time, and scores between 2 and 3 occur even less, only 2.1% of the time.

Normal distributions always have these properties, even when they have different means and standard deviations. For example, take a look at the normal distribution in Figure **??** that has a mean = 100, and standard deviation = 25.

Now we are looking at a normal distribution with mean = 100 and standard deviation = 25. Notice that the region between 100 and 125 contains 34.1% of the scores. This region is 1 standard deviation away from the mean (the standard deviation is 25, the mean is 100, so 25 is one whole standard deviation away from 100). As you can see, the very same proportions

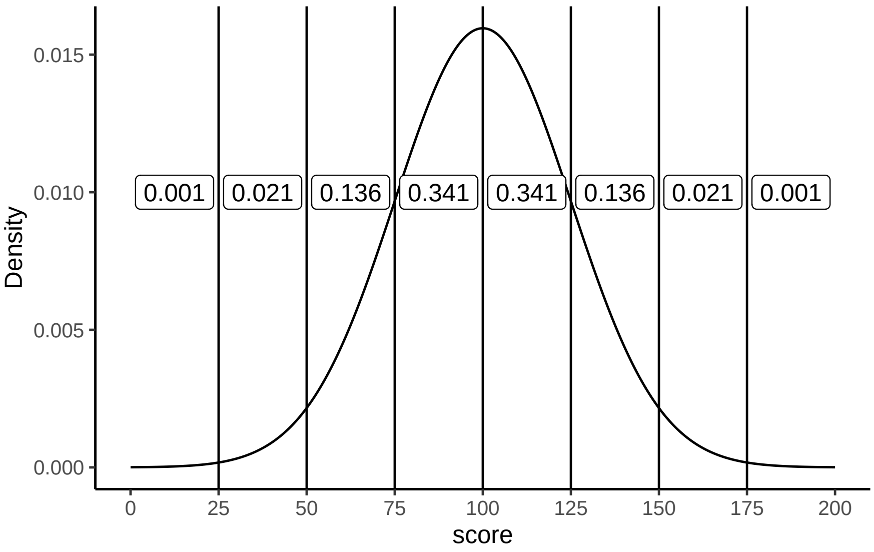


Figure 4.18: A normal distribution. Each line represents a standard deviation from the mean. The labels show the proportions of scores that fall between each bar.

occur between each of the standard deviations, as they did when our standard deviation was set to 1 (with a mean of 0).

4.12.1 Idea behind z-scores

Sometimes it can be convenient to transform your original scores into different scores that are easier to work with. For example, if you have a bunch of proportions, like .3, .5, .6, .7, you might want to turn them into percentages like 30%, 50%, 60%, and 70%. To do that you multiply the proportions by a constant of 100. If you want to turn percentages back into proportions, you divide by a constant of 100. This kind of transformation just changes the scale of the numbers from between 0-1, and between 0-100. Otherwise, the pattern in the numbers stays the same.

The idea behind z-scores is a similar kind of transformation. The idea is to express each raw score in terms of it's standard deviation. For example, if I told you I got a 75% on test, you wouldn't know how well I did compared to the rest of the class. But, if I told you that I scored 2 standard deviations above the mean, you'd know I did quite well compared to the rest of the class, because you know that most scores (if they are distributed normally) fall below 2 standard deviations of the mean.

We also know, now thanks to the central limit theorem, that many of our measures, such as sample means, will be distributed normally. So, it can often be desirable to express the raw scores in terms of their standard deviations.

Let's see how this looks in a table without showing you any formulas. We will look at some scores that come from a normal distribution with mean = 100, and standard deviation = 25. We will list some raw scores, along with the z-scores

raw	z
25	-3
50	-2
75	-1
100	0
125	1
150	2
175	3

Remember, the mean is 100, and the standard deviation is 25. How many standard deviations away from the mean is a score of 100? The answer is 0, it's right on the mean. You can see the z-score for 100, is 0. How many standard deviations is 125 away from the mean? Well the standard deviation is 25, 125 is one whole 25 away from 100, that's a total of 1 standard deviation, so the z-score for 125 is 1. The z-score for 150 is 2, because 150 is two 25s away from 100. The z-score for 50 is -2, because 50 is two 25s away from 100 in the opposite direction. All we are doing here is re-expressing the raw scores in terms of how many standard deviations they are from the mean. Remember, the mean is always right on target, so the center of the z-score distribution is always 0.

4.12.2 Calculating z-scores

To calculate z-scores all you have to do is figure out how many standard deviations from the mean each number is. Let's say the mean is 100, and the standard deviation is 25. You have a score of 97. How many standard deviations from the mean is 97?

First compute the difference between the score and the mean:

$97 - 100 = -3$

Alright, we have a total difference of -3. How many standard deviations does -3 represent if 1 standard deviation is 25? Clearly -3 is much smaller than 25, so it's going to be much less than 1. To figure it out, just divide -3 by the standard deviation.

$\frac{-3}{25} = -.12$

Our z-score for 97 is -.12.

Here's the general formula:

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

So, for example if we had these 10 scores from a normal distribution with mean = 100, and standard deviation =25

```
#>  [1] 111.83 114.87  75.28  61.08  98.05 103.29  70.79 125.17  86.41 152.67
```

The z-scores would be:

```
#>  [1]  0.4732  0.5948 -0.9888 -1.5568 -0.0780  0.1316 -1.1684  1.0068 -0.5436
#> [10]  2.1068
```

Once you have the z-scores, you could use them as another way to describe your data. For example, now just by looking at a score you know if it is likely or unlikely to occur, because you know how the area under the normal curve works. z-scores between -1 and 1 happen pretty often, scores greater than 1 or -1 still happen fairly often, but not as often. And, scores bigger than 2 or -2 don't happen very often. This is a convenient thing to do if you want to look at your numbers and get a general sense of how often they happen.

Usually you do not know the mean or the standard deviation of the population that you are drawing your sample scores from. So, you could use the mean and standard deviation of your sample as an estimate, and then use those to calculate z-scores.

Finally, z-scores are also called **standardized scores**, because each raw score is described in terms of it's standard deviation. This may well be the last time we talk about z-scores in this book. You might wonder why we even bothered telling you about them. First, it's worth knowing they are a thing. Second, they become important as your statistical prowess becomes more advanced. Third, some statistical concepts, like correlation, can be re-written in terms of z-scores, and this illuminates aspects of those statistics. Finally, they are super useful when you are dealing with a normal distribution that has a known mean and standard deviation.

4.13 Estimating population parameters

Let's pause for a moment to get our bearings. We're about to go into the topic of **estimation**. What is that, and why should you care? First, population parameters are things about a distribution. For example, distributions have means. The mean is a parameter of the distribution. The standard deviation of a distribution is a parameter. Anything that can describe a distribution is a potential parameter.

OK fine, who cares? This I think, is a really good question. There are some good concrete reasons to care. And there are some great abstract reasons to care. Unfortunately, most of the time in research, it's the abstract reasons that matter most, and these can be the most difficult to get your head around.

4.13.1 Concrete population parameters

First some concrete reasons. There are real populations out there, and sometimes you want to know the parameters of them. For example, if you are a shoe company, you would want to know about the population parameters of feet size. As a first pass, you would want to know the mean and standard deviation of the population. If your company knew this, and other companies did not, your company would do better (assuming all shoes are made equal). Why would your company do better, and how could it use the parameters? Here's one good reason. As a shoe company you want to meet demand with the right amount of supply. If you make too many big or small shoes, and there aren't enough people to buy them, then you're making extra shoes that don't sell. If you don't make enough of the most popular sizes, you'll be leaving money on the table. Right? Yes. So, what would be an optimal thing to do? Perhaps, you would make different amounts of shoes in each size, corresponding to how the demand for each shoe size. You would know something about the demand by figuring out the frequency of each size in the population. You would need to know the population parameters to do this.

Fortunately, it's pretty easy to get the population parameters without measuring the entire population. Who has time to measure every-bodies feet? Nobody, that's who. Instead, you would just need to randomly pick a bunch of people, measure their feet, and then measure the parameters of the sample. If you take a big enough sample, we have learned that the sample mean gives a very good estimate of the population mean. We will learn shortly that a version of the standard deviation of the sample also gives a good estimate of the standard deviation of the population. Perhaps shoe-sizes have a slightly different shape than a normal distribution. Here too, if you collect a big enough sample, the shape of the distribution of the sample will be a good estimate of the shape of the populations. All of these are good reasons to care about estimating population parameters. But, do you run a shoe company? Probably not.

4.13.2 Abstract population parameters

Even when we think we are talking about something concrete in Psychology, it often gets abstract right away. Instead of measuring the population of feet-sizes, how about the population of human happiness. We all think we know what happiness is, everyone has more or less of it, there are a bunch of people, so there must be a population of happiness right? Perhaps, but it's not very concrete. The first problem is figuring out how to measure happiness. Let's use a questionnaire. Consider these questions:

> How happy are you right now on a scale from 1 to 7? How happy are you in general on a scale from 1 to 7? How happy are you in the mornings on a scale from 1 to 7? How happy are you in the afternoons on a scale from 1 to 7?

1. = very unhappy
2. = unhappy
3. = sort of unhappy

4. = in the middle
5. = sort of happy
6. = happy
7. = very happy

Forget about asking these questions to everybody in the world. Let's just ask them to lots of people (our sample). What do you think would happen? Well, obviously people would give all sorts of answers right. We could tally up the answers and plot them in a histogram. This would show us a distribution of happiness scores from our sample. "Great, fantastic!", you say. Yes, fine and dandy.

So, on the one hand we could say lots of things about the people in our sample. We could say exactly who says they are happy and who says they aren't, after all they just told us!

But, what can we say about the larger population? Can we use the parameters of our sample (e.g., mean, standard deviation, shape etc.) to estimate something about a larger population. Can we infer how happy everybody else is, just from our sample? HOLD THE PHONE.

4.13.2.1 Complications with inference

Before listing a bunch of complications, let me tell you what I think we can do with our sample. Provided it is big enough, our sample parameters will be a pretty good estimate of what another sample would look like. Because of the following discussion, this is often all we can say. But, that's OK, as you see throughout this book, we can work with that!

Problem 1: Multiple populations: If you looked at a large sample of questionnaire data you will find evidence of multiple distributions inside your sample. People answer questions differently. Some people are very cautious and not very extreme. Their answers will tend to be distributed about the middle of the scale, mostly 3s, 4s, and 5s. Some people are very bi-modal, they are very happy and very unhappy, depending on time of day. These people's answers will be mostly 1s and 2s, and 6s and 7s, and those numbers look like they come from a completely different distribution. Some people are entirely happy or entirely unhappy. Again, these two "populations" of people's numbers look like two different distributions, one with mostly 6s and 7s, and one with mostly 1s and 2s. Other people will be more random, and their scores will look like a uniform distribution. So, is there a single population with parameters that we can estimate from our sample? Probably not. Could be a mixture of lots of populations with different distributions.

Problem 2: What do these questions measure?: If the whole point of doing the questionnaire is to estimate the population's happiness, we really need wonder if the sample measurements actually tell us anything about happiness in the first place. Some questions: Are people accurate in saying how happy they are? Does the measure of happiness depend on the scale, for example, would the results be different if we used 0-100, or -100 to +100, or no numbers? Does the measure of happiness depend on the wording in the question? Does a measure like this one tell us everything we want to know about happiness (probably not), what is it missing (who knows? probably lots). In short, nobody knows if these kinds of questions measure what we want them to measure. We just hope that they do. Instead, we have a very good idea of the kinds of things that they actually measure. It's really quite obvious, and staring you in the face. Questionnaire measurements measure how people answer questionnaires. In other words, how people behave and answer questions when they are given a questionnaire. This might also measure something about happiness, when the question has to do about happiness. But, it turns out people are remarkably consistent in how they answer questions, even when the questions are total nonsense, or have no questions at all (just numbers to choose!) Maul (2017).

The take home complications here are that we can collect samples, but in Psychology, we often don't have a good idea of the populations that might be linked to these samples. There might be lots of populations, or the populations could be different depending on who you ask. Finally, the "population" might not be the one you want it to be.

4.13.3 Experiments and Population parameters

OK, so we don't own a shoe company, and we can't really identify the population of interest in Psychology, can't we just skip this section on estimation? After all, the "population" is just too weird and abstract and useless and contentious. HOLD THE PHONE AGAIN!

It turns out we can apply the things we have been learning to solve lots of important problems in research. These allow us to answer questions with the data that we collect. Parameter estimation is one of these tools. We just need to be a little bit more creative, and a little bit more abstract to use the tools.

Here is what we know already. The numbers that we measure come from somewhere, we have called this place "distributions". Distributions control how the numbers arrive. Some numbers happen more than others depending on the distribution. We assume, even if we don't know what the distribution is, or what it means, that the numbers came from one. Second, when get some numbers, we call it a sample. This entire chapter so far has taught you one thing. When your sample is big, it resembles the distribution it came from. And, when your sample is big, it will resemble very closely what another big sample of the same thing will look like. We can use this knowledge!

Very often as Psychologists what we want to know is what causes what. We want to know if X causes something to change in Y. Does eating chocolate make you happier? Does studying improve your grades? There a bazillions of these kinds of questions. And, we want answers to them.

I've been trying to be mostly concrete so far in this textbook, that's why we talk about silly things like chocolate and happiness, at least they are concrete. Let's give a go at being abstract. We can do it.

So, we want to know if X causes Y to change. What is X? What is Y? X is something you change, something you manipulate, the independent variable. Y is something you measure. So, we will be taking samples from Y. "Oh I get it, we'll take samples from Y, then we can use the sample parameters to estimate the population parameters of Y!" NO, not really, but yes sort of. We will take sample from Y, that is something we absolutely do. In fact, that is really all we ever do, which is why talking about the population of Y is kind of meaningless. We're more interested in our samples of Y, and how they behave.

So, what would happen if we removed X from the universe altogether, and then took a big sample of Y. We'll pretend Y measures something in a Psychology experiment. So, we know right away that Y is variable. When we take a big sample, it will have a distribution (because Y is variable). So, we can do things like measure the mean of Y, and measure the standard deviation of Y, and anything else we want to know about Y. Fine. What would happen if we replicated this measurement. That is, we just take another random sample of Y, just as big as the first. What should happen is that our first sample should look a lot like our second example. After all, we didn't do anything to Y, we just took two big samples twice. Both of our samples will be a little bit different (due to sampling error), but they'll be mostly the same. The bigger our samples, the more they will look the same, especially when we don't do anything to cause them to be different. In other words, we can use the parameters of one sample to estimate the parameters of a second sample, because they will tend to be the same, especially when they are large.

We are now ready for step two. You want to know if X changes Y. What do you do? You make X go up and take a big sample of Y then look at it. You make X go down, then take a second big sample of Y and look at it. Next, you compare the two samples of Y. If X does nothing then what should you find? We already discussed that in the previous paragraph. If X does nothing, then both of your big samples of Y should be pretty similar. However, if X does something to Y, then one of your big samples of Y will be different from the other. You will have changed something about Y. Maybe X makes the mean of Y change. Or maybe X makes the variation in Y change. Or, maybe X makes the whole shape of the distribution change. If we find any big changes that can't be explained by sampling error, then we can conclude that something about X caused a change in Y! We could use this approach to learn about what causes what!

The very important idea is still about estimation, just not population parameter estimation exactly. We know that when we take samples they naturally vary. So, when we estimate a parameter of a sample, like the mean, we know we are off by some amount. When we find that two samples are different, we need to find out if the size of the difference is consistent with what sampling error can produce, or if the difference is bigger than that. If the difference is bigger, then we can be confident that sampling error didn't produce the difference. So, we can confidently infer that something else (like an X) did cause the difference. This bit of abstract thinking is what most of the rest of the textbook is about. Determining whether there is a difference caused by your manipulation. There's more to the story, there always is. We can get more specific than just, is there a difference, but for introductory purposes, we will focus on the finding of differences as a foundational concept.

4.13.4 Interim summary

We've talked about estimation without doing any estimation, so in the next section we will do some estimating of the mean and of the standard deviation. Formally, we talk about this as using a sample to estimate a parameter of the population. Feel free to think of the "population" in different ways. It could be concrete population, like the distribution of feet-sizes. Or, it could be something more abstract, like the parameter estimate of what samples usually look like when they come from a distribution.

4.13.5 Estimating the population mean

Suppose we go to Brooklyn 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 Brooklyn? Obviously, we don't know the answer to that question. It could be 97.2, but if could also be 103.5. Our sampling isn't exhaustive so we cannot give a definitive answer. Nevertheless if forced to give a "best guess" I'd have to say 98.5. That's the essence of statistical estimation: giving a best guess. We're using the sample mean as the best guess of the population mean.

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 we'll explain the statistical justification for this intuitive answer. However, for the moment let's make sure you recognize 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 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:

Symbol	What is it?	Do we know what it is?
$\bar{X}$	Sample mean	Yes, calculated from the raw data
μ	True population mean	Almost never known for sure
$\hat{\mu}$	Estimate of the population mean	Yes, identical to the sample mean

4.13.6 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 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**.

If the error is systematic, that means it is **biased**. For example, imagine if the sample mean was always smaller than the population mean. If this was true (it's not), then we couldn't use the sample mean as an estimator. It would be biased, we'd be using the wrong number.

It turns out the sample standard deviation is a **biased estimator** of the population standard deviation. We can sort of anticipate this by what we've been discussing. When the sample size is 1, the standard deviation is 0, which is obviously to small. When the sample size is 2, the standard deviation becomes a number bigger than 0, but because we only have two sample, we suspect it might still be too small. Turns out this intuition is correct.

It would be nice to demonstrate this somehow. There are in fact mathematical proofs that confirm this intuition, but unless you have the right mathematical background they don't help very much. Instead, what I'll do is use R to simulate the results of some experiments. With that in mind, let's return to our IQ studies. Suppose the true population mean IQ is 100 and the standard deviation is 15. I can use the **rnorm()** function to generate the the results of an experiment in which I measure $N = 2$ IQ scores, and calculate the sample standard deviation. If I do this over and over again, and plot a histogram of these sample standard deviations, what I have is the **sampling distribution of the standard deviation**. I've plotted this distribution in Figure **??**.

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 from when we were plotting sampling distributions of the sample mean, those were always centered around the mean of the population.

Now let's extend the simulation. Instead of restricting ourselves to the situation where we have a sample size of $N = 2$, let's repeat the exercise for sample sizes from 1 to 10. If we plot the average sample mean and average sample standard deviation as a function of sample size, you get the following results.

@fig-estimatorbiasA shows the sample mean as a function of sample size. Notice it's a flat line. The sample mean doesn't underestimate or overestimate the population mean. It is an unbiased estimate!

Figure **??** shows the sample standard deviation as a function of sample size. Notice it is not a flat line. The sample standard deviation systematically underestimates the population standard deviation!

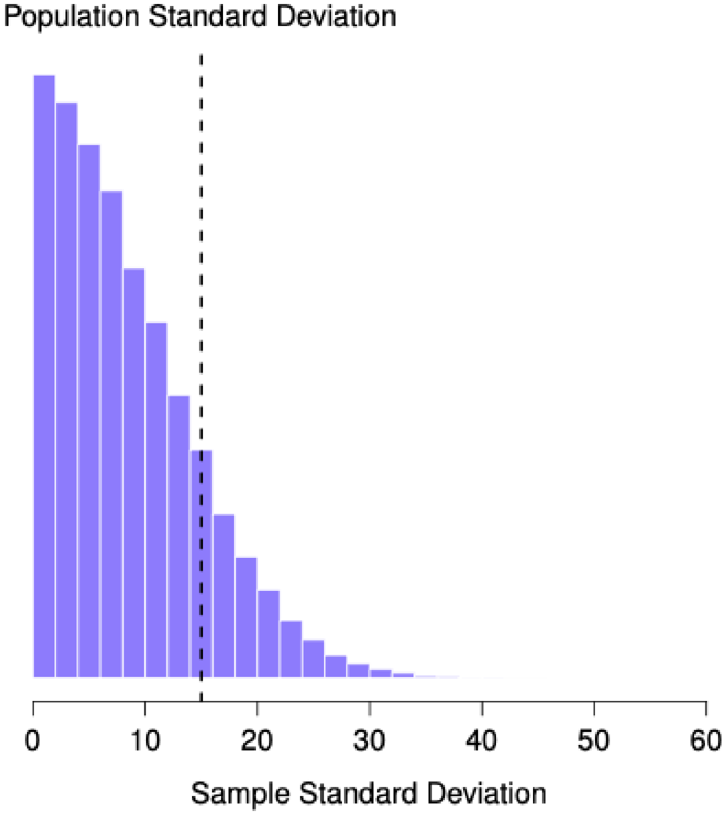


Figure 4.19: The sampling distribution of the sample standard deviation for a two IQ scores experiment. The true population standard deviation is 15 (dashed line), but as you can see from the histogram, the vast majority of experiments will produce a much smaller sample standard deviation than this. On average, this experiment would produce a sample standard deviation of only 8.5, well below the true value! In other words, the sample standard deviation is a biased estimate of the population standard deviation.

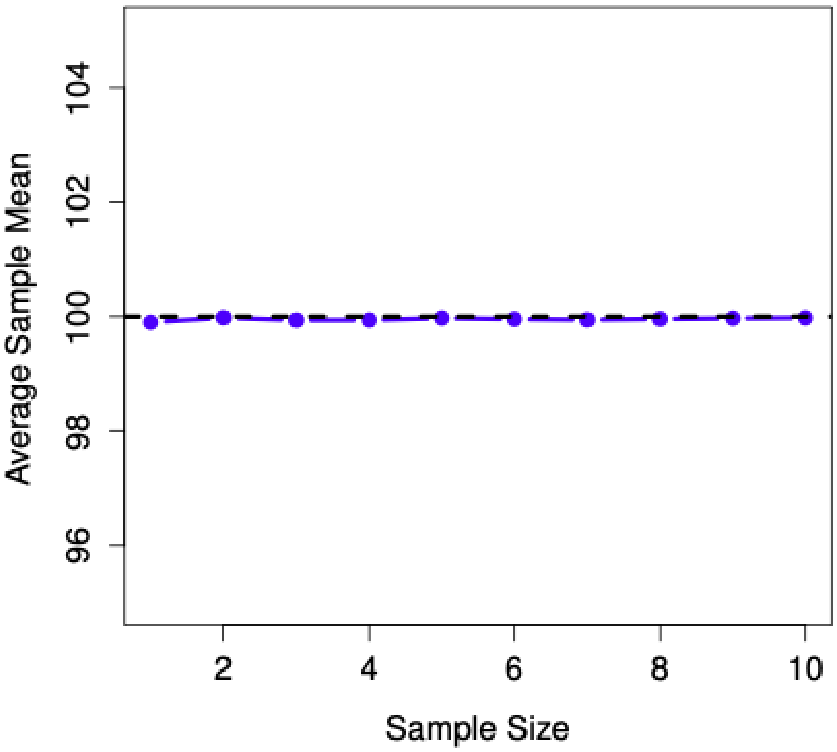


Figure 4.20: An illustration of the fact that the sample mean is an unbiased estimator of the population mean.

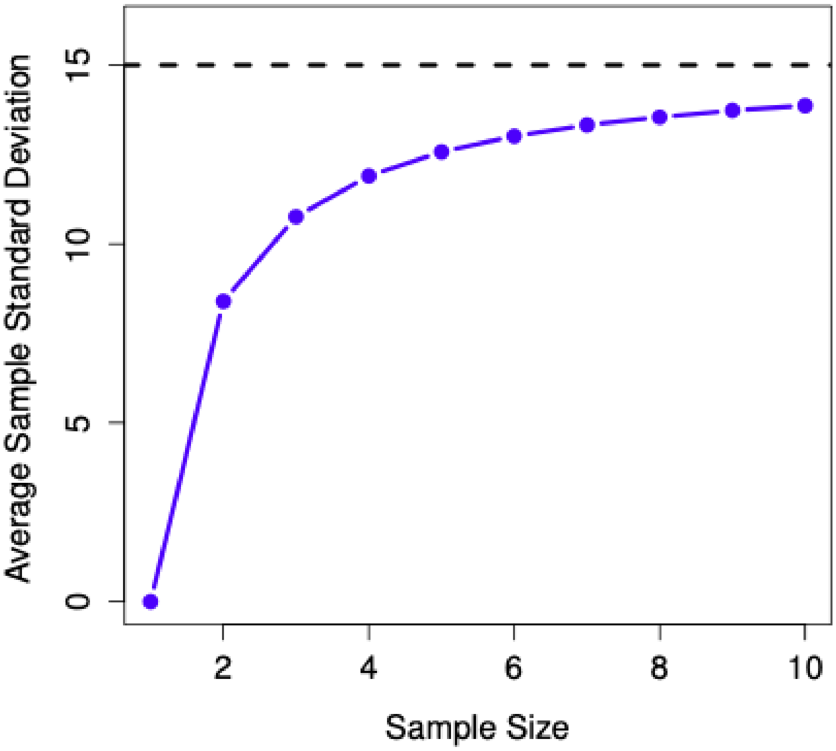


Figure 4.21: An illustration of the fact that the the sample standard deviation is a biased estimator of the population standard deviation.

In other words, if we want to make a “best guess” ($\hat{\sigma}$, our estimate of the population standard deviation) about the value of the population standard deviation σ, we should make sure our guess is a little bit larger than the sample standard deviation s.

The fix to this systematic bias turns out to be very simple. Here’s how it works. Before tackling the standard deviation, let’s look at the variance. If you recall from the second chapter, the sample variance is defined to be the average of the squared deviations from the sample mean. That is:

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

The sample variance s^2 is a biased estimator of the population variance σ^2. But as it turns out, we only need to make a tiny tweak to transform this into an unbiased estimator. All we have to do is divide by $N-1$ rather than by N. If we do that, we obtain the following formula:

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

This is an unbiased estimator of the population variance σ.

A similar story applies for the standard deviation. If we divide by $N-1$ rather than N, our estimate of the population standard deviation becomes:

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

.

It is worth pointing out that software programs make assumptions **for you**, about which variance and standard deviation **you** are computing. Some programs automatically divide by $N-1$, some do not. You need to check to figure out what they are doing. Don’t let the software tell you what to do. Software is for you telling it what to do.

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.

ℹ Note

Note, whether you should divide by N or N-1 also depends on your philosophy about what you are doing. For example, if you don’t think that what you are doing is estimating a population parameter, then why would you divide by N-1? Also, when N is large, it

> doesn't matter too much. The difference between a big N, and a big N-1, is just -1.

This is the right number to report, of course, it's that people tend to get a little bit imprecise about terminology when they write it up, because "sample standard deviation" is shorter than "estimated population standard deviation". It's no big deal, and in practice I do the same thing everyone else does. Nevertheless, I think it's important to keep the two **concepts** separate: it's never a good idea to confuse "known properties of your sample" with "guesses about the population from which it came". The moment you start thinking that s and $\hat{\sigma}$ are the same thing, you start doing exactly that.

To finish this section off, here's another couple of tables to help keep things clear:

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

4.14 Estimating a confidence interval

> Statistics means never having to say you're certain – Unknown origin

Up to this point in this chapter, we'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 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 in our estimate. It's not enough to be able guess that 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 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 that there is a 95% chance that a normally-distributed quantity will fall within two standard deviations of the true mean. To be more precise, we can use the **qnorm()** function to compute the 2.5th and 97.5th percentiles of the normal distribution

```
qnorm( p = c(.025, .975) ) [1] -1.959964 1.959964
```

Okay, so I lied earlier on. The more correct answer is that 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. Oof, that is a lot of mathy talk there. We'll clear it up, don't worry.

Mathematically, we write this as:

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

where the SEM is equal to $\sigma / \sqrt{N}$, and we can be 95% confident that this is true.

However, that's not answering the question that we're actually interested in. The equation above tells us what we should expect about the sample mean, given that we know what the population parameters are. What we **want** is to have this work the other way around: we want to know what we should believe about the population parameters, given that we have observed a particular sample. However, it's not too difficult to do this. Using a little high school algebra, a sneaky way to rewrite our equation is like this:

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

What this is telling 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:

$$\text{CI}_{95} = \bar{X} \pm \left(1.96 \times \frac{\sigma}{\sqrt{N}} \right)$$

Of course, there's nothing special about the number 1.96: it just happens to be the multiplier you need to use if you want a 95% confidence interval. If I'd wanted a 70% confidence interval, I could have used the **qnorm()** function to calculate the 15th and 85th quantiles:

```
qnorm( p = c(.15, .85) ) [1] -1.036433 1.036433
```

and so the formula for CI_{70} would be the same as the formula for CI_{95} except that we'd use 1.04 as our magic number rather than 1.96.

4.14.1 A slight mistake in the formula

As usual, I lied. The formula that I've given above for the 95% confidence interval is approximately correct, but I glossed over an important detail in the discussion. Notice my formula requires you to use the standard error of the mean, SEM, which in turn requires you to use the true population standard deviation σ.

Yet, before we stressed the fact that we don't actually **know** the true population parameters. Because we don't know the true value of σ, we have to use an estimate of the population standard deviation $\hat{\sigma}$ instead. This is pretty straightforward to do, but this has the consequence that we need to use the quantiles of the t-distribution rather than the normal distribution to calculate our magic number; and the answer depends on the sample size. Plus, we haven't really talked about the t distribution yet.

When we use the t distribution instead of the normal distribution, we get bigger numbers, indicating that we have more uncertainty. And why do we have that extra uncertainty? Well, because our estimate of the population standard deviation $\hat{\sigma}$ might be wrong! If it's wrong, it implies that we're a bit less sure about what our sampling distribution of the mean actually looks like… and this uncertainty ends up getting reflected in a wider confidence interval.

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

- Basic ideas about samples, sampling and populations
- Statistical theory of sampling: the law of large numbers, sampling distributions and the central limit theorem.
- Estimating means and standard deviations
- confidence intervals

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.

4.16 Videos

4.16.1 Introduction to Probability

Jeff has several more videos on probability that you can view on his statistics playlist.

4.16.2 Chebychev's Theorem

4.16.3 Z-scores

4.16.4 Normal Distribution I

4.16.5 Normal Distribution II

5 Foundations for inference

> Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. —Katie Crawford

So far we have been talking about describing data and looking for possible relationships between things we measure. We began with the problem of having too many numbers and discussed how they could be summarized with descriptive statistics, and communicated in graphs. We also looked at the idea of relationships between things. If one thing causes change in another thing, then if we measure how one thing goes up and down we should find that other thing goes up and down, or does something systematically following the first thing. At the end of the chapter on correlation, we showed how correlations, which imply a relationship between two things, are very difficult to interpret. Why? Because an observed correlation can be caused by a hidden third variable, or could be a spurious finding "caused" by random chance. In the last chapter, we talked about sampling from distributions, and we saw how samples can be different because of random error introduced by the sampling process.

Now we begin our journey into **inferential statistics**. These are tools used to make inferences about where our data came from, and to make inferences about what causes what.

In this chapter we provide some foundational ideas. We will stay mostly at a conceptual level, and use lots of simulations like we did in the last chapters. In the remaining chapters we formalize the intuitions built here to explain how some common inferential statistics work.

5.1 Brief review of Experiments

In chapter one we talked a about research methods and experiments. Experiments are a structured way of collecting data that can permit inferences about causality. If we wanted to know whether something like watching cats on YouTube increases happiness we would need an experiment. We already found out that just finding a bunch of people and measuring number of hours watching cats, and level of happiness, and correlating the two will not permit inferences about causation. For one, the causal flow could be reversed. Maybe being happy causes people to watch more cat videos. We need an experiment.

An experiment has two parts. A manipulation and a measurement. The manipulation is under the control of the experimenter. Manipulations are also called **independent variables**. For example, we could manipulate time spent watching cat videos: 1 hour versus 2 hours of cat videos. The measurement is the data that is collected. We could measure how happy people are after watching cat videos on a scale from 1 to 100. Measurements are also called **dependent variables**. So, in a basic experiment like the one above, we take measurements of happiness from people in one of two experimental conditions defined by the independent variable. Let's say we ran 50 subjects. 25 subjects would be randomly assigned to watch 1 hour of cat videos, and the other 25 subjects would be randomly assigned to watch 2 hours of cat videos. We would measure happiness for each subject at the end of the videos. Then we could look at the data.

What would we want to look at? If watching cat videos caused a change in happiness, then we would expect the measures of happiness for people watching 1 hour of cat videos to be different from the measures of happiness for people watching 2 hours of cat videos. If watching cat videos does not change happiness, then we would expect no differences in measures of happiness between conditions. Causal forces cause change, and the experiment is set up to detect the change.

Now we can state one overarching question, how do we know if the data changed between conditions? If we can be confident that there was a change between conditions, we can infer that our manipulation caused a changed in the measurement. If we cannot be confident there was a change, then we cannot infer that our manipulation caused a change in the measurement. We need to build some change detection tools so we can know a change when we find one.

"Hold on, if we are just looking for a change, wouldn't that be easy to see by looking at the numbers and seeing if they are different, what's so hard about that?". Good question. Now we must take a detour. The short answer is that there will always be change in the data (remember variance).

5.2 The data came from a distribution

In the last chapter we discussed samples and distributions, and the idea that you can take samples from distributions. So, from now on when you see a bunch of numbers, you should wonder, "where did these numbers come from?". What caused some kinds of numbers to happen more than other kinds of numbers. The answer to this question requires us to again veer off into the abstract world of distributions. A **distribution** a place where numbers can come from. The distribution sets the constraints. It determines what numbers are likely to occur, and what numbers are not likely to occur. Distributions are abstract ideas. But, they can be made concrete, and we can draw them with pictures that you have seen already, called histograms.

The next bit might seem slightly repetitive from the previous chapter. We again look at sampling numbers from a uniform distribution. We show that individual samples can look quite different from each other. Much of the beginning part of this chapter will already be familiar to you, but we take the concepts in a slightly different direction. The direction is how to make inferences about the role of chance in your experiment.

5.2.1 Uniform distribution

As a reminder from last chapter, Figure **??** shows that the shape of a uniform distribution is completely flat.

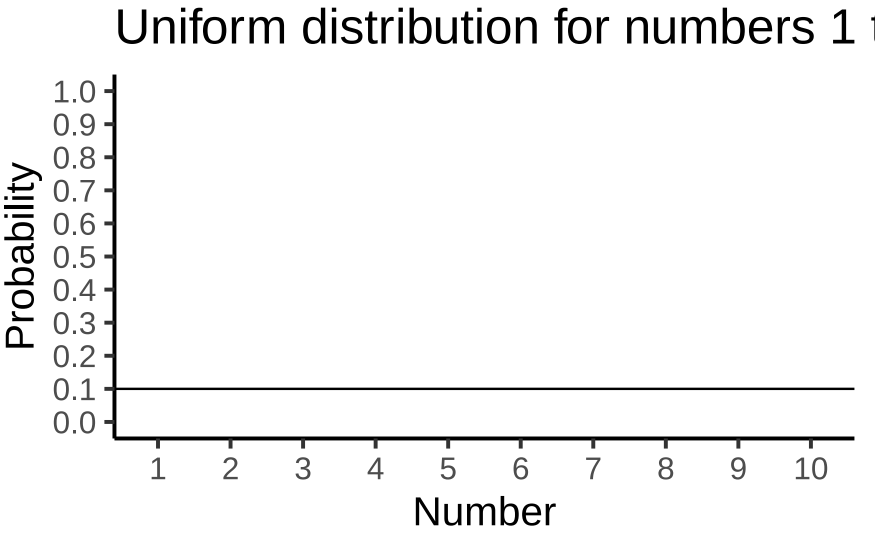


Figure 5.1: Uniform distribution showing that the numbers from 1 to 10 have an equal probability of being sampled

OK, so that doesn't look like much. What is going on here? The y-axis is labelled `probability`, and it goes from 0 to 1. The x-axis is labelled `Number`, and it goes from one to 10. There is a horizontal line drawn straight through. This line tells you the probability of each number from 1 to 10. Notice the line is flat. This means all of the numbers have the same probability of occurring. More specifically, there are 10 numbers from 1 to 10 (1,2,3,4,5,6,7,8,9,10), and they all have an equal chance of occurring. 1/10 = .1, which is the probability indicated by the horizontal line.

"So what?". Imagine that this uniform distribution is a number generating machine. It spits out numbers, but it spits out each number with the probability indicated by the line. If this distribution was going to start spitting out numbers, it would spit out 10% 1s, 10% 2s, 10% 3s, and so on, up to 10% 10s. Wanna see what that would look like? Let's make it spit out 100 numbers and put them in Table **??**.

We used the uniform distribution to generate these numbers. Officially, we call this **sampling** from a **distribution**. Sampling is what you do at a grocery store when there is free food. You

Table 5.1: 100 numbers randomly sampled from a uniform distribution.

9	6	5	8	3	6	7	8	6	2
8	6	7	7	2	6	4	3	1	3
6	6	7	3	6	8	8	4	7	2
8	5	2	2	3	5	10	5	7	8
4	7	4	9	4	9	3	1	2	3
7	6	2	10	9	8	8	5	6	6
2	7	4	4	8	5	8	4	8	9
4	7	3	9	4	5	2	4	7	7
7	2	4	2	4	8	1	4	1	10
8	4	2	10	6	3	8	6	7	3

can keep taking more. However, if you take all of the samples, then what you have is called the **population**. We'll talk more about samples and populations as we go along.

Because we used the uniform distribution to create numbers, we already know where our numbers came from. However, we can still pretend for the moment that someone showed up at your door, showed you these numbers, and then you wondered where they came from. Can you tell just by looking at these numbers that they came from a uniform distribution? What would need to look at? Perhaps you would want to know if all of the numbers occur with roughly equal frequency, after all they should have right? That is, if each number had the same chance of occurring, we should see that each number occurs roughly the same number of times.

We already know what a histogram is, so we can put our sample of 100 numbers into a histogram and see what the counts look like. If all of the numbers from 1 to 10 occur with equal frequency, then each individual number should occur about 10 times. Figure **??** shows the histogram:

Uh oh, as you can see, not all of the number occurred 10 times each. All of the bars are not the same height. This shows that randomly sampling numbers from this distribution does not guarantee that our numbers will be exactly like the distribution they came from. We can call this sampling error, or sampling variability.

5.2.2 Not all samples are the same, they are usually quite different

Let's look at sampling error more closely. We will sample 20 numbers from the uniform distribution. We should expect that each number between 1 and 10 occurs about two times each. As before, this expectation can be visualized in a histogram. To get a better sense of sampling error, let's repeat the above process ten times. Figure **??** has 10 histograms, each showing what 10 different samples of twenty numbers looks like:

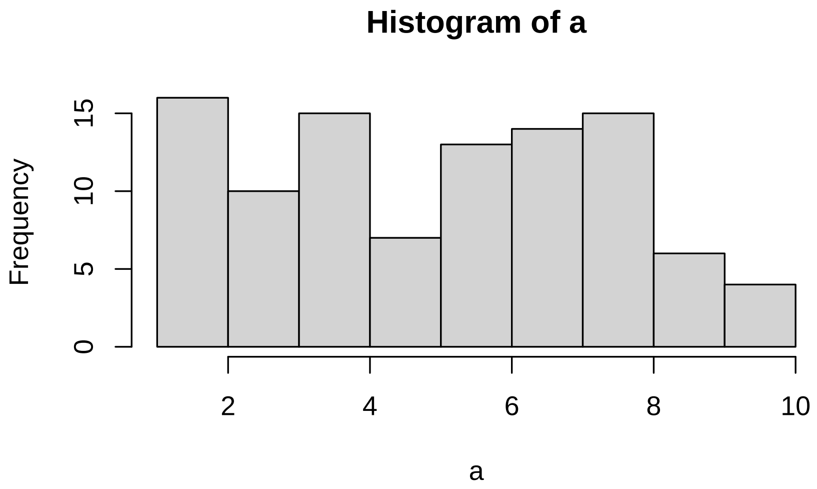


Figure 5.2: Histogram of 100 numbers randomly sampled from the uniform distribution containing the integers from 1 to 10

You might notice right away that none of the histograms are the same. Even though we are randomly taking 20 numbers from the very same uniform distribution, each sample of 20 numbers comes out different. This is sampling variability, or sampling error.

?@fig-5expectedUnif shows an animated version of the process of repeatedly choosing 20 new random numbers and plotting a histogram. The horizontal line shows the flat-line shape of the uniform distribution. The line crosses the y-axis at 2; and, we expect that each number (from 1 to 10) should occur about 2 times each in a sample of 20. However, each sample bounces around quite a bit, due to random chance.

Looking at the above histograms shows us that figuring out where our numbers came from can be difficult. In the real world, our measurements are samples. We usually only have the luxury of getting one sample of measurements, rather than repeating our own measurements 10 times or more. If you look at the histograms, you will see that some of them look like they could have come from the uniform distribution: most of the bars are near two, and they all fall kind of on a flat line. But, if you happen to look at a different sample, you might see something that is very bumpy, with some numbers happening way more than others. This could suggest to you that those numbers did not come from a uniform distribution (they're just too bumpy). But let me remind you, all of these samples came from a uniform distribution, this is what samples from that distribution look like. This is what chance does to samples, it makes the individual data points noisy.

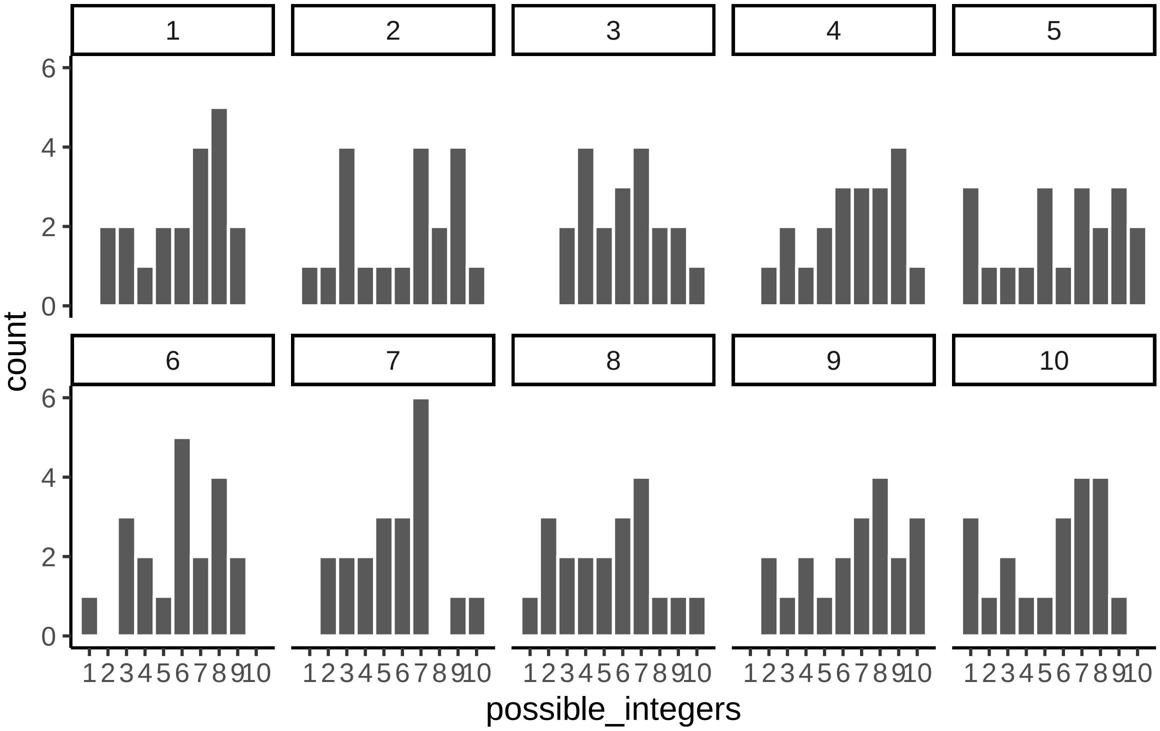


Figure 5.3: Histograms for 10 different samples from the uniform distribution. Each sample contains 20 numbers. The histograms all look quite different. The differences between the samples are due to sampling error or random chance.

5.2.3 Large samples are more like the distribution they came from

Let's refresh the question. Which of the two samples in Figure **??** do you think came from a uniform distribution?

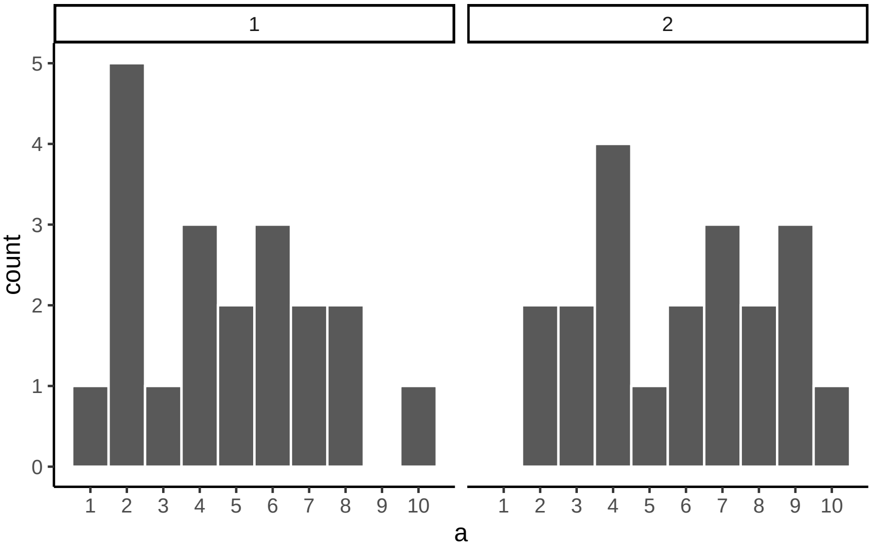


Figure 5.4: Which of these two samples came from a uniform distribution?

The answer is that they both did. But, neither of them look like they did.

Can we improve things, and make it easier to see if a sample came from a uniform distribution? Yes, we can. All we need to do is increase the **sample-size**. We will often use the letter `n` to refer to sample-size. N is the number of observations in the sample.

So let's increase the number of observations in each sample from 20 to 100. We will again create 10 samples (each with 100 observations), and make histograms for each of them. All of these samples will be drawn from the very same uniform distribution. This, means we should expect each number from 1 to 10 to occur about 10 times in each sample. The histograms are shown in Figure **??**.

Again, most of these histograms don't look very flat, and all of the bars seem to be going up or down, and they are not exactly at 10 each. So, we are still dealing with sampling error. It's a pain. It's always there.

Let's bump up the N from 100 to 1000 observations per sample. Now we should expect every number to appear about 100 times each. What happens?

Figure **??** shows the histograms are starting to flatten out. The bars are still not perfectly at 100, because there is still sampling error (there always will be). But, if you found a histogram

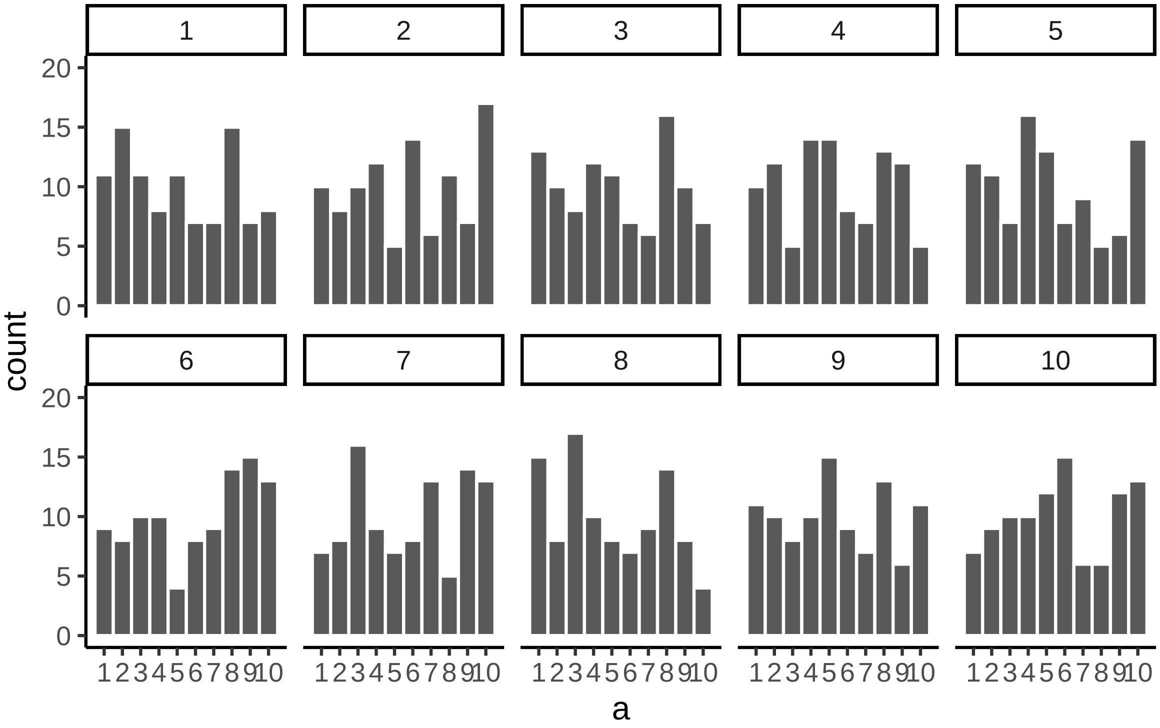


Figure 5.5: Histograms for different samples from a uniform distribution. N = 100 for each sample.

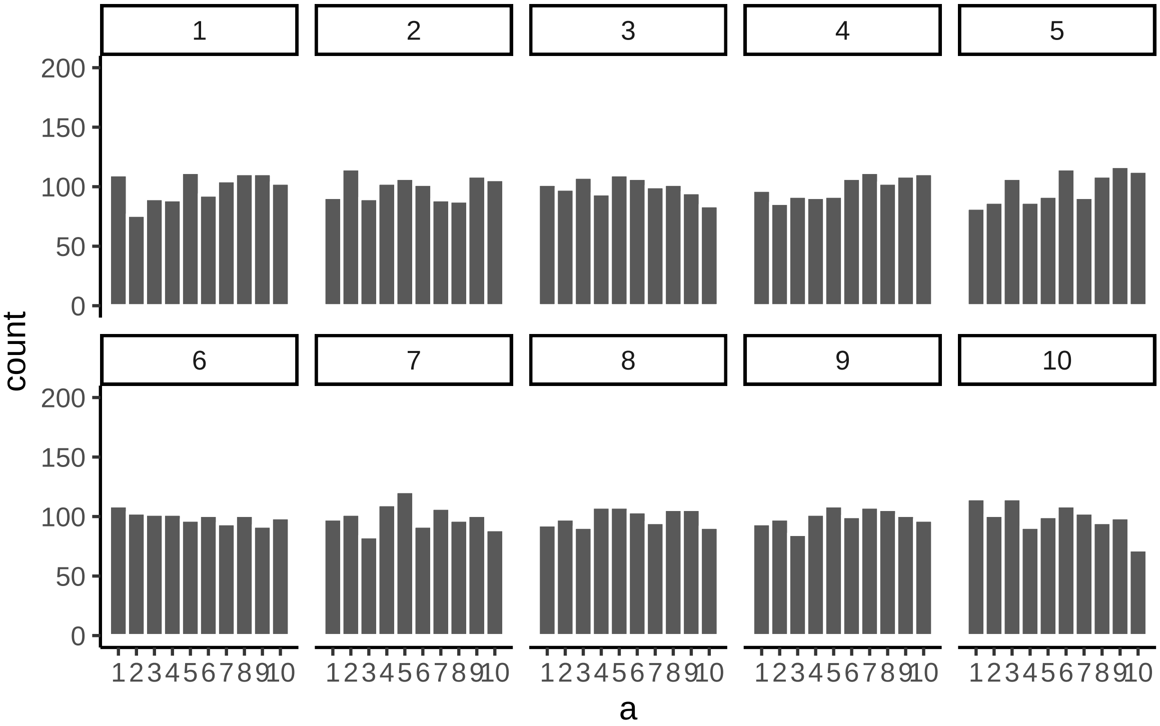


Figure 5.6: Histograms for different samples from a uniform distribution. N = 1000 for each sample.

that looked flat and knew that the sample contained many observations, you might be more confident that those numbers came from a uniform distribution.

Just for fun let's make the samples really big. Say 100,000 observations per sample. Here, we should expect that each number occurs about 10,000 times each. What happens?

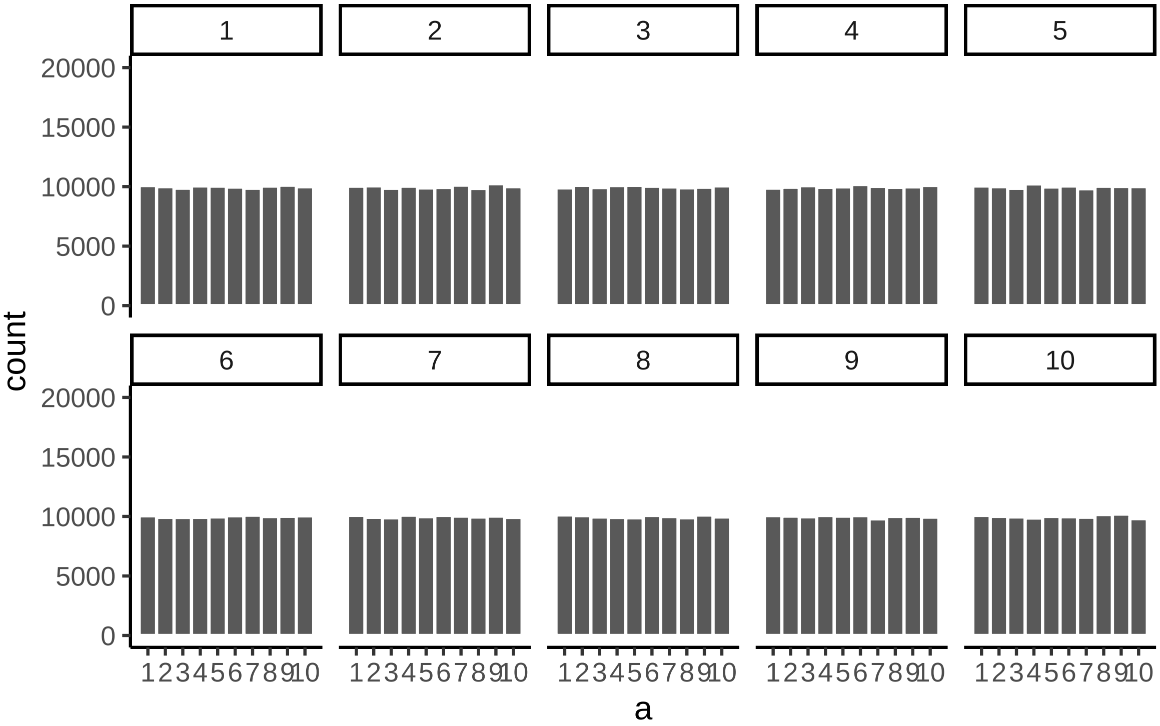


Figure 5.7: Histograms for different samples from a uniform distribution. N = 100,000 for each sample.

Figure **??** shows that the histograms for each sample are starting to look the same. They all have 100,000 observations, and this gives chance enough opportunity to equally distribute the numbers, roughly making sure that they all occur very close to the same amount of times. As you can see, the bars are all very close to 10,000, which is where they should be if the sample came from a uniform distribution.

Pro tip

The pattern behind a sample will tend to stabilize as sample-size increases. Small samples will have all sorts of patterns because of sampling error (chance).

Before getting back to the topic of experiments that we started with, let's ask two more questions. First, which of the two samples in Figure **??** do you think came from a uniform distribution? FYI, each of these samples had 20 observations each.

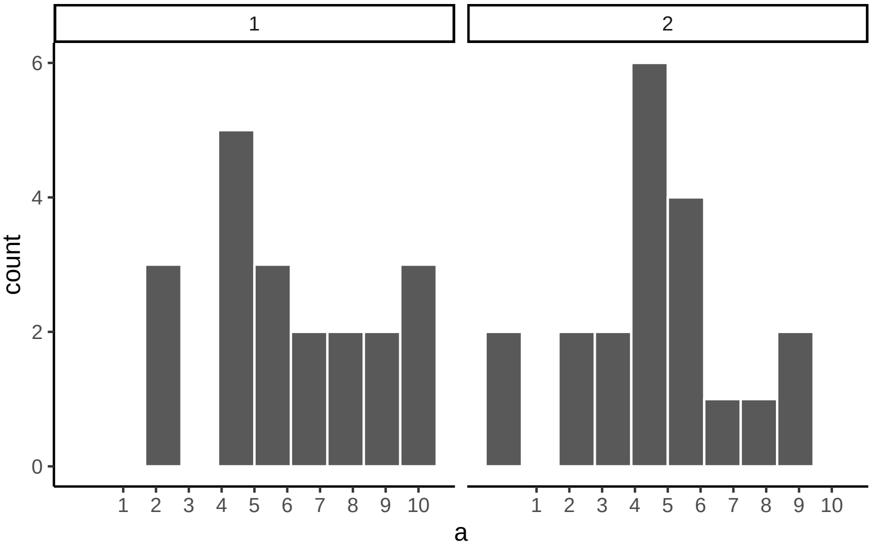


Figure 5.8: Which of these samples came from a uniform distribution?

If you are not confident in the answer, this is because **sampling error** (randomness) is fuzzing with the histograms.

Here is the very same question, only this time we will take 1,000 observations for each sample. Which histogram in Figure **??** do you think came from a uniform distribution, which one did not?

Now that we have increased N, we can see the pattern in each sample becomes more obvious. The histogram for sample 1 has bars near 100, not perfectly flat, but it resembles a uniform distribution. The histogram for sample 2 is not flat looking at all.

Congratulations to Us! We have just made some statistical inferences without using formulas!

"We did?" Yes, by looking at our two samples we have inferred that sample 2 did not come from a uniform distribution. We have also inferred that sample 1 could have come form a uniform distribution. Fantastic. These are the same kinds of inferences we will be making for the rest of the course. We will be looking at some numbers, wondering where they came from, then we will arrange the numbers in such a way so that we can make inferences about the kind of distribution they came from. That's it.

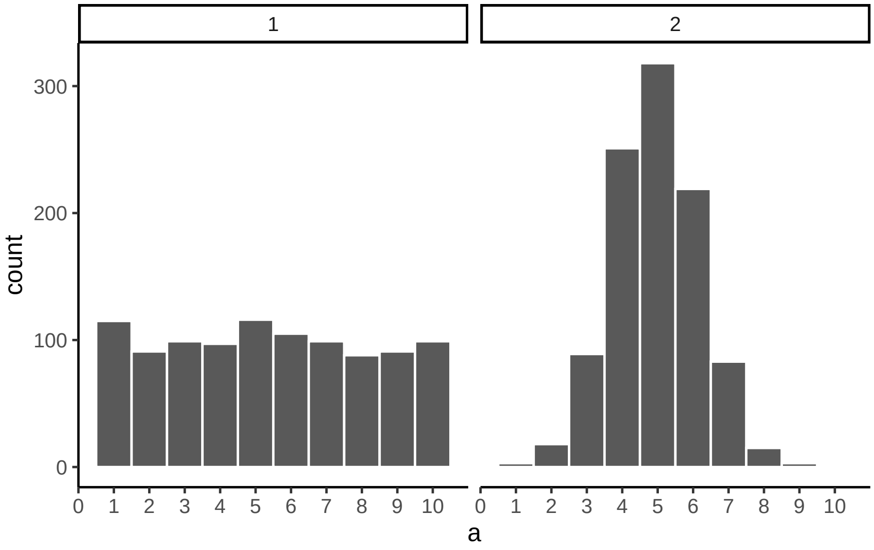


Figure 5.9: Which of these samples came from a uniform distribution?

5.3 Is there a difference?

Let's get back to experiments. In an experiment we want to know if an independent variable (our manipulation) causes a change in a dependent variable (measurement). If this occurs, then we will expect to see some differences in our measurement as a function of the manipulation.

Consider the light switch example:

Light Switch Experiment: You manipulate the switch up (condition 1 of independent variable), light goes on (measurement). You manipulate the switch down (condition 2 of independent variable), light goes off (another measurement). The measurement (light) changes (goes off and on) as a function of the manipulation (moving switch up or down).

You can see the change in measurement between the conditions, it is as obvious as night and day. So, when you conduct a manipulation, and can see the difference (change) in your measure, you can be pretty confident that your manipulation is causing the change.

> note: to be cautious we can say "something" about your manipulation is causing the change, it might not be what you think it is if your manipulation is very complicated and involves lots of moving parts.

5.3.1 Chance can produce differences

Do you think random chance can produce the appearance of differences, even when there really aren't any? I hope so. We have already shown that the process of sampling numbers from a distribution is a chancy process that produces different samples. Different samples are different, so yes, chance can produce differences. This can muck up our interpretation of experiments.

Let's conduct a fictitious experiment where we expect to find no differences, because we will manipulate something that shouldn't do anything. Here's the set-up:

You are the experimenter standing in front of a gumball machine. It is very big, has thousands of gumballs. 50% of the gumballs are green, and 50% are red. You want to find out if picking gumballs with your right hand vs. your left hand will cause you to pick more green gumballs. Plus, you will be blindfolded the entire time. The independent variable is Hand: right hand vs. left hand. The dependent variable is the measurement of the color of each gumball.

You run the experiment as follows. 1) put on blind fold. 2) pick 10 gumballs randomly with left hand, set them aside. 3) pick 10 gumballs randomly with right hand, set them aside. 4) count the number of green and red gumballs chosen by your left hand, and count the number of green and red gumballs chosen by your right hand. Hopefully you will agree that your hands will not be able to tell the difference between the gumballs. If you don't agree, we will further stipulate the gumballs are completely identical in every way except their color, so it would be impossible to tell them apart using your hands. So, what should happen in this experiment?

"Umm, maybe you get 5 red gum balls and 5 green balls from your left hand, and also from your right hand?". Sort of yes, this is what you would usually get. But, it is not all that you can get. Here is some data showing what happened from one pretend experiment:

hand	gumball
left	0
left	0
left	1
left	1
left	0
left	0
left	1
left	1
left	0
left	0
right	0
right	1
right	0
right	0
right	0
right	1
right	1
right	1
right	0
right	0

"What am I looking at here". This is a long-format table. Each row is one gumball. The first column tells you what hand was used. The second column tells you what kind of gumball. We will say 1s stand for green gum balls, and 0s stand for red gumballs. So, did your left hand cause you to pick more green gumballs than your right hand?

It would be easier to look at the data using a bar graph (Figure **??**). To keep things simple, we only count the green gumballs (the other gumballs must be red). So, all we need to do is sum up the 1s. The 0s won't add anything.

Oh look, the bars are not the same. One hand picked more green gum balls than the other. Does this mean that one of your hands secretly knows how to find green gumballs? No, it's just another case of sampling error, that thing we call luck or chance. The difference here is caused by chance, not by the manipulation (which hand you use). **Major problem for inference alert**. We run experiments to look for differences so we can make inferences about whether our manipulations cause change in our measures. However, this example demonstrates that we can find differences by chance. How can we know if a difference is real, or just caused by chance?

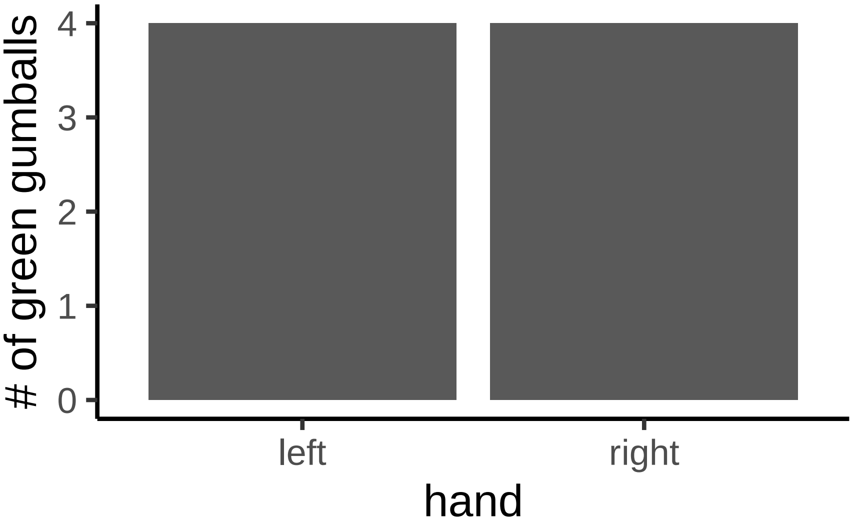


Figure 5.10: Counts of green gumballs picked randomly by each hand.

5.3.2 Differences due to chance can be simulated

Remember when we showed that chance can produce correlations. We also showed that chance is restricted in its ability to produce correlations. For example, chance more often produces weak correlations than strong correlations. Remember the window of chance? We found out before that correlations falling outside the window of chance were very unlikely. We can do the same thing for differences. Let's find out just what chance can do in our experiment. Once we know what chance is capable of we will be in a better position to judge whether our manipulation caused a difference, or whether it could have been chance.

The first thing to do is pretend you conduct the gumball experiment 10 times in a row. This will produce 10 different sets of results. Figure **??** shows bar graphs for each replication of the experiment. Now we can look at whether the left hand chose more green gumballs than red gumballs.

These 10 experiments give us a better look at what chance can do. It should also mesh well with your expectations. If everything is determined by chance (as we have made it so), then sometimes your left hand will choose more green balls, sometimes your right hand will choose more green gumballs, and sometimes they will choose the same amount of gumballs. Right? Right.

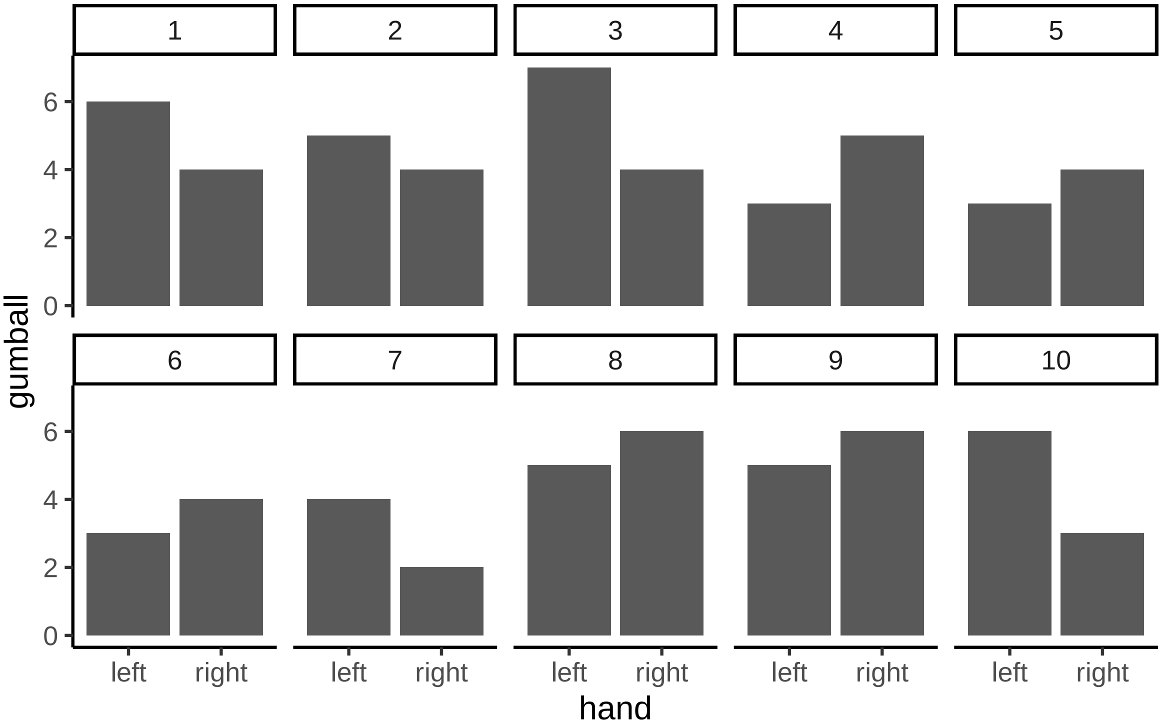


Figure 5.11: 10 simulated replications of picking gumballs. Each replication gives a slightly different answer. Any difference between the bars is due to chance, or sampling error. This shows that chance alone can produce differences, just by the act of sampling.

5.4 Chance makes some differences more likely than others

OK, we have seen that chance can produce differences here. But, we still don't have a good idea about what chance usually does and doesn't do. For example, if we could find the window of opportunity here, we would be able find out that chance usually does not produce differences of a certain large size. If we knew what the size was, then if we ran experiment and our difference was bigger than what chance can do, we could be confident that chance did not produce our difference.

Let's think about our measure of green balls in terms of a difference. For example, in each experiment we counted the green balls for the left and right hand. What we really want to know is if there is a difference between them. So, we can calculate the **difference score**. Let's decide that the difference score = # of green gumballs in left hand - # of green gumballs in right hand. Figure **??** redraws the 10 bar graphs from above; however, now there is only one bar for each experiment. This bar represents the difference in number of green gumballs drawn by the left and right hand.

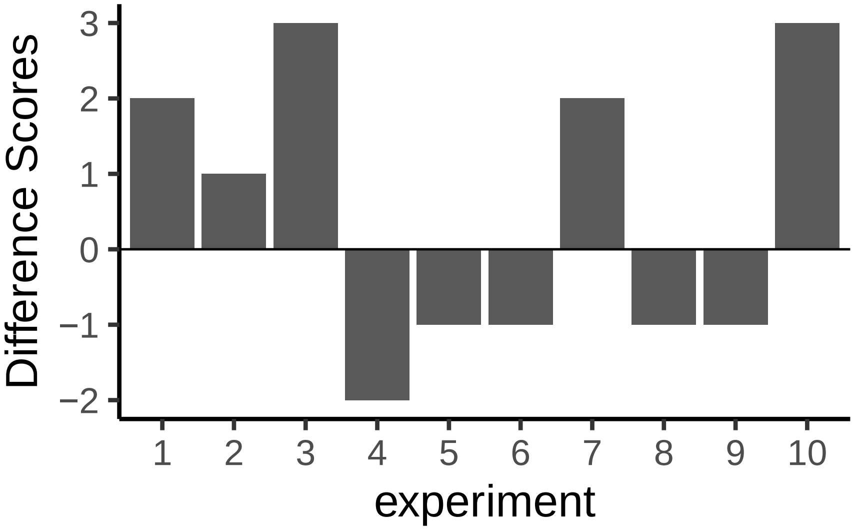


Figure 5.12: A look at the differences between number of each kind of gumball for the different replications. The difference should be zero, but sampling error produces non-zero differences.

Missing bars mean that there were an equal number of green gumballs chosen by the left and right hands (difference score is 0). A positive value means that more green gumballs were chosen by the left than right hand. A negative value means that more green gumballs were chosen by the right than left hand. Note that if we decided (and we get to decide) to calculate the difference in reverse (right hand - left hand), the signs of the differences scores would flip around.

We are starting to see more of the differences that chance can produce. The difference scores are mostly between -2 to +2. We could get an even better impression by running this pretend experiment 100 times instead of only 10 times. The results are shown in Figure **??**.

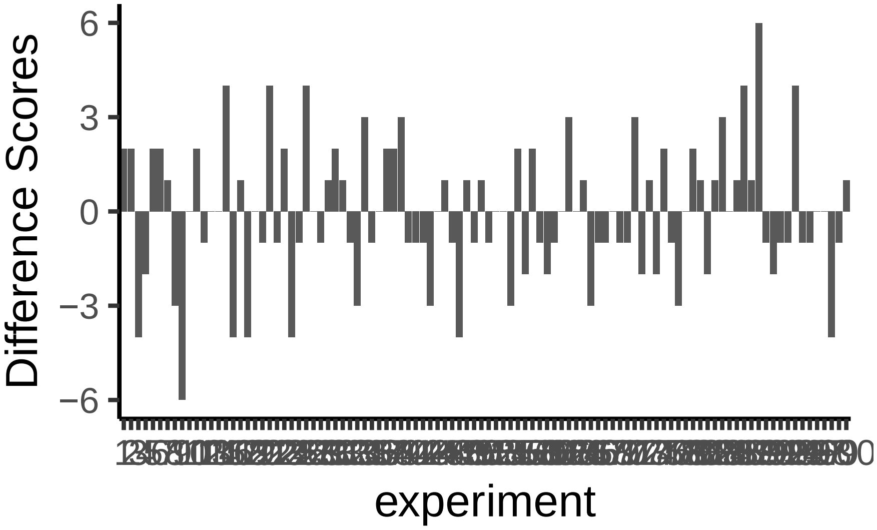


Figure 5.13: Replicating the experiment 100 times, and looking at the differences each time. There are mnay kinds of differences that chance alone can produce.

Ooph, we just ran so many simulated experiments that the x-axis is unreadable, but it goes from 1 to 100. Each bar represents the difference of number of green balls chosen randomly by the left or right hand. Beginning to notice anything? Look at the y-axis, this shows the size of the difference. Yes, there are lots of bars of different sizes, this shows us that many kinds of differences do occur by chance. However, the y-axis is also restricted. It does not go from -10 to +10. Big differences greater than 5 or -5 don't happen very often.

Now that we have a method for simulating differences due to chance, let's run 10,000 simulated experiments. But, instead of plotting the differences in a bar graph for each experiment, how about we look at the histogram of difference scores. The histogram in Figure **??** provides a clearer picture about which differences happen most often, and which ones do not. This will be another window into observing what kinds of differences chance is capable of producing.

Our computer simulation allows us to force chance to operate hundreds of times, each time it produces a difference. We record the difference, then at the end of the simulation we plot the histogram of the differences. The histogram begins to show us the where the differences came from. Remember the idea that numbers come from a distribution, and the distribution says how often each number occurs. We are looking at one of these distributions. It is showing us that chance produces some differences more often than others. First, chance usually produces 0 differences, that's the biggest bar in the middle. Chance also produces larger differences, but as the differences get larger (positive or negative), they occur less frequently. The shape

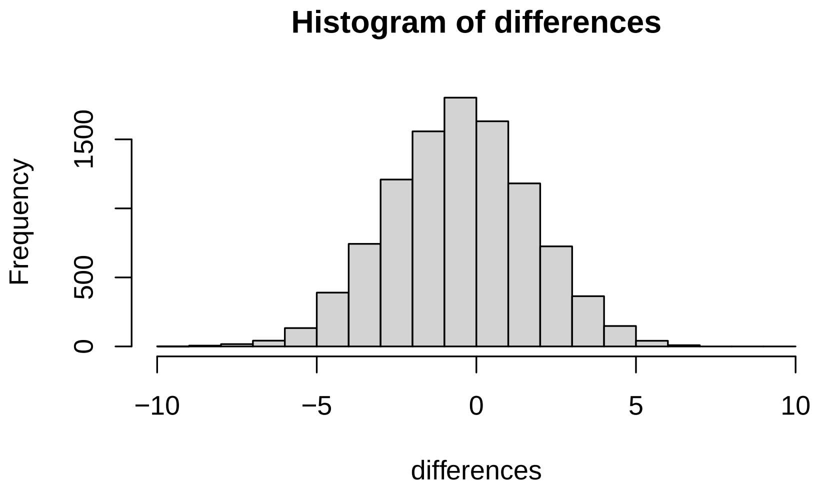


Figure 5.14: A histogram of the differences obtained by chance over 10,000 replications. The most frequency difference is 0, which is what we expect by chance. But the differences can be as large as -10 or +10. Larger differences occur less often by chance. Chance can't do everything.

of this histogram is your chance window, it tells you what chance can do, it tells you what chance usually does, and what it usually does not do.

You can use this chance window to help you make inferences. If you ran yourself in the gumball experiment and found that your left hand chose 2 more green gumballs than red gumballs, would you conclude that you left hand was special, and caused you to choose more green gumballs? Hopefully not. You could look at the chance window and see that differences of size +2 do happen fairly often by chance alone. You should not be surprised if you got a +2 difference. However, what if your left chose 5 more green gumballs than red gumballs. Well, chance doesn't do this very often, you might think something is up with your left hand. If you got a whopping 9 more green gumballs than red gumballs, you might really start to wonder. This is the kind of thing that could happen (it's possible), but virtually never happens by chance. When you get things that almost never happen by chance, you can be more confident that the difference reflects a causal force that is not chance.

5.5 The Crump Test

We are going to be doing a lot of inference throughout the rest of this course. Pretty much all of it will come down to one question. Did chance produce the differences in my data? We will be talking about experiments mostly, and in experiments we want to know if our manipulation caused a difference in our measurement. But, we measure things that have natural variability, so every time we measure things we will always find a difference. We want to know if the difference we found (between our experimental conditions) could have been produced by chance. If chance is a very unlikely explanation of our observed difference, we will make the inference that chance did not produce the difference, and that something about our experimental manipulation did produce the difference. This is it (for this textbook).

> ℹ Note
>
> Statistics is not only about determining whether chance could have produced a pattern in the observed data. The same tools we are talking about here can be generalized to ask whether any kind of distribution could have produced the differences. This allows comparisons between different models of the data, to see which one was the most likely, rather than just rejecting the unlikely ones (e.g., chance). But, we'll leave those advanced topics for another textbook.

This chapter is about building intuitions for making these kinds of inferences about the role of chance in your data. It's not clear to me what are the best things to say, to build up your intuitions for how to do statistical inference. So, this chapter tries different things, some of them standard, and some of them made up. What you are about to read, is a made up way of doing statistical inference, without using the jargon that we normally use to talk about it. The goal is to do things without formulas, and without probabilities, and just work with some ideas using simulations to see what happens. We will look at what chance can do, then we will talk about what needs to happen in your data in order for you to be confident that chance didn't do it.

5.5.1 Intuitive methods

Warning, this is an unofficial statistical test made up by Matt Crump. It makes sense to him (me), and if it turns out someone else already made this up, then Crump didn't do his homework, and we will change the name of this test to it's original author later on. The point of this test is to show how simple arithmetic operations that you already understand can be used to create a statistic tool for inference. This test uses:

1. Sampling numbers randomly from a distribution
2. Adding and subtracting
3. Division, to find the mean
4. Counting
5. Graphing and drawing lines
6. NO FORMULAS

5.5.2 Part 1: Frequency based intuition about occurrence

Question: How many times does something need to happen for it to happen a lot? Or, how many times does something need to happen for it to happen not very much, or even really not at all? Small enough for you to not worry about it at all happening to you?

Would you go outside everyday if you thought that you would get hit by lightning 1 out of 10 times? I wouldn't. You'd probably be hit by lightning more than once per month, you'd be dead pretty quickly. 1 out of 10 is a lot (to me, maybe not to you, there's no right answer here).

Would you go outside everyday if you thought that you would get hit by lightning 1 out of every 100 days? Jeez, that's a tough one. What would I even do? If I went out everyday I'd probably be dead in a year! Maybe I would go out 2 or 3 times per year, I'm risky like that, but I'd probably live longer if I stayed at home forever. It would massively suck.

Would you go outside everyday if you thought you would get hit by lightning 1 out of every 1000 days? Well, you'd probably be dead in 3-6 years if you did that. Are you a gambler? Maybe go out once per month, still sucks.

Would you go outside everyday if you thought lightning would get you 1 out every 10,000 days? 10,000 is a bigger number, harder to think about. It translates to getting hit about once every 27 years. Ya, I'd probably go out 150 days per year, and keep my fingers crossed.

Would you go outside everyday if you thought lightning would get you 1 out every 100,000 days? How many years is that? It's about 273 years. With those odds, I'd probably go out all the time and forget about being hit by lightning. It doesn't happen very often.

The point of considering these questions is to get a sense for yourself of what happens a lot, and what doesn't happen a lot, and how you would make important decisions based on what happens a lot and what doesn't.

5.5.3 Part 2: Simulating chance

This next part could happen a bunch of ways, I'll make loads of assumptions that I won't defend, and I won't claim the Crump test has problems. I will claim it helps us make an inference about whether chance could have produced some differences in data. We've already been introduced to simulating things, so we'll do that again. Here is what we will do. I am a cognitive psychologist who happens to be measuring X. Because of prior research in the field, I know that when I measure X, my samples will tend to have a particular mean and standard deviation. Let's say the mean is usually 100, and the standard deviation is usually 15. In this case, I don't care about using these numbers as estimates of the population parameters, I'm just thinking about what my samples usually look like. What I want to know is how they behave when I sample them. I want to see what kind of samples happen a lot, and what kind of samples don't happen a lot. Now, I also live in the real world, and in the real world when I run experiments to see what changes X, I usually only have access to some number of participants, who I am very grateful too, because they participate in my experiments. Let's say I usually can run 20 subjects in each condition in my experiments. Let's keep the experiment simple, with two conditions, so I will need 40 total subjects.

I would like to learn something to help me with inference. One thing I would like to learn is what the sampling distribution of the sample mean looks like. This distribution tells me what kinds of mean values happen a lot, and what kinds don't happen very often. But, I'm actually going to skip that bit. Because what I'm really interested in is what the **sampling distribution of the difference between my sample means** looks like. After all, I am going to run an experiment with 20 people in one condition, and 20 people in the other. Then I am going to calculate the mean for group A, and the mean for group B, and I'm going to look a the difference. I will probably find a difference, but my question is, did my manipulation cause this difference, or is this the kind of thing that happens a lot by chance. If I knew what chance can do, and how often it produces differences of particular sizes, I could look at the difference I observed, then look at what chance can do, and then I can make a decision! If my difference doesn't happen a lot (we'll get to how much not a lot is in a bit), then I might be willing to believe that my manipulation caused a difference. If my difference happens all the time by chance alone, then I wouldn't be inclined to think my manipulation caused the difference, because it could have been chance.

So, here's what we'll do, even before running the experiment. We'll do a simulation. We will sample numbers for group A and Group B, then compute the means for group A and group B, then we will find the difference in the means between group A and group B. But, we will do one very important thing. We will pretend that we haven't actually done a manipulation. If we do this (do nothing, no manipulation that could cause a difference), then we know that **only sampling error** could cause any differences between the mean of group A and group B. We've eliminated all other causes, only chance is left. By doing this, we will be able to see exactly what chance can do. More importantly, we will see the kinds of differences that occur a lot, and the kinds that don't occur a lot.

Before we do the simulation, we need to answer one question. How much is a lot? We could pick any number for a lot. I'm going to pick 10,000. That is a lot. If something happens only 1 times out 10,000, I am willing to say that is not a lot.

OK, now we have our number, we are going to simulate the possible mean differences between group A and group B that could arise by chance. We do this 10,000 times. This gives chance a lot of opportunity to show us what it does do, and what it does not do.

This is what I did: I sampled 20 numbers into group A, and 20 into group B. The numbers both came from the same normal distribution, with mean = 100, and standard deviation = 15. Because the samples are coming from the same distribution, we expect that on average they will be similar (but we already know that samples differ from one another). Then, I compute the mean for each sample, and compute the difference between the means. I save the **mean difference score**, and end up with 10,000 of them. Then, I draw the histogram in Figure **??**.

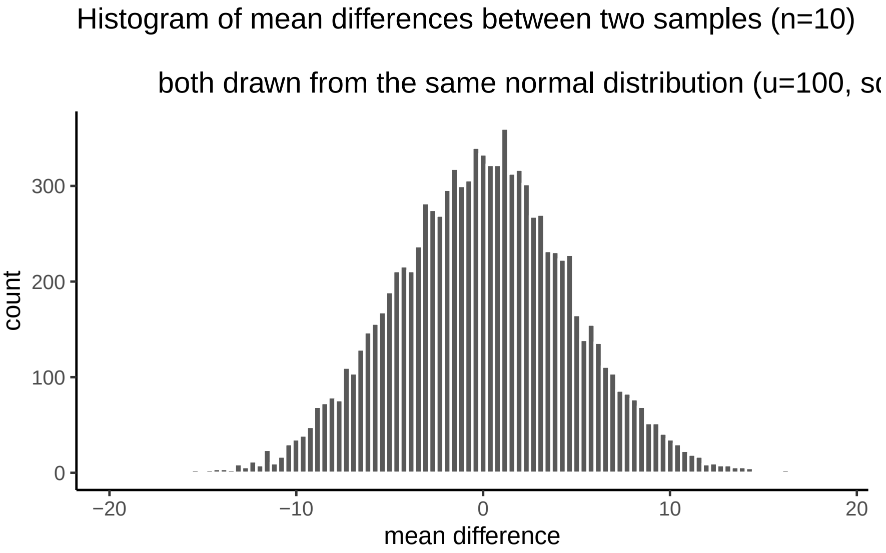


Figure 5.15: Histogram of mean differences arising by chance.

Note

Of course, we might recognize that chance could do a difference greater than 15. We just didn't give it the opportunity. We only ran the simulation 10,000 times. If we ran it a million times, maybe a difference greater than 15 or even 20 would happen a couple times. If we ran it a bazillion gazillion times, maybe a difference greater than 30 would happen a couple times. If we go out to infinity, then chance might produce all sorts of bigger differences once in a while. But, we've already decided that 1/10,000 is not a lot. So things that happen 0 out of 10,000 times, like differences greater than 15, are considered to be extremely unlikely.

Now we can see what chance can do to the size of our mean difference. The x-axis shows the size of the mean difference. We took our samples from the sample distribution, so the difference between them should usually be 0, and that's what we see in the histogram.

Pause for a second. Why should the mean differences usually be zero, wasn't the population mean = 100, shouldn't they be around 100? No. The mean of group A will tend to be around 100, and the mean of group B will tend be around 100. So, the difference score will tend to be 100-100 = 0. That is why we expect a mean difference of zero when the samples are drawn from the same population.

So, differences near zero happen the most, that's good, that's what we expect. Bigger or smaller differences happen increasingly less often. Differences greater than 15 or -15 never happen at all. For our purposes, it looks like chance only produces differences between -15 to 15.

OK, let's ask a couple simple questions. What was the biggest negative number that occurred in the simulation? We'll use R for this. All of the 10,000 difference scores are stored in a variable I made called `difference`. If we want to find the minimum value, we use the `min` function. Here's the result.

```
min(difference)
#> [1] -19.67353
```

OK, so what was the biggest positive number that occurred? Let's use the `max` function to find out. It finds the biggest (maximum) value in the variable. FYI, we've just computed the range, the minimum and maximum numbers in the data. Remember we learned that before. Anyway, here's the max.

```
max(difference)
#> [1] 18.46874
```

Both of these extreme values only occurred once. Those values were so rare we couldn't even see them on the histogram, the bar was so small. Also, these biggest negative and positive numbers are pretty much the same size if you ignore their sign, which makes sense because the distribution looks roughly symmetrical.

So, what can we say about these two numbers for the min and max? We can say the min happens 1 times out of 10,000. We can say the max happens 1 times out of 10,000. Is that a lot of times? Not to me. It's not a lot.

So, how often does a difference of 30 (much larger larger than the max) occur out of 10,000. We really can't say, 30s didn't occur in the simulation. Going with what we got, we say 0 out of 10,000. That's never.

We're about to move into part three, which involves drawing decision lines and talking about them. The really important part about part 3 is this. What would you say if you ran this experiment once, and found a mean difference of 30? I would say it happens 0 times of out 10,000 by chance. I would say chance did not produce my difference of 30. That's what I would say. We're going to expand upon this right now.

5.5.4 Part 3: Judgment and Decision-making

Remember, we haven't even conducted an experiment. We're just simulating what could happen if we did conduct an experiment. We made a histogram. We can see that chance produces some differences more than others, and that chance never produced really big differences. What should we do with this information?

What we are going to do is talk about judgment and decision making. What kind of judgment and decision making? Well, when you finally do run an experiment, you will get two means for group A and B, and then you will need to make some judgments, and perhaps even a decision, if you are so inclined. You will need to judge whether chance (sampling error) could have produced the difference you observed. If you judge that it did it not, you might make the decision to tell people that your experimental manipulation actually works. If you judge that it could have been chance, you might make a different decision. These are important decisions for researchers. Their careers can depend on them. Also, their decisions matter for the public. Nobody wants to hear fake news from the media about scientific findings.

So, what we are doing is preparing to make those judgments. We are going to draw up a plan, before we even see the data, for how we will make judgments and decisions about what we find. This kind of planning is extremely important, because we discuss in part 4, that your planning can help you design an even better experiment than the one you might have been intending to run. This kind of planning can also be used to interpret other people's results, as a way of double-checking checking whether you believe those results are plausible.

The thing about judgement and decision making is that reasonable people disagree about how to do it, unreasonable people really disagree about it, and statisticians and researchers disagree about how to do it. I will propose some things that people will disagree with. That's OK, these things still make sense. And, the disagreeable things point to important problems that are very real for any "real" statistical inference test.

Let's talk about some objective facts from our simulation of 10,000 things that we definitely know to be true. For example, we can draw some lines on the graph, and label some different regions. We'll talk about two kinds of regions.

1. Region of chance. Chance did it. Chance could have done it
2. Region of not chance. Chance didn't do it. Chance couldn't have done it.

The regions are defined by the minimum value and the maximum value. Chance never produced a smaller or bigger number. The region inside the range is what chance did do, and the the region outside the range on both sides is what chance never did. It looks like Figure **??**:

We have just drawn some lines, and shaded some regions, and made one plan we could use to make decisions. How would the decisions work. Let's say you ran the experiment and found a mean difference between groups A and B of 25. Where is 25 in the figure? It's in the green part. What does the green part say? NOT CHANCE. What does this mean. It means chance never made a difference of 25. It did that 0 out of 10,000 times. If we found a difference of 25, perhaps we could confidently conclude that chance did not cause the difference. If I found a difference of 25 with this kind of data, I'd be pretty confident chance did not cause the difference; and, I would give myself license to consider that my experimental manipulation may be causing the difference.

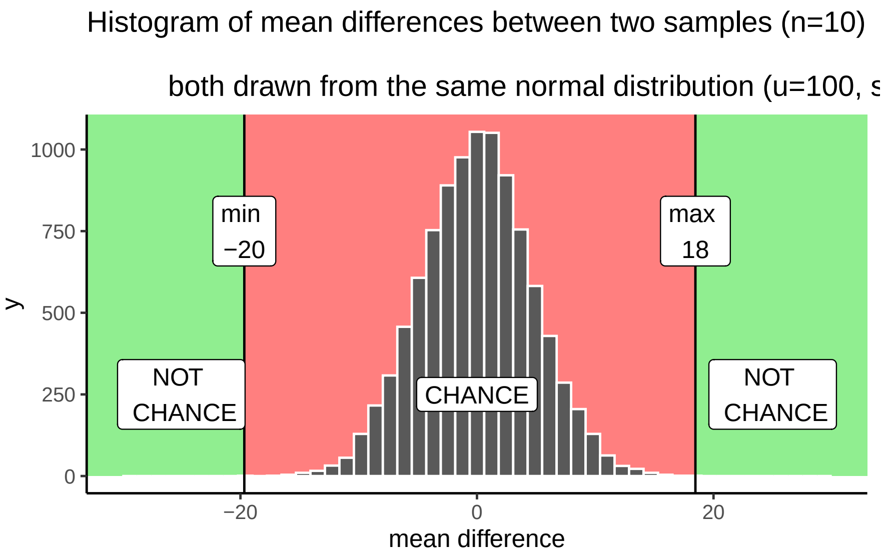


Figure 5.16: Applying decision boundaries to the histogrm of mean differences. The boundaries identify what differences chance did or did not produce in the simulation.

What about a difference of +10? That's in the red part, where chance lives. Chance could have done a difference of +10 because we can see that it did do that sometimes. The red part is the window of what chance did in our simulation. Anything inside the window could have been a difference caused by chance. If I found a difference of +10, I'd say, "ya, it coulda been chance." I would also be less confident that the difference was only caused by my experimental manipulation.

Statistical inference could be this easy. The number you get from your experiment could be in the chance window (then you can't rule out chance as a cause), or it could be outside the chance window (then you can rule out chance). Case closed. Let's all go home.

5.5.4.1 Grey areas

So what's the problem? Depending on who you are, and what kinds of risks you're willing to take, there might not be a problem. But, if you are just even a little bit risky then there is a problem that makes clear judgments about the role of chance difficult. We would like to say chance did or did not cause our difference. But, we're really always in the position of admitting that it could have sometimes, or wouldn't have most times. These are wishy washy statements, they are in between yes or no. That's OK. Grey is a color too, let's give grey some respect.

"What grey areas are you talking about?, I only see red or green, am I grey blind?". Let's look at where some grey areas might be. I say might be, because people disagree about where the grey is. People have different comfort levels with grey. Figure **??** shows my opinion on grey areas.

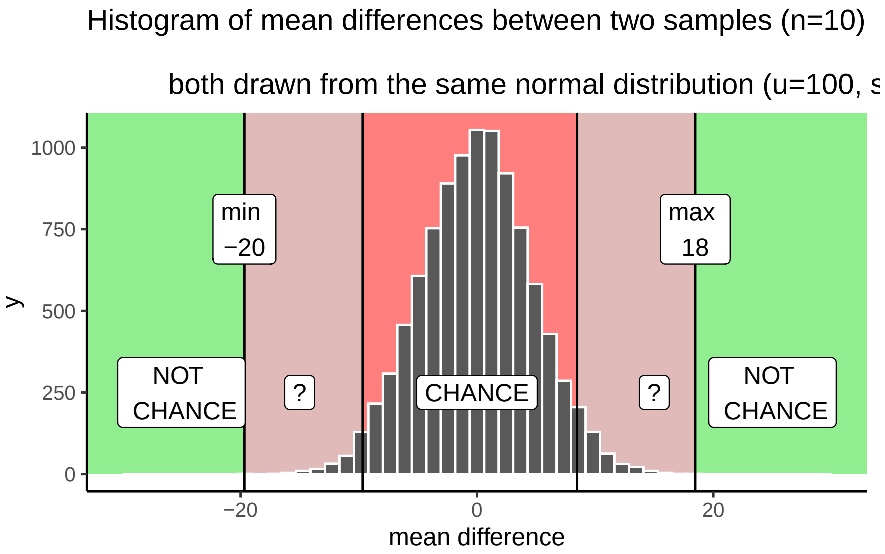


Figure 5.17: The question marks refer to an area where you have some uncertainty. Differences inside the question mark region do not happen very often by chance. When you find differences of these sizes, should you reject the idea that chance caused your difference? You will always have some uncertainty associated with this decision because it is clear that chance could have caused the difference. But, chance usually does not produce differences of these sizes.

I made two grey areas, and they are reddish grey, because we are still in the chance window. There are question marks (?) in the grey areas. Why? The question marks reflect some uncertainty that we have about those particular differences. For example, if you found a difference that was in a grey area, say a 15. 15 is less than the maximum, which means chance did create differences of around 15. But, differences of 15 don't happen very often.

What can you conclude or say about this 15 you found? Can you say without a doubt that chance did not produce the difference? Of course not, you know that chance could have. Still, it's one of those things that doesn't happen a lot. That makes chance an unlikely explanation. Instead of thinking that chance did it, you might be willing to take a risk and say that your experimental manipulation caused the difference. You'd be making a bet that it wasn't chance…but, could be a safe bet, since you know the odds are in your favor.

You might be thinking that your grey areas aren't the same as the ones I've drawn. Maybe you want to be more conservative, and make them smaller. Or, maybe you're more risky, and would make them bigger. Or, maybe you'd add some grey area going in a little bit to the green area (after all, chance could probably produce some bigger differences sometimes, and to avoid those you would have to make the grey area go a bit into the green area).

Another thing to think about is your decision policy. What will you do, when your observed difference is in your grey area? Will you always make the same decision about the role of chance? Or, will you sometimes flip-flop depending on how you feel. Perhaps, you think that there shouldn't be a strict policy, and that you should accept some level of uncertainty. The difference you found could be a real one, or it might not. There's uncertainty, hard to avoid that.

So let's illustrate one more kind of strategy for making decisions. We just talked about one that had some lines, and some regions. This makes it seem like a binary choice: we can either rule out, or not rule out the role of chance. Another perspective is that everything is a different shade of grey, like in Figure **??**.

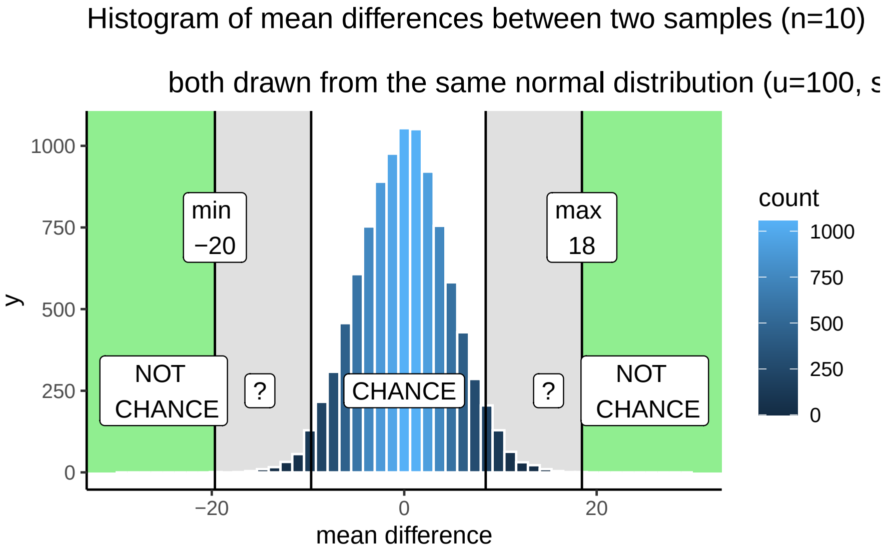


Figure 5.18: The shading of the blue bars indicates levels of confidence in whether a difference could have been produced by chance. Darker bars represent increased confidence that the difference was not produced by chance. Bars get darker as the mean difference increases in absolute value.

OK, so I made it shades of blue (because it was easier in R). Now we can see two decision plans at the same time. Notice that as the bars get shorter, they also get become a darker stronger blue. The color can be used as a guide for your confidence. That is, your confidence in the belief that your manipulation caused the difference rather than chance. If you found a difference near a really dark bar, those don't happen often by chance, so you might be really confident that chance didn't do it. If you find a difference near a slightly lighter blue bar, you might be slightly less confident. That is all. You run your experiment, you get your data, then you have some amount of confidence that it wasn't produced by chance. This way of thinking is elaborated to very interesting degrees in the Bayesian world of statistics. We don't wade too much into that, but mention it a little bit here and there. It's worth knowing it's out there.

5.5.4.2 Making decisions and being wrong

No matter how you plan to make decisions about your data, you will always be prone to making some mistakes. You might call one finding real, when in fact it was caused by chance. This is called a **type I** error, or a false positive. You might ignore one finding, calling it chance, when in fact it wasn't chance (even though it was in the window). This is called a **type II** error, or a false negative.

How you make decisions can influence how often you make errors over time. If you are a researcher, you will run lots of experiments, and you will make some amount of mistakes over time. If you do something like the very strict method of only accepting results as real when they are in the "no chance" zone, then you won't make many type I errors. Pretty much all of your result will be real. But, you'll also make type II errors, because you will miss things real things that your decision criteria says are due to chance. The opposite also holds. If you are willing to be more liberal, and accept results in the grey as real, then you will make more type I errors, but you won't make as many type II errors. Under the decision strategy of using these cutoff regions for decision-making there is a necessary trade-off. The Bayesian view get's around this a little bit. Bayesians talk about updating their beliefs and confidence over time. In that view, all you ever have is some level of confidence about whether something is real, and by running more experiments you can increase or decrease your level of confidence. This, in some fashion, avoids some trade-off between type I and type II errors.

Regardless, there is another way to reduce type I and type II errors, and to increase your confidence in your results, even before you do the experiment. It's called "knowing how to design a good experiment".

5.5.5 Part 4: Experiment Design

We've seen what chance can do. Now, let's venture into an experiment. We make a change between ecosystems A and B, gather the data, assess the average outcomes, and then observe the variance. Then we keep our fingers crossed, hoping that the variance is significant enough to be beyond natural fluctuations. Yes, nature keeps us guessing.

Here's the catch, we aren't always certain about the magnitude of our environmental interventions. So, even if an intervention induces a change, pinning down its exact magnitude can be challenging. And that's the essence of our experiment. Many interventions in Environmental Science might not cause large-scale shifts. This poses a challenge in identifying these subtle, yet potentially crucial, environmental effects. In a hypothetical scenario, introducing a certain pollinator species might influence plant growth, but to what extent? If the difference is marginal, differentiating between natural variation and the effect of our intervention becomes tricky. Let's say our intervention involves introducing shade in one ecosystem versus none in the other. While shade can influence plant growth, if the effect is only marginal, it becomes hard to ascertain if it wasn't just a natural occurrence. And, it's not straightforward to intensify the shading to amplify its impact, without risking other unintended consequences.

EXPERIMENT DESIGN TO THE RESCUE! Newsflash, it is often possible to change how you run your experiment so that it is **more sensitive** to smaller effects. How do you think we can do this? Here is a hint. It's the stuff you learned about the sampling distribution of the sample mean, and the role of sample-size. What happens to the sampling distribution of the sample mean when N (sample size)? The distribution gets narrower and narrower, and starts to look the a single number (the hypothetical mean of the hypothetical population). That's great. If you switch to thinking about mean difference scores, like the distribution we created in this test, what do you think will happen to that distribution as we increase N? It will will also shrink. As we increase N to infinity, it will shrink to 0. Which means that, when N is infinity, chance never produces any differences at all. We can use this.

For example, we could run our experiment with 20 subjects in each group. Or, we could decide to invest more time and run 40 subjects in each group, or 80, or 150. When you are the experimenter, you get to decide the design. These decisions matter big time. Basically, the more subjects you have, the more sensitive your experiment. With bigger N, you will be able to reliably detect smaller mean differences, and be able to confidently conclude that chance did not produce those small effects.

Check out the histograms in Figure **??**. This is the same simulation as before, but with four different sample-sizes: 20, 40, 80, 160. We are doubling our sample-size across each simulation just to see what happens to the width of the chance window.

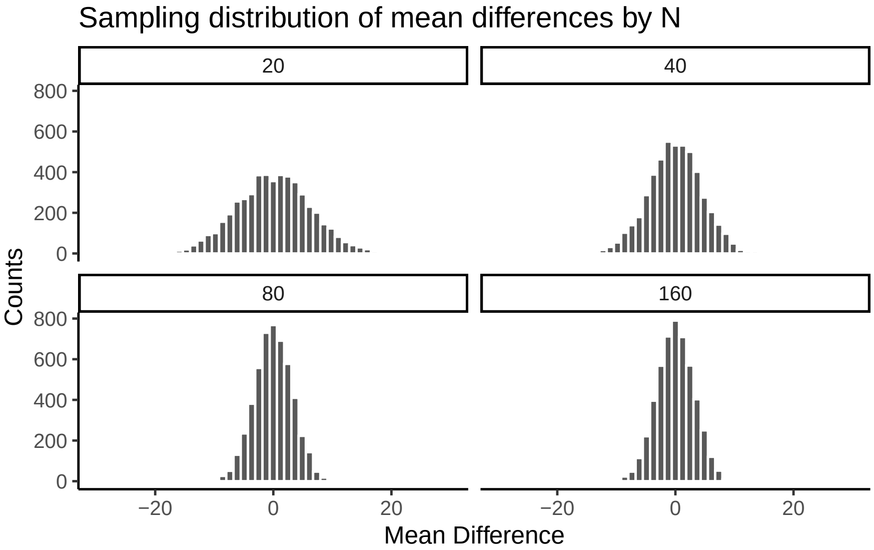


Figure 5.19: The range or width of the differences produced by chance shrinks as sample-size increases.

Table 5.2: The smallest and largest mean differences produced by chance as a function of sample-size.

sample_size	smallest	largest
20	-25.04651	24.51859
40	-16.74573	15.47603
80	-10.77141	10.97880
160	-10.90443	11.86731

There you have it. The **sampling distribution of the mean differences** shrinks toward 0 as sample-size increases. This means if you run an experiment with a larger sample-size, you will be able to detect smaller mean differences, and be confident they aren't due to chance. Table **??** contains the minimum and maximum values that chance produced across the four sample-sizes:

The table shows the range of chance behavior is very wider for smaller N and narrower for larger N. Consider what this narrowing means for your experiment design. For example, one aspect of the design is the choice of sample size, N, or in a psychology experiment the number of participants.

If it turns out your manipulation will cause a difference of +11, then what should you do? Run an experiment with N = 20 people? I hope not. If you did that, you could get a mean difference of +11 fairly often by chance. However, if you ran the experiment with 160 people, then you would definitely be able to say that +11 was not due to chance, it would be outside the range of what chance can do. You could even consider running the experiment with 80 subjects. A +11 there wouldn't happen often by chance, and you'd be cost-effective, spending less time on the experiment.

The point is: **the design of the experiment determines the sizes of the effects it can detect**. If you want to detect a small effect. Make your sample size bigger. It's really important to say this is not the only thing you can do. You can also make your cell-sizes bigger. For example, often times we take several measurements from a single subject. The more measurements you take (cell-size), the more stable your estimate of the subject's mean. We discuss these issues more later. You can also make a stronger manipulation, when possible.

5.5.6 Part 5: I have the power

> By the power of greyskull, I HAVE THE POWER - He-man

The last topic in this section is called **power**. Later we will define power in terms of some particular ideas about statistical inference. Here, we will just talk about the big idea. And, we'll show how to make sure your design has 100% power. Because, why not. Why run a design that doesn't have the power?

The big idea behind power is the concept of sensitivity. The concept of sensitivity assumes that there is something to be sensitive to. That is, there is some real difference that can be measured. So, the question is, how sensitive is your experiment? We've already seen that the number of subjects (sample-size), changes the sensitivity of the design. More subjects = more sensitivity to smaller effects.

Let's take a look at one more plot. What we will do is simulate a measure of sensitivity across a whole bunch of sample sizes, from 10 to 300. We'll do this in steps of 10. For each simulation, we'll compute the mean differences as we have done. But, rather than showing the histogram, we'll just compute the smallest value and the largest value. This is a pretty good measure of the outer reach of chance. Then we'll plot those values as a function of sample size and see what we've got.

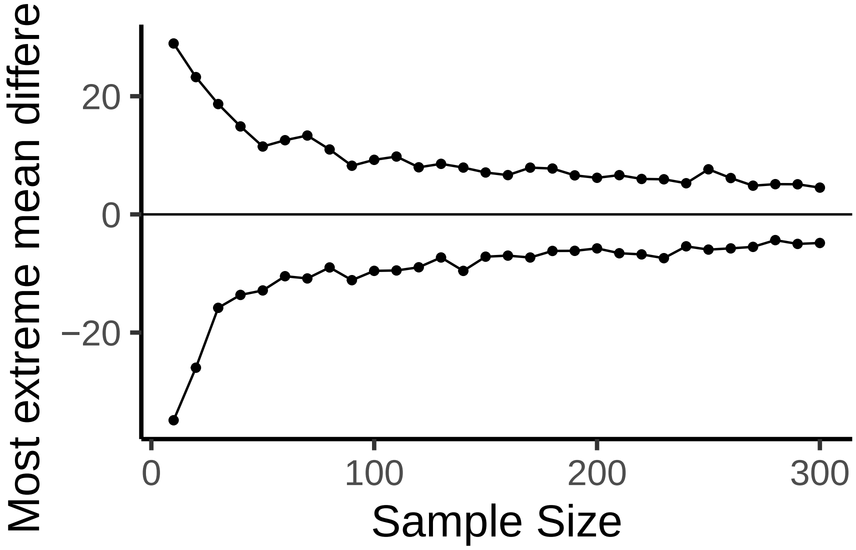


Figure 5.20: A graph of the maximum and minimum mean differences produced by chance as a function of sample-size. The range narrows as sample-size increases showing that chance alone produces a smaller range of mean differences as sample-size increases.

Figure **??** shows a reasonably precise window of sensitivity as a function of sample size. For each sample size, we can see the maximum difference that chance produced and the minimum difference. In those simulations, chance never produced bigger or smaller differences. So, each design is sensitive to any difference that is underneath the bottom line, or above the top line.

Here's another way of putting it. Which of the sample sizes will be sensitive to a difference of +10 or -10. That is, if a difference of +10 or -10 was observed, then we could very confidently say that the difference was not due to chance, because according to these simulations, chance never produced differences that big. To help us see which ones are sensitive, Figure **??** draws horizontal lines at -10 and +10.

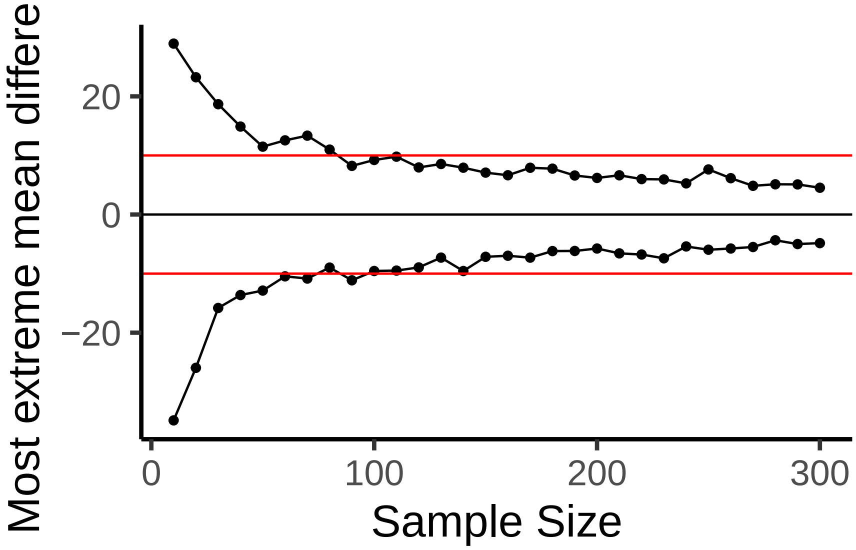


Figure 5.21: The red line represents the size of a mean difference that a researcher may be interested in detecting. All of the dots outside (above or below) the red line represent designs with small sample-sizes. When a difference of 10 occurs for these designs, we can rule out chance with confidence. The dots between the red lines represent designs with larger sample-sizes. These designs never produce differences as large as 10, so when those differences occur, we can be confident chance did not produce them.

Based on visual guesstimation, the designs with sample-size >= 100 are all sensitive to real differences of 10. Designs with sample-size > 100 all failed to produce extreme differences outside of the red lines by chance alone. If these designs were used, and if an effect of 10 or larger was observed, then we could be confident that chance alone did not produce the effect. Designing your experiment so that you know it is sensitive to the thing you are looking for is the big idea behind power.

5.5.7 Summary of Crump Test

What did we learn from this so-called fake Crump test that nobody uses? Well, we learned the basics of what we'll be doing moving forward. And, we did it all without any hard math or formulas. We sampled numbers, we computed means, we subtracted means between experimental conditions, then we repeated that process many times and counted up the mean differences and put them in a histogram. This showed us what chance do in an experiment. Then, we discussed how to make decisions around these facts. And, we showed how we can control the role of chance just by changing things like sample size.

5.6 The randomization test (permutation test)

Welcome to the first official inferential statistic in this textbook. Up till now we have been building some intuitions for you. Next, we will get slightly more formal and show you how we can use random chance to tell us whether our experimental finding was likely due to chance or not. We do this with something called a randomization test. The ideas behind the randomization test are the very same ideas behind the rest of the inferential statistics that we will talk about in later chapters. And, surprise, we have already talked about all of the major ideas already. Now, we will just put the ideas together, and give them the name **randomization test**.

Here's the big idea. When you run an experiment and collect some data you get to find out what happened that one time. But, because you ran the experiment only once, you don't get to find out what **could have happened**. The randomization test is a way of finding out what **could have happened**. And, once you know that, you can compare **what did happen** in your experiment, with **what could have happened**.

5.6.1 Pretend example does chewing gum improve your grades?

Let's say you run an experiment to find out if chewing gum causes students to get better grades on statistics exams. You randomly assign 20 students to the chewing gum condition, and 20 different students to the no-chewing gum condition. Then, you give everybody statistics tests and measure their grades. If chewing gum causes better grades, then the chewing gum group should have higher grades on average than the group who did not chew gum.

Let's say the data looked like this:

student	gum	no_gum
1	77	76
2	93	79
3	91	88
4	81	59
5	70	79
6	92	79
7	86	77
8	99	69
9	100	48
10	82	78
11	93	66
12	85	52
13	76	81
14	98	69
15	96	45
16	98	78
17	91	70
18	91	70
19	81	62
20	75	42
Sums	1755	1367
Means	87.75	68.35

So, did the students chewing gum do better than the students who didn't chew gum? Look at the mean test performance at the bottom of the table. The mean for students chewing gum was 87.75, and the mean for students who did not chew gum was 68.35. Just looking at the means, it looks like chewing gum worked!

"STOP THE PRESSES, this is silly". We already know this is silly because we are making pretend data. But, even if this was real data, you might think, "Chewing gum won't do anything, this difference could have been caused by chance, I mean, maybe the better students just happened to be put into the chewing group, so because of that their grades were higher, chewing gum didn't do anything…". We agree. But, let's take a closer look. We already know how the data come out. What we want to know is how they could have come out, what are all the possibilities?

For example, the data would have come out a bit different if we happened to have put some of the students from the gum group into the no gum group, and vice versa. Think of all the ways you could have assigned the 40 students into two groups, there are lots of ways. And, the means for each group would turn out differently depending on how the students are assigned to each group.

Practically speaking, it's not possible to run the experiment every possible way, that would take too long. But, we can nevertheless estimate how all of those experiments might have turned out using simulation.

Here's the idea. We will take the 40 measurements (exam scores) that we found for all the students. Then we will randomly take 20 of them and pretend they were in the gum group, and we'll take the remaining 20 and pretend they were in the no gum group. Then we can compute the means again to find out what would have happened. We can keep doing this over and over again. Every time computing what happened in that version of the experiment.

5.6.1.1 Doing the randomization

Before we do that, let's show how the randomization part works. We'll use fewer numbers to make the process easier to look at. Here are the first 5 exam scores for students in both groups.

student	gum	no_gum
1	77	76
2	93	79
3	91	88
4	81	59
5	70	79
Sums	412	381
Means	82.4	76.2

Things could have turned out differently if some of the subjects in the gum group were switched with the subjects in the no gum group. Here's how we can do some random switching. We will do this using R.

```
all_scores       <- c(gum[1:5],no_gum[1:5])
randomize_scores <- sample(all_scores)
new_gum          <- randomize_scores[1:5]
new_no_gum       <- randomize_scores[6:10]
print(new_gum)
#> [1] 59 79 79 70 88
print(new_no_gum)
#> [1] 91 81 77 93 76
```

We have taken the first 5 numbers from the original data, and put them all into a variable called `all_scores`. Then we use the `sample` function in R to shuffle the scores. Finally, we take the first 5 scores from the shuffled numbers and put them into a new variable called `new_gum`. Then, we put the last five scores into the variable `new_no_gum`. Then we printed them, so we can see them.

If we do this a couple of times and put them in a table, we can indeed see that the means for gum and no gum would be different if the subjects were shuffled around. Check it out:

student	gum	no_gum	gum2	no_gum2	gum3	no_gum3
1	77	76	79	81	93	76
2	93	79	88	59	79	81
3	91	88	79	91	79	91
4	81	59	93	76	77	59
5	70	79	77	70	70	88
Sums	412	381	416	377	398	395
Means	82.4	76.2	83.2	75.4	79.6	79

5.6.1.2 Simulating the mean differences across the different randomizations

In our pretend experiment we found that the mean for students chewing gum was 87.75, and the mean for students who did not chew gum was 68.35. The mean difference (gum - no gum) was 19.4. This is a pretty big difference. This is **what did happen**. But, **what could have happened**? If we tried out all of the experiments where different subjects were switched around, what does the distribution of the possible mean differences look like? Let's find out. This is what the randomization test is all about.

When we do our randomization test we will measure the mean difference in exam scores between the gum group and the no gum group. Every time we randomize we will save the mean difference.

Let's look at a short animation of what is happening in the randomization test. **?@fig-5randtest** shows data from a different fake experiment, but the principles are the same. We'll return to the gum no gum experiment after the animation. The animation is showing three important things. First, the purple dots show the mean scores in two groups (didn't study vs study). It looks like there is a difference, as 1 dot is lower than the other. We want to know if chance could produce a difference this big. At the beginning of the animation, the light green and red dots show the individual scores from each of 10 subjects in the design (the purple dots are the means of these original scores). Now, during the randomizations, we randomly shuffle the original scores between the groups. You can see this happening throughout the animation, as the green and red dots appear in different random combinations. The moving yellow dots show you the new means for each group after the randomization. The differences between the yellow dots show you the range of differences that chance could produce.

We are engaging in some visual statistical inference. By looking at the range of motion of the yellow dots, we are watching what kind of differences chance can produce. In this animation, the purple dots, representing the original difference, are generally outside of the range of chance. The yellow dots don't move past the purple dots, as a result chance is an unlikely explanation of the difference.

If the purple dots were inside the range of the yellow dots, then when would know that chance is capable of producing the difference we observed, and that it does so fairly often. As a result, we should not conclude the manipulation caused the difference, because it could have easily occurred by chance.

Let's return to the gum example. After we randomize our scores many times, and computed the new means, and the mean differences, we will have loads of mean differences to look at, which we can plot in a histogram. The histogram gives a picture of **what could have happened**. Then, we can compare **what did happen** with **what could have happened**.

Here's the histogram of the mean differences from the randomization test. For this simulation, we randomized the results from the original experiment 1000 times. This is what could have happened. The blue line in Figure **??** shows where the observed difference lies on the x-axis.

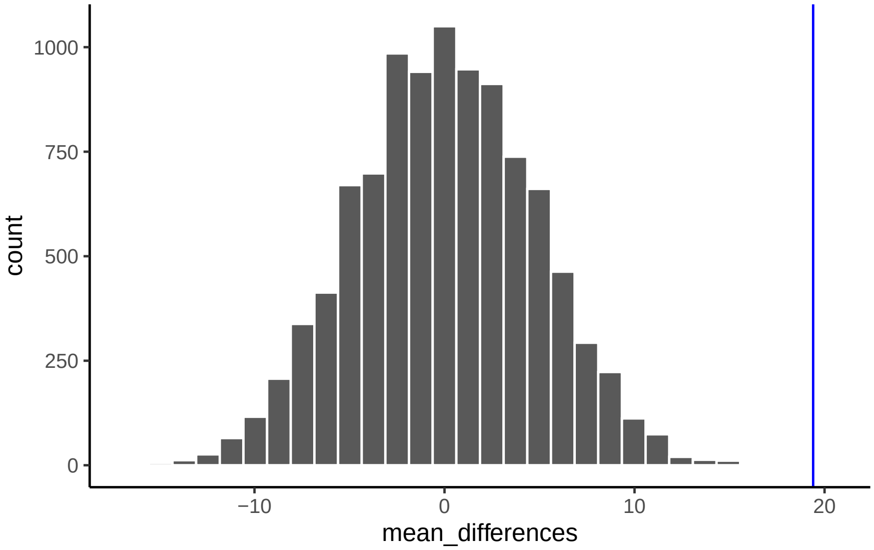


Figure 5.22: A histogram of simulated mean differences for a randomization test

What do you think? Could the difference represented by the blue line have been caused by chance? My answer is probably not. The histogram shows us the window of chance. The blue line is not inside the window. This means we can be pretty confident that the difference we observed was not due to chance.

We are looking at another window of chance. We are seeing a histogram of the kinds of mean differences that could have occurred in our experiment, if we had assigned our subjects to the gum and no gum groups differently. As you can see, the mean differences range from negative to positive. The most frequent difference is 0. Also, the distribution appears to be symmetrical about zero, which shows we had roughly same the chances of getting a positive or negative difference. Also, notice that as the differences get larger (in the positive or negative direction, they become less frequent). The blue line shows us **the observed difference**, this is the one we found in our fake experiment. Where is it? It's way out to the right. It is is well outside the histogram. In other words, when we look at **what could have happened**, we see that **what did happen** doesn't occur very often.

IMPORTANT: In this case, when we speak of **what could have happened**. We are talking about what could have happened **by chance**. When we compare what did happen to what chance could have done, we can get a better idea of whether our result was caused by chance.

OK, let's pretend we got a much smaller mean difference when we first ran the experiment. We can draw new lines (blue and red) to represent a smaller mean that we might have found.

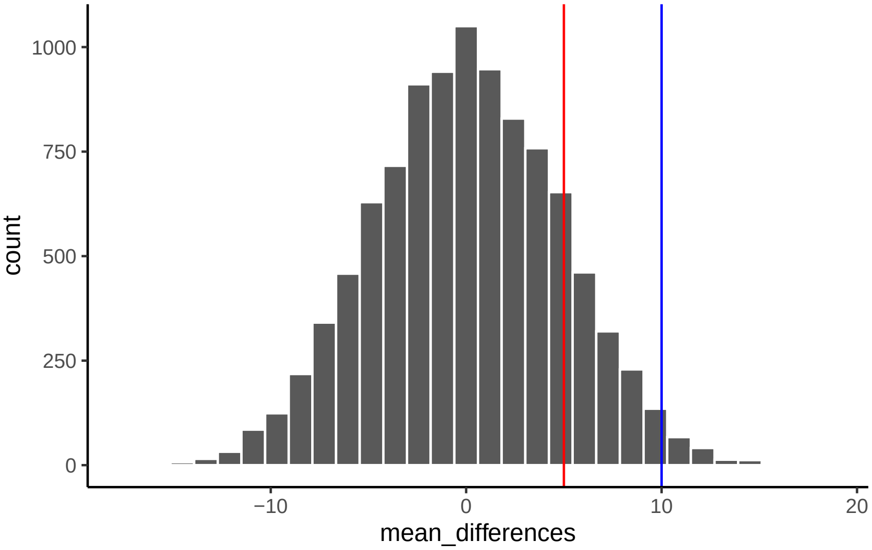


Figure 5.23: Would you expect a mean difference represented by the blue line to occur more or less often by chance compared to the mean difference represented by the red line?

Look at the blue line in Figure **??**. If you found a mean difference of 10, would you be convinced that your difference was not caused by chance? As you can see, the blue line is inside the chance window. Notably, differences of +10 don't very often. You might infer that your difference was not likely to be due to chance (but you might be a little bit skeptical, because it could have been). How about the red line? The red line represents a difference of

+5. If you found a difference of +5 here, would you be confident that your difference was not caused by chance? I wouldn't be. The red line is totally inside the chance window, this kind of difference happens fairly often. I'd need some more evidence to consider the claim the some independent variable actually caused the difference. I'd be much more comfortable assuming that sampling error probably caused the difference.

5.6.2 Take homes so far

Have you noticed that we haven't used any formulas yet, but we have been able to accomplish inferential statistics. We will see some formulas as we progress, but these aren't as the idea behind the formulas.

Inferential statistics is an attempt to solve the problem: **where did my data from?**. In the randomization test example, our question was: **where did the differences between the means in my data come from?**. We know that the differences could be produced by chance alone. We simulated what chance can due using randomization. Then we plotted what chance can do using a histogram. Then, we used to picture to help us make an inference. Did our observed difference come from the distribution, or not? When the observed difference is clearly inside the chance distribution, then we can infer that our difference **could have been produced by chance**. When the observed difference is not clearly inside the chance distribution, then we can infer that our difference was **probably not produced by chance**.

In my opinion, these pictures are very, very helpful. If one of our goals is to help ourselves summarize a bunch of complicated numbers to arrive at an inference, then the pictures do a great job. We don't even need a summary number, we just need to look at the picture and see if the observed difference is inside or outside of the window. This is what it is all about. Creating intuitive and meaningful ways to make inferences from our data. As we move forward, the main thing that we will do is formalize our process, and talk more about "standard" inferential statistics. For example, rather than looking at a picture (which is a good thing to do), we will create some helpful numbers. For example, what if you wanted to the probability that your difference could have been produced by chance? That could be a single number, like 95%. If there was a 95% probability that chance can produce the difference you observed, you might not be very confident that something like your experimental manipulation was causing the difference. If there was only 1% probability that chance could produce your difference, then you might be more confident that **chance did not** produce the difference; and, you might instead be comfortable with the possibility that your experimental manipulation actually caused the difference. So, how can we arrive at those numbers? In order to get there we will introduce you to some more foundational tools for statistical inference.

5.7 Videos

5.7.1 Null and Alternate Hypotheses

5.7.2 Types of Errors

6 Hypothesis Testing

6.1 Hypothesis Testing - The Nuts & Bolts

Hypothesis testing helps us figure out if what we believe about a whole group is likely true, just by looking at a small part of it (a sample).

6.1.1 Clarifying Alpha, P-value, and Confidence Level

Before diving deep, let's clear up some terms you'll come across often.

Alpha (α)

Alpha (α) is the significance level of a statistical test, and it quantifies the risk of committing a Type I error. A Type I error happens when we incorrectly reject a true null hypothesis. The standard value for alpha is often set at 0.05, implying a 5% chance of making a Type I error. In other words, we are willing to accept a 5% risk of concluding that a difference exists when there is no actual difference.

P-value

The p-value is another crucial concept in hypothesis testing. It represents the probability of observing the obtained results, or something more extreme, assuming that the null hypothesis is true. A small p-value (usually 0.05) suggests that the observed data is inconsistent with the null hypothesis, and thus, you have evidence to reject it.

Confidence Level

The confidence level is related but distinct from alpha and p-value. While alpha quantifies the risk of a Type I error, the confidence level indicates how confident we are in our statistical estimates. The confidence level is calculated as the complement of alpha:

$$\text{Confidence Level} = 1 - \alpha$$

For example, if α is 0.05, the confidence level would be (1 - 0.05 = 0.95) or 95%. This means we are 95% confident that our results fall within a specific range.

Bringing It All Together

- **Alpha (α)**: Risk of Type I error (usually 5%)
- **P-value**: Probability of observed data given the null is true
- **Confidence Level**: Confidence in the range of our estimates (usually 95%)

Grasping how these three terms connect and differ is key to making sense of the stats we'll discuss.

6.1.2 The Steps of Hypothesis Testing Applied to an Example

Let's say we want to know if the average pollution in a set of water samples is above the legal limit. Or if young deer in a region are, on average, healthy.

Step 1: Define Your Hypotheses: First, we need to define two hypotheses: the **research hypothesis** and the **null hypothesis**.

- **Research Hypothesis ($\mathbf{H_a}$)**: This is what we aim to support. **Keep in mind, we can't exactly "prove" $\mathbf{H_a}$ is correct, we can only say that $\mathbf{H_0}$ isn't likely**. It can take a few forms based on the question:
  - H_a: average pollution > legal limit (pollution is too high)
  - H_a: average pollution < legal limit (pollution is too low)
  - H_a: average pollution legal limit (pollution is just different)
- **Null Hypothesis ($\mathbf{H_0}$)**: This is the default or 'no change' scenario. It's opposite to the research hypothesis.
  - H_0: average pollution legal limit (for the first H_a)
  - H_0: average pollution legal limit (for the second H_a)
  - H_0: average pollution = legal limit (for the third H_a)

Step 2: Choose Your Test Statistic: Based on the data, we'll compute a **test statistic**. This number will help us decide which hypothesis seems more likely.

Step 3: Determine the Rejection Region: Before running the test, we decide on a **rejection region**. If our test statistic falls in this region, we'll reject the null hypothesis.

Step 4: Check Assumptions: Before drawing conclusions, ensure that the test's conditions and assumptions are satisfied.

Step 5: Draw Conclusions: Finally, based on the test statistic and the rejection region, decide whether to reject the null hypothesis.

6.1.3 Errors in Hypothesis Testing

Sometimes, even with the best methods, we make incorrect decisions.

- **Type I Error** (α): This happens when we mistakenly reject the true null hypothesis. Imagine sending an innocent person to jail. Typically, α is set at 0.05 (5%).
- **Type II Error** (β): Here, we mistakenly accept a false null hypothesis. Think of it as letting a guilty person go free.

Decision	If the null hypothesis is True	If the null hypothesis is False
Reject $\mathbf{H_0}$	Type I error (prob = α)	Correct (prob = 1 - β)
Fail to reject $\mathbf{H_0}$	Correct (prob = 1 - α)	Type II error (prob = β)

Key Takeaway: As α gets smaller, β gets bigger, and vice-versa.

6.1.4 Deciphering Significance with P-values

The p-value is like a reality-check. It tells us how weird our results are if we assume the starting belief (null hypothesis) is spot on.

- **One-Tailed Test**: The p-value shows the likelihood of observing an average as extreme as our sample's if the null hypothesis stands.
- **Two-Tailed Test**: This p-value represents the odds of spotting an average as different from the null value as our sample's.

Rule of Thumb: If the p-value is less than α, we opt to reject the null hypothesis.

6.2 Graphical Review

6.2.1 Key Players in Hypothesis Testing Visualization

We define and visualize the core components essential to understanding the graphical representations of hypothesis testing: