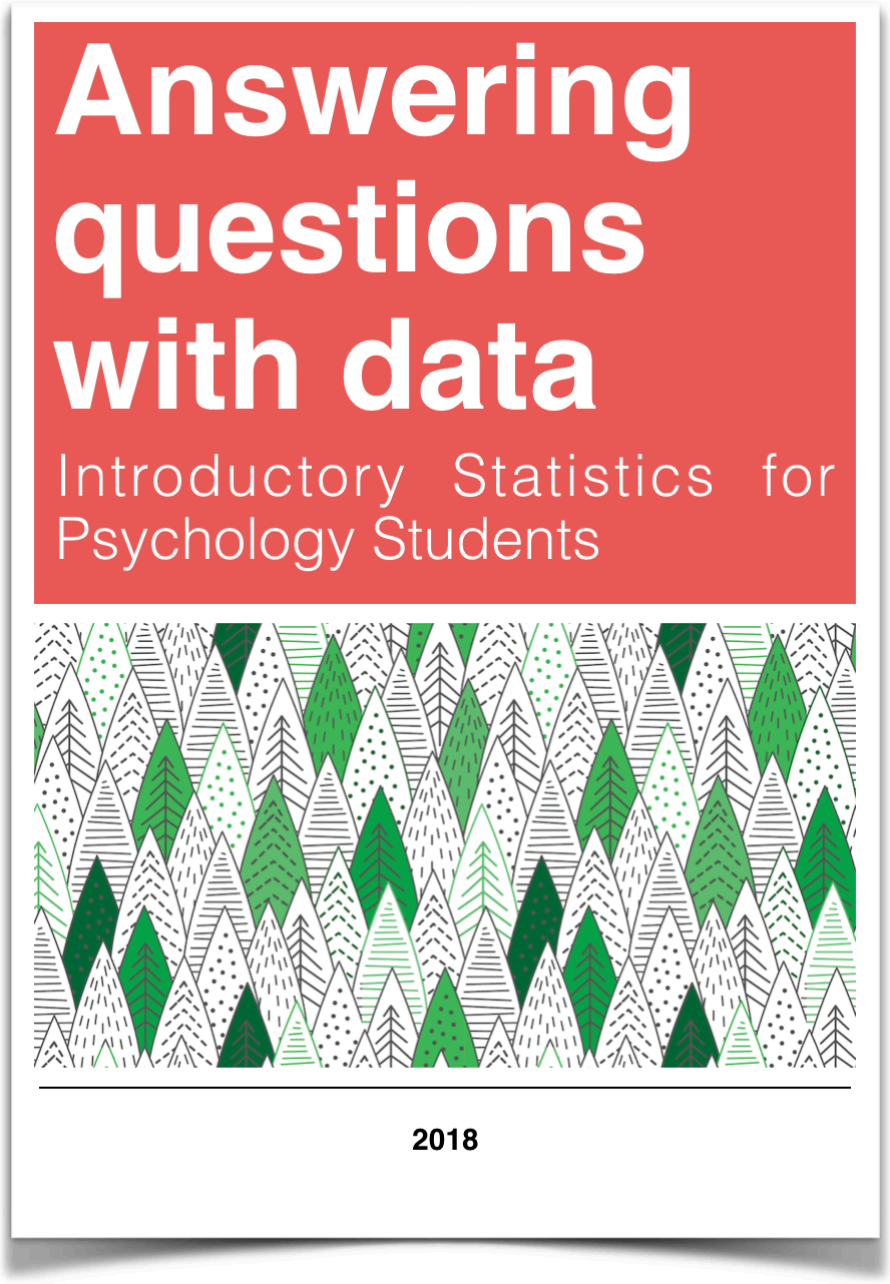
Answering questions with data

Lead Author: Matthew J. C. Crump
Chapters 2 and 4 adapted from Navarro, D.
Videos: Jeffrey Suzuki

2018 Last Compiled 2020-07-24

Contents

Preface



First Draft (version 0.9 = August 7th, 2018) Last Compiled: 2020-07-24

Online Textbook: https://crumplab.github.io/statistics/

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

0.1 Important notes

This is a free textbook teaching introductory statistics for undergraduates in Psychology. This textbook is part of a larger OER course package for teaching undergraduate statistics in Psychology, including this textbook, a lab manual, and a course website. All of the materials are free and copiable, with source code maintained in Github repositories. The links below connect to various components of the project. The main OSF project link is https://osf.io/3s68c/.

0.1.1 Textbook

- website: https://crumplab.github.io/statistics/
- OSF: https://osf.io/jze52/
- Github: https://github.com/CrumpLab/statistics
- DOI: 10.17605/OSF.IO/JZE52

0.1.2 Lab Manual

- website: https://crumplab.github.io/statisticsLab/
- OSF: https://osf.io/m2npj/
- Github: https://github.com/CrumpLab/statisticsLab
- DOI: 10.17605/OSF.IO/M2NPJ

0.1.3 Course website

- website: https://crumplab.github.io/psyc3400/
- OSF: https://osf.io/awxu9/
- Github: https://github.com/CrumpLab/psyc3400
- DOI: 10.17605/OSF.IO/AWXU9

0.1.4 Contributors

Team members contributing new content include, Matthew Crump, Alla Chavarga, Anjali Krishnan, Jeffrey Suzuki, and Stephen Volz. This textbook was primarily written by Matthew J. C. Crump. Jeff contributed the YouTube videos peppered throughout the textbook. All of Jeff's statistics videos are available on his Youtube channel: Statistics Video playlist.

Alla, Anjali, and Stephen wrote the lab manual exercises for SPSS, JAMOVI, and Excel. Matt Crump wrote the lab manual exercises for R.

Matt Crump wrote a free and copiable course website, in R Markdown. The course website also contains slide decks for the lectures.

0.1.5 Attributions

Two of the chapters were adapted from Danielle Navarro's wonderful (and bigger) free textbook, also licensed under the same creative commons license. The citation for that textbook is: Navarro, D. (2018). Learning statistics with R: A tutorial for psychology students and other beginners (version 0.6). The website is https://compcogscisydney.org/learning-statistics-with-r/

Chapter notes within the book are provided to indicate sections where material from Navarro was included. A short summary is here

Chapter 1: Why statistics, Adapted nearly verbatim with some editorial changes from Chapters 1 and 2, Navarro, D.

Chapter 4: Probability, Sampling, and Estimation, Adapted and expanded from Chapters 9 and 10, Navarro D.

0.1.6 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.

0.2 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 in the GitHub repository for this book:

https://github.com/CrumpLab/statistics

In principle, anybody could fork or otherwise download this repository. 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.

0.2.1 Acknowledgments

Thanks to the librarians at Brooklyn College of CUNY, especially Miriam Deutch, and Emily Fairey, for their support throughout the process. Thanks to CUNY for supporting OER development, and for the grant we received to develop this work. Thanks to Jenn Richler for letting me talk about statistics all summer long.

0.2.2 Why we did this

Why write another statistics textbook, aren't there already loads of those? Yes, there are. We had a couple reasons. First, we would like to make R more accessible for the undergraduate population, and we wrote this textbook around the capabilities of R. The textbook was written entirely in R-Studio, and most of the examples have associated R-code. R is not much of a focus in the textbook, but there is an introduction to using R to solve data-analysis problems in the lab manual. Many instructors still use SPSS, Excel, or newer free GUIs like JAMOVI, so we also made lab exercises for each of those as well.

This is a mildly opinionated, non-traditional introduction to statistics. It acknowledges some of the major ideas from traditional frequentist approaches, and some Bayesian approaches. Much of the conceptual foundation is rooted in simulations that can be conducted in R. We us some formulas, but mostly explain things without formulas. The textbook was written with math-phobia in mind, and attempts to reduce the phobia associated with arithmetic computations. There are many things missing that should probably be added. We will do our best to add necessary things as we update the textbook.

0.2.3 Hypothes.is

Hypothesis is a web-browser plug-in that lets you make comments on websites by highlighting text, and then making comments. Feel free to use hypothesis with this textbook. We will read your comments.

1. Use Hypothes.is, an amazing tool for annotating the web.

a. Go to Hypothes.is, and "get-started"

b. Install the the add-on for chrome, or other browser

c. That's it, turn on Hypothes.is when you are reading this textbook, and you will see all public annotations made by anyone else.

Chapter 1

Why Statistics?

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 psychology of statistics

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

A big part of this issue at hand relates to the very idea of statistics. What is it? What's it there for? And why are scientists so bloody obsessed with it? These are all good questions, when you think about it. So let's start with the last one. As a group, scientists seem to be bizarrely fixated on running statistical tests on everything. In fact, we use statistics so often that we sometimes forget to explain to people why we do. It's a kind of article of faith among scientists – and especially social scientists – that your findings can't be trusted until you've done some stats. Undergraduate students might be forgiven for thinking that we're all completely mad, because no-one takes the time to answer one very simple question:

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

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

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

1.1.1 The curse of belief bias

People are mostly pretty smart. We're 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:

	conlusion 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, **?** 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:

	conlusion 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:

	conlusion 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 judgements were correct (you'd expect 50% by chance). Even so, if you were a professional "evaluator of evidence", and someone came along and offered you a magic tool that improves your chances of making the right decision from 60% to (say) 95%, you'd probably jump at it, right? Of course you would. Thankfully, we actually do have a tool that can do this. But it's not magic, it's statistics. So that's reason #1 why scientists love statistics. It's just *too easy* for us to "believe what we want to believe"; so 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 post-graduate courses. Specifically, the thing that caused the problem was that the gender breakdown of their admissions, which looked like this:

	Number of applicants	Percent admitted
Males	8442	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 favour of women (sort of!), you'd probably think that I was either crazy or sexist.

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. A big thank you to Wilfried Van Hirtum for pointing this out to me!

When people started looking more carefully at the admissions data (**?**) 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 all 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, counterintuitive traps lying in wait for the unwary. That's reason #2 why scientists love statistics, and why we teach research methods. Because science is hard, and the truth is sometimes cunningly hidden in the nooks and crannies of complicated data.

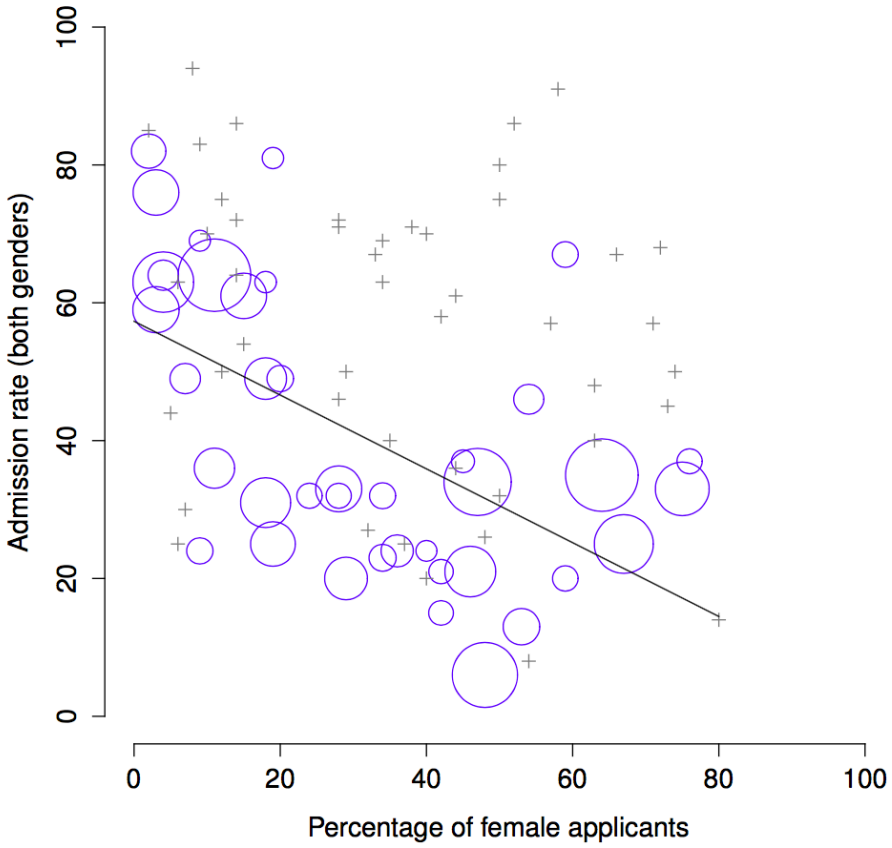


Figure 1.1: The Berkeley 1973 college admissions data. This figure plots the admission rate for the 85 departments that had at least one female applicant, as a function of the percentage of applicants that were female. The plot is a redrawing of Figure 1 from Bickel et al. (1975). Circles plot departments 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.

Before leaving this topic entirely, I want to point out something else really critical that is often overlooked in a research methods class. Statistics only solves

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

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

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

- **Why does psychology have so much statistics?**

To be perfectly honest, there's a few different reasons, some of which are better than others. The most important reason is that psychology is a statistical science. What I mean by that is that the "things" that we study are *people*. Real, complicated, gloriously messy, infuriatingly perverse people. The "things" of physics include objects like electrons, and while there are all sorts of complexities that arise in physics, electrons don't have minds of their own. They don't have opinions, they don't differ from each other in weird and arbitrary ways, they don't get bored in the middle of an experiment, and they don't get angry at the experimenter and then deliberately try to sabotage the data set. At a fundamental level psychology is harder than physics.

Basically, we teach statistics to you as psychologists because you need to be better at stats than physicists. There's actually a saying used sometimes in physics, to the effect that "if your experiment needs statistics, you should have done a better experiment". They have the luxury of being able to say that because their objects of study are pathetically simple in comparison to the vast mess that confronts social scientists. It's not just psychology, really: most social sciences are desperately reliant on statistics. Not because we're bad experimenters, but because we've picked a harder problem to solve. We teach you stats because you really, really need it.

- **Can't someone else do the statistics?**

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

1. There's the fundamental reason: statistics is deeply intertwined with research design. If you want to be good at designing psychological studies, you need to at least understand the basics of stats.

2. If you want to be good at the psychological side of the research, then you need to be able to understand the psychological literature, right? But almost every paper in the psychological literature reports the results of statistical analyses. So if you really want to understand the psychology, you need to be able to understand what other people did with their data. And that means understanding a certain amount of statistics.

3. There's a big practical problem with being dependent on other people to do all your statistics: statistical analysis is *expensive*. In almost any real life situation where you want to do psychological research, the cruel facts will be that you don't have enough money to afford a statistician. So the economics of the situation mean that you have to be pretty self-sufficient.

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

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

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

1.4 Statistics in everyday life

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

When I started writing up my lecture notes I took the 20 most recent news articles posted to the ABC news website. Of those 20 articles, it turned out that 8 of them involved a discussion of something that I would call a statistical topic; 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 **?** for the discussion of study design, and **?** for the discussion of scales of measurement. Later versions will attempt to be more precise in the citations.

1.7 Introduction to psychological measurement

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

1.7.1 Some thoughts about psychological measurement

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

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

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 very likely to be *male* or *female*, but it doesn't have to agree with my chromosomal gender. I may also choose to identify with *neither*, or to explicitly call myself *transgender*.

As you can see, for some things (like age) it seems fairly obvious what the set of possible measurements should be, whereas for other things it gets a bit tricky. But I want to point out that even in the case of someone's age, it's much more subtle than this. For instance, in the example above, I assumed that it was okay to measure age in years. But if you're a developmental psychologist, that's way too crude, and so you often measure age in *years and months* (if a child is 2 years and 11 months, this is usually written as "2;11"). If you're interested in newborns, you might want to measure age in *days since birth*, maybe even *hours since birth*. In other words, the way in which you specify the allowable measurement values is important.

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

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

- You could just ask people "how old are you?" The method of self-report is fast, cheap and easy, but it only works with people old enough to understand the question, and some people lie about their age.
- You could ask an authority (e.g., a parent) "how old is your child?" This method is fast, and when dealing with kids it's not all that hard since the parent is almost always around. It doesn't work as well if you want to know "age since conception", since a lot of parents can't say for sure when

conception took place. For that, you might need a different authority (e.g., an obstetrician).

- You could look up official records, 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 behavioural 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 colour". 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 colour". Similarly, gender is nominal too: male isn't better or worse than female, neither does it make sense to try to talk about an "average gender". In short, nominal scale variables are those for which the only thing you can say about the different possibilities is that they are different. That's it.

Let's take a slightly closer look at this. Suppose I was doing research on how people commute to and from work. One variable I would have to measure would be what kind of transportation people use to get to work. This "transport type" variable could have quite a few possible values, including: "train", "bus", "car",

"bicycle", 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
(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 analysing 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.

	continuous	discrete

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 potatos 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, *Computational Cognitive Science*, has an assessment structure that has a research component and an exam component (plus other things). The exam component is *intended* to measure something different from the research component, so the assessment as a whole has low internal consistency. However, within the exam there are several questions that are intended to (approximately) measure the same things, and those tend to produce similar outcomes; so the exam on its own has a fairly high internal consistency. Which is as it should be. You should only demand reliability in those situations where you want to be 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 analysing 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 analysing 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 is a broad term that covers "any study in which the researcher doesn't have quite as much control as they do in an experiment". Obviously, control is something that scientists like to have, but as the previous example illustrates, there are lots of situations in which you can't or shouldn't try to obtain that control. Since it's grossly unethical (and almost certainly criminal) to force people to smoke in order to find out if they get cancer, this is a good example of a situation in which you really shouldn't try to obtain experimental control. But there are other reasons too. Even leaving aside the ethical issues, our "smoking experiment" does have a few other issues. For instance, when I suggested that we "force" half of the people to become smokers, I must have been talking about *starting* with a sample of non-smokers, and then forcing them to become smokers. While this sounds like the kind of solid, evil experimental design that a mad scientist would love, it might not be a very sound way of investigating the effect in the real world. For instance, suppose that smoking only causes lung cancer when people have poor diets, and suppose also that people who normally smoke do tend to have poor diets. However, since the "smokers" in our experiment aren't "natural" smokers (i.e., we forced non-smokers to become smokers; they didn't take on all of the other normal, real life characteristics that smokers might tend to possess) they probably have better diets. As such, in this silly example they wouldn't get lung cancer, and our experiment will fail, because it violates the structure of the "natural" world (the technical name for this is an "artifactual" result; see later).

One distinction worth making between two types of non-experimental research is the difference between **quasi-experimental research** and **case studies**. The example I discussed earlier – in which we wanted to examine incidence of lung cancer among smokers and non-smokers, without trying to control who smokes and who doesn't – is a quasi-experimental design. That is, it's the same as an

experiment, but we don't control the predictors (IVs). We can still use statistics to analyse the results, it's just that we have to be a lot more careful.

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

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 theatre so that I can count them. When I do this with a class of 300 students, 0 people claim to be cheaters. So I therefore conclude that the proportion of cheaters in my class is 0%. Clearly this is a bit ridiculous. But the point here is not that this is a very deep methodological example, but rather to explain what construct validity is. The problem with my measure is that while I'm *trying* to measure "the proportion of people who cheat" what I'm actually measuring is "the proportion of people stupid enough to own up to cheating, or bloody minded enough to pretend that they do". Obviously, these aren't the

same thing! So my study has gone wrong, because my measurement has very poor construct validity.

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 rigour to it. To be ecologically valid, the study has to look right in a fairly specific way. The idea behind it is the intuition that a study that is ecologically valid is more likely to be externally valid. It's no guarantee, of course. But the nice thing about ecological validity is that it's much easier to check whether a study is ecologically valid than it is to check whether a study is externally valid. 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 behaviour 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

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

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. In the example I gave above, the differential attrition would be homogeneous if (and only if) the easily bored participants are dropping out of all of the conditions in my experiment at about the same rate. In general, the main effect of homogeneous attrition is likely to be that it makes your sample unrepresentative. As such, the biggest worry that you'll have is that the generalisability of the results decreases: in other words, you lose external validity.

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

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 intro-

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

Regression to the mean is surprisingly common. For instance, if two very tall people have kids, their children will tend to be taller than average, but not as

tall as the parents. The reverse happens with very short parents: two very short parents will tend to have short children, but nevertheless those kids will tend to be taller than the parents. It can also be extremely subtle. For instance, there have been studies done that suggested that people learn better from negative feedback than from positive feedback. However, the way that people tried to show this was to give people positive reinforcement whenever they did good, and negative reinforcement when they did bad. And what you see is that after the positive reinforcement, people tended to do worse; but after the negative reinforcement they tended to do better. But! Notice that there's a selection bias here: when people do very well, you're selecting for "high" values, and so you should *expect* (because of regression to the mean) that performance on the next trial should be worse, regardless of whether reinforcement is given. Similarly, after a bad trial, people will tend to improve all on their own. The apparent superiority of negative feedback is an artifact caused by regression to the mean (**?**)

1.13.8 Experimenter bias

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

The general solution to the problem of experimenter bias is to engage in double blind studies, where neither the experimenter nor the participant knows which condition the participant is in, or knows what the desired behaviour is. This provides a very good solution to the problem, but it's important to recognize that it's not quite ideal, and hard to pull off perfectly. For instance, the obvious way that I could try to construct a double blind study is to have one of my Ph.D. students (one who doesn't know anything about the experiment) run the study. That feels like it should be enough. The only person (me) who knows all the details (e.g., correct answers to the questions, assignments of participants to conditions) has no interaction with the participants, and the person who does all the talking to people (the Ph.D. student) doesn't know anything. Except,

that last part is very unlikely to be true. In order for the Ph.D. student to run the study effectively, they need to have been briefed by me, the researcher. And, as it happens, the Ph.D. student also knows me, and knows a bit about my general beliefs about people and psychology (e.g., I tend to think humans are much smarter than psychologists give them credit for). As a result of all this, it's almost impossible for the experimenter to avoid knowing a little bit about what expectations I have. And even a little bit of knowledge can have an effect: suppose the experimenter accidentally conveys the fact that the participants are expected to do well in this task. Well, there's a thing called the "Pygmalion effect": if you expect great things of people, they'll rise to the occasion; but if you expect them to fail, they'll do that too. In other words, the expectations become a self-fulfilling prophesy.

1.13.9 Demand effects and reactivity

When talking about experimenter bias, the worry is that the experimenter's knowledge or desires for the experiment are communicated to the participants, and that these effect people's behaviour **?**. However, even if you manage to stop this from happening, it's almost impossible to stop people from knowing that they're part of a psychological study. And the mere fact of knowing that someone is watching/studying you can have a pretty big effect on behaviour. This is generally referred to as **reactivity** or **demand effects**. The basic idea is captured by the Hawthorne effect: people alter their performance because of the attention that the study focuses on them. The effect takes its name from a the "Hawthorne Works" factory outside of Chicago (**?**). 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.

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

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

so much so that his or her behaviour becomes highly unnatural, or overly socially desirable.

1.13.10 Placebo effects

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

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". They refer to the fact that the choice of subpopulation from which you draw your participants, the location, timing and manner in which you run your study (including who collects the data) and the tools that you use to make your measurements might all be influencing the results. Specifically, the worry is that these things might be influencing the results in such a way that the results won't generalize to a wider array of people, places and measures.

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 hypothesising**. 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 reanalyse 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 behaviour into account. If you don't, then they can't. And that makes your behaviour 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 **?** with the depressing title "Why most published research findings are false". I'd also suggest taking a look at work by **?** 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

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

- **Introduction to psychological measurement**: What does it mean to operationalize a theoretical construct? What does it mean to have variables and take measurements?
- **Scales of measurement and types of variables**: Remember that there are *two* different distinctions here: there's the difference between discrete and continuous data, and there's the difference between the four different scale types (nominal, ordinal, interval and ratio).
- **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**: What makes an experiment an experiment? Is it a nice white lab coat, or does it have something to do with researcher control over variables?
- **Validity and its threats**: Does your study measure what you want it to? How might things go wrong? And is it my imagination, or was that a

very long list of possible ways in which things can go wrong?

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

1.15 Videos

1.15.1 Terms of Statistics

Chapter 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

Chapter by Matthew Crump

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:

-1020

1136
-172
-157
-384
142
398
-367
457
-79
407
541
-250
-301
-151
291
-609
722
-769
211
-996
-344
-283
-358
-616
-849
-1058
-175
-842
-345
16
285

851
-12
120
434
752
-316
-193
-441
398
1156
-798
105
320
-662
-50
375
151
401
620
-552
130
-614
-297
1128
-541
-448
290
638
-466
-335
246

-218
1460
1358
-389
-347
-134
476
134
820
614
174
494
-618
-568
385
477
227
456
643
-664
-589
-566
-212
-22
-24
54
924
-3
848
-17
327

-13
656
42
521
231
-199
-1020
311
-776
-110
717
374
419
444
-360
191
55
-665
1232
-125
-334
-425
-256
-158
364
492
9
268
425
-483
145

-12
1095
-427
-546
683
308
-1057
-374
-358
-126
-701
549
622
327
472
577
550
364
166
637
-54
202
50
852
301
125
-558
263
278
-379
-572

316
665
-320
-225
61
665
-1085
355
1008
-198
-158
123
-470
-290
-778
-240
1178
297
-205
-711
208
-456
368
293
72
-100
-342
-67
-403
-722
9

464
-740
-67
-163
356
189
-486
1467
451
611
426
-147
-140
-549
929
-405
-7
-36
285
-42
-67
231
-127
-620
1065
96
101
-575
-205
-242
591

-541
58
-244
1333
52
301
-198
-106
100
96
-151
457
-770
194
-203
135
-33
100
152
-172
-1384
799
268
-389
-650
312
-78
-809
-351
-483
-136

459
894
112
578
204
21
-574
14
135
-487
-615
272
219
550
245
226
-496
628
-288
658
750
122
-199
-378
215
345
800
106
502
-619
19

-773
449
275
-388
-34
-66
-85
273
-192
525
640
1061
326
-594
427
361
361
414
-203
876
853
-645
-156
256
249
-588
-48
-666
721
-357
-808

111
-160
862
-88
-642
564
-668
449
408
-295
-238
53
400
786
35
-460
511
1146
878
1088
-599
94
54
159
-160
-98
746
121
483
-738
264

987
218
924
579
-207
385
570
485
387
-175
-799
179
-580
794
-596
-96
84
-1
486
2
-113
233
-262
-283
82
157
-332
986
-402
-919
331

671
236
920
380
-619
-133
-45
-302
-258
179
-913
68
-525
-276
873
-341
48
729
370
-130
-7
511
1018
791
-579
160
-232
-565
-348
54
-929

641
429
-156
332
-202
701
49
-457
-22
57
-299
305
-281
220
-197
-86
397
-185
-544
789
-64
233
-453
-26
1222
645
-776
-129
-42
-77
-402

196
291
-73
-160
-808
-381
861
-454
267
313
1125
522
397
217
353
32
627
-345
1241
770
671
1101
8
139
-283
1770
-438
396
616
622
84

-595
-385
7
239
58
-239
-144
-575
567
29
212
-635
-113
361
154
89
-610
377
-78
186
-216
493
99
-395
-282
-107
-163
-45
199
-397
-625

-826

465

232

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, plotting time (o o oh) U can plot this

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 **??** 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.

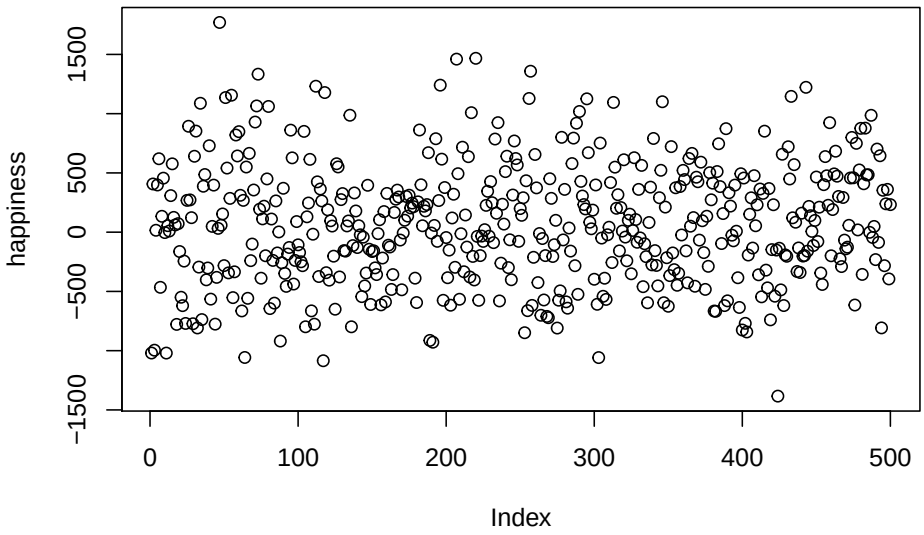
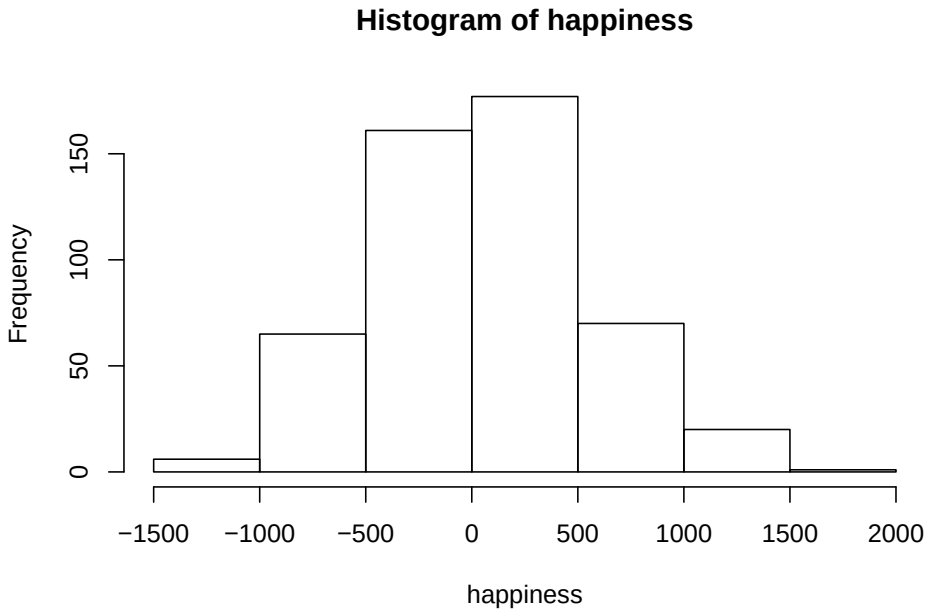


Figure 2.1: Pretend happiness ratings from 500 people

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 a histogram of the happiness data, and then explain it.



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.

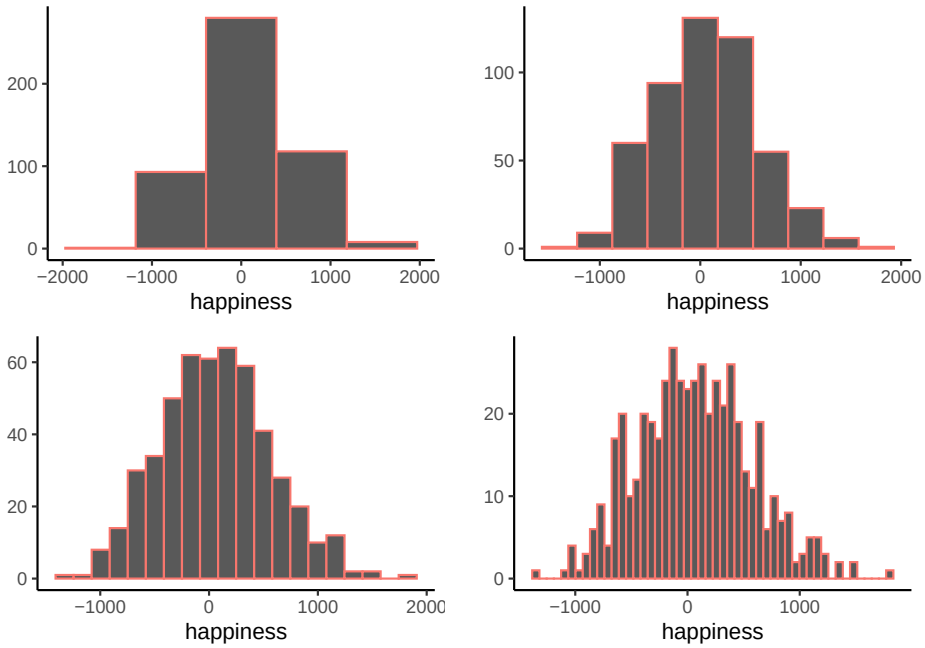


Figure 2.2: Four histograms of the same data using different bin widths

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 differentness: 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. 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.

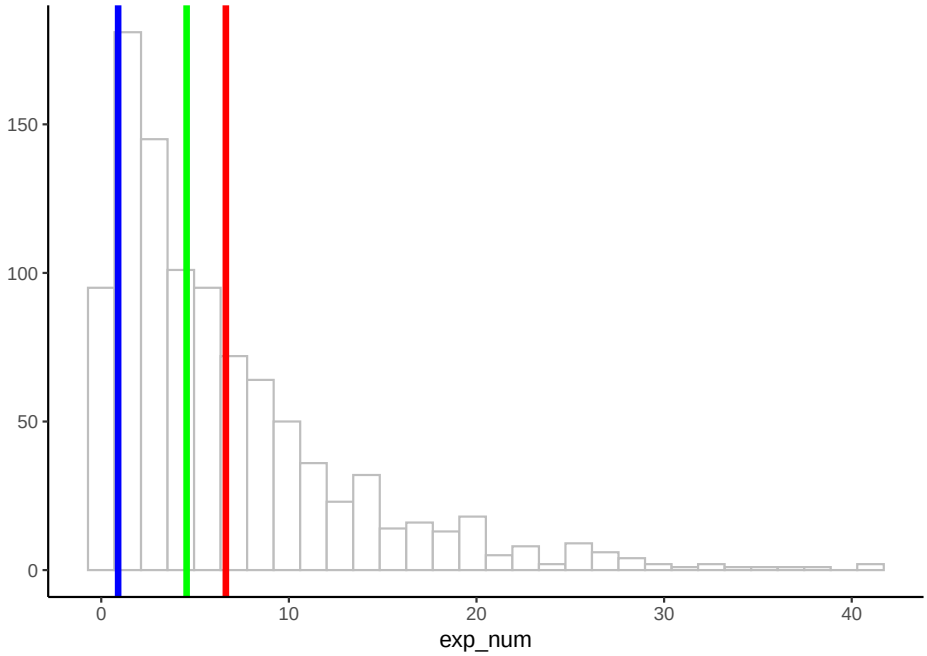


Figure 2.3: A histogram with the mean (red), the median (green), and the mode (blue)

2.5 Measures of Variation (Differentness)

What did you do when you wrote essayss 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 differentness 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	2	3
1	0		
2	1	0	
3	2	1	0

	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.

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:

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

We could also expand this to say:

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

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.

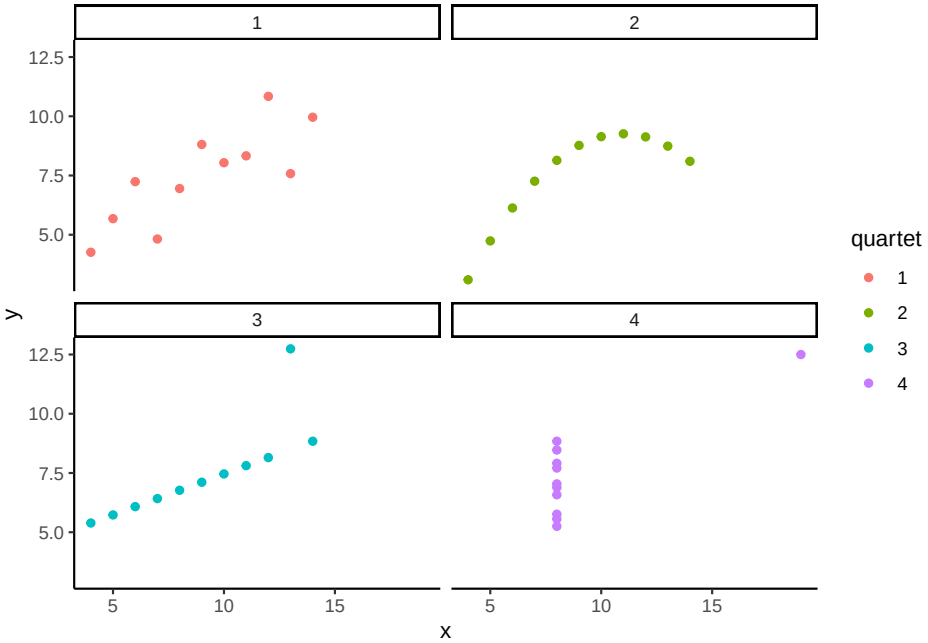
2.8 Remember to look at your data

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 shortcomings. 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 **?**. 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?



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 (**?**). 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

Chapter 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, belieifs, personality, color of your shoes, eyebrow length, number of cat's you see per day, frequency of subway delay, a lifetime supply of chocolate, etcetera etcetera (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

3

4

5

3

3

6

6

3

7

4

7

8

7

8

9

6

7

10

7

7

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. The plot below shows 100 dots for each subject.

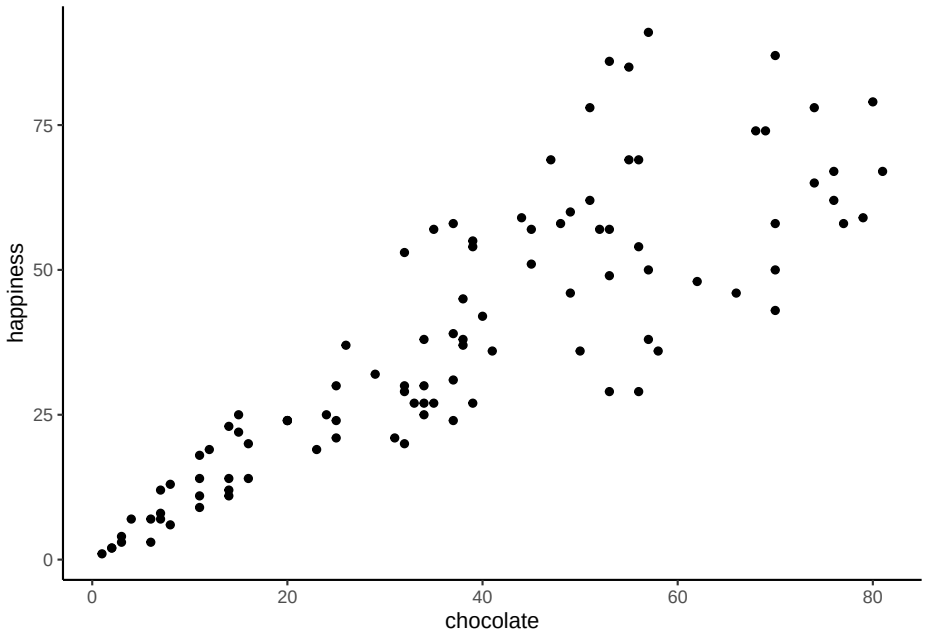


Figure 3.1: Imaginary data showing a positive correlation between amount of chocolate and amount happiness

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.

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 these graphs:

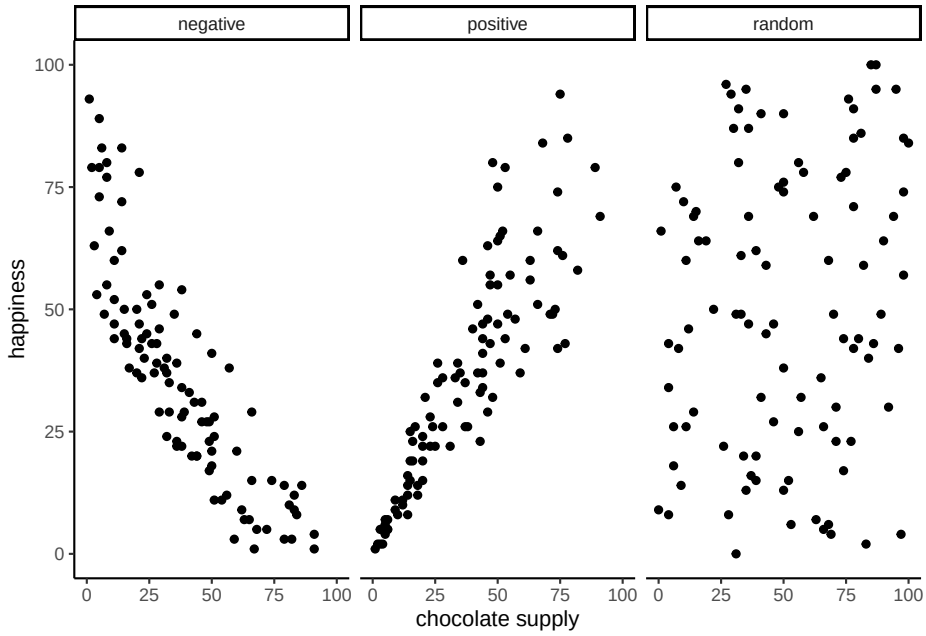


Figure 3.2: Three scatterplots showing negative, positive, and zero correlation

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.

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

If Beyoncé was a statistician, she might look at these scatter plots and want to "put a number on it". We think this is a good idea too. We've already learned how to create descriptive statistics for a single measure, like chocolate, or happiness (i.e., means, variances, etc.). Is it possible to create a descriptive statistic that summarized the relationship between two measures, all in one number? Can it be done? Karl Pearson to the rescue.

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

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 stats fights…why can't we all just get along in statistics.

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

3

4

5

3

3

6

6

3

7

4

7

8

7

8

9

6

7

10

7

7

Sums

41

44

Means

4.1

4.4

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. We 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: Succesfully playing paddycake occurs when two people coordinate their actions so they have postively 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
3
4
12
5
3
3
9
6
6
3
18
7
4
7

28

8

7

8

56

9

6

7

42

10

7

7

49

Sums

41

44

223

Means

4.1

4.4

22.3

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

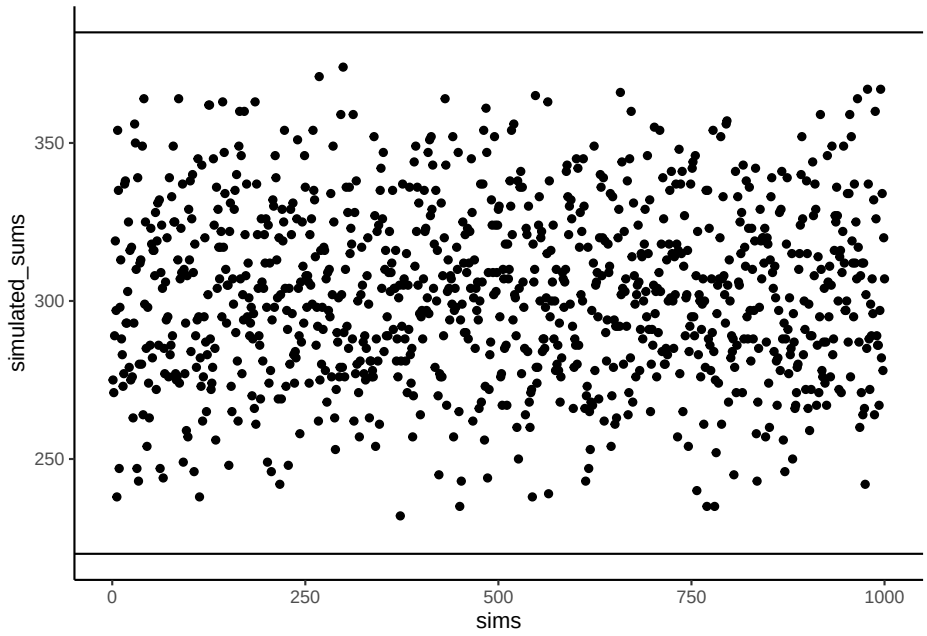
38.5

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:



The above graph 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 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

-3.4
10.54
2
2
2
-2.1
-2.4
5.04
3
2
2
-2.1
-2.4
5.04
4
3
4
-1.1
-0.4
0.44
5
3
3
-1.1
-1.4
1.54
6
6
3
1.9
-1.4

-2.66

7

4

7

-0.1

2.6

-0.26

8

7

8

2.9

3.6

10.44

9

6

7

1.9

2.6

4.94

10

7

7

2.9

2.6

7.54

Sums

41

44

0

0

43

Means

4.1

4.4

0

0

4.26

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.

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 to.

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 all of the dots are between -1 and 1. Neat huh.

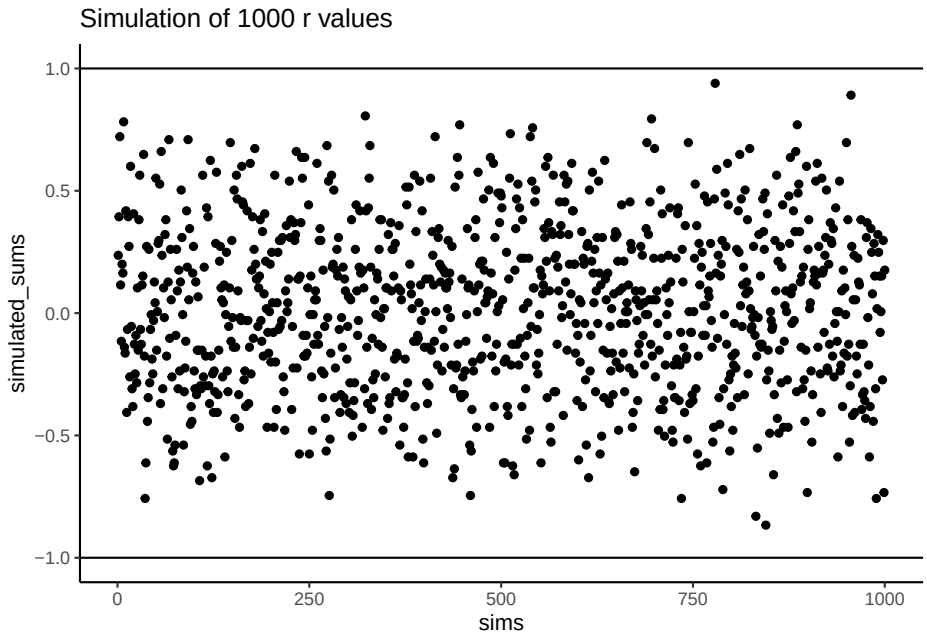


Figure 3.3: A simulation of of correlations. Each dot represents the r-value for the correlation between an X and Y variable that each contain the numbers 1 to 10 in random orders. The figure ilustrates that many r-values can be obtained by this random process

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. Here is a scatterplot showing the relationship between these two measures. Each dot represents means for different countries.

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:

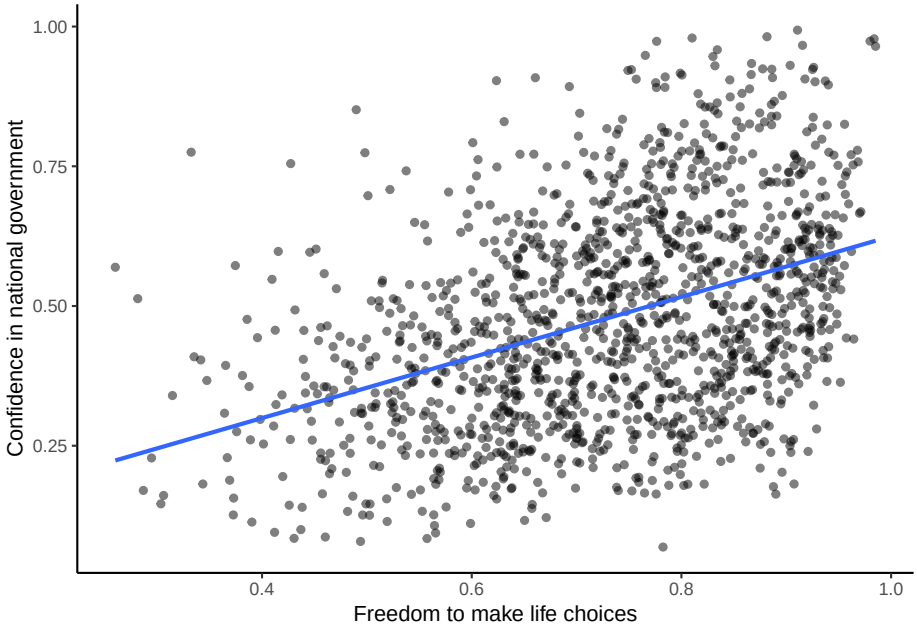


Figure 3.4: Relationship between freedom to make life choices and confidence in national government. Data from the world happiness report for 2018

```
## [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. Here's some scatter plots, same one's you've already seen. But,

we've added something new! Lines.

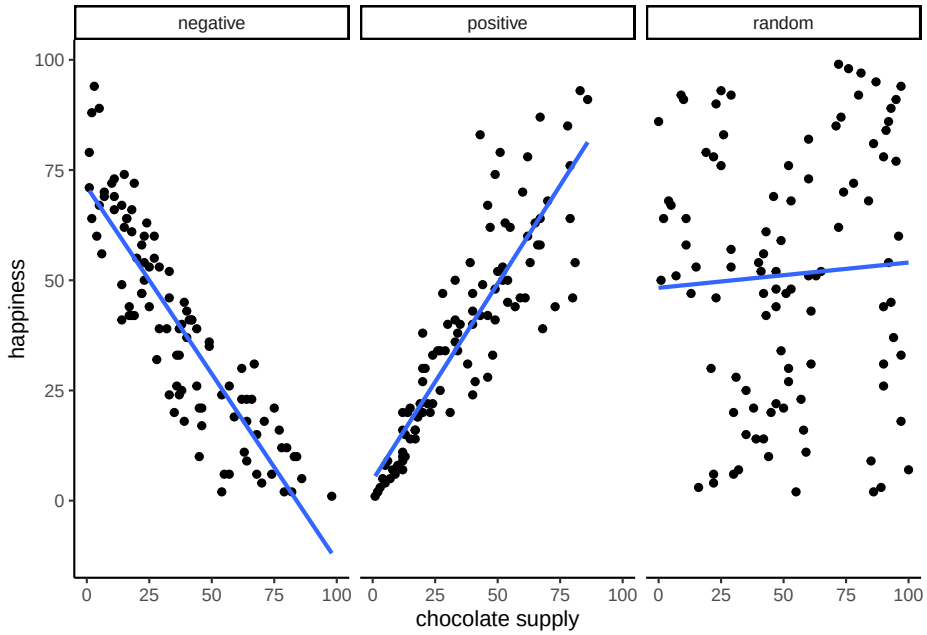


Figure 3.5: Three scatterplots showing negative, positive, and a random correlation (where the r-value is expected to be 0), along with the best fit regression line

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.

(NB: 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.

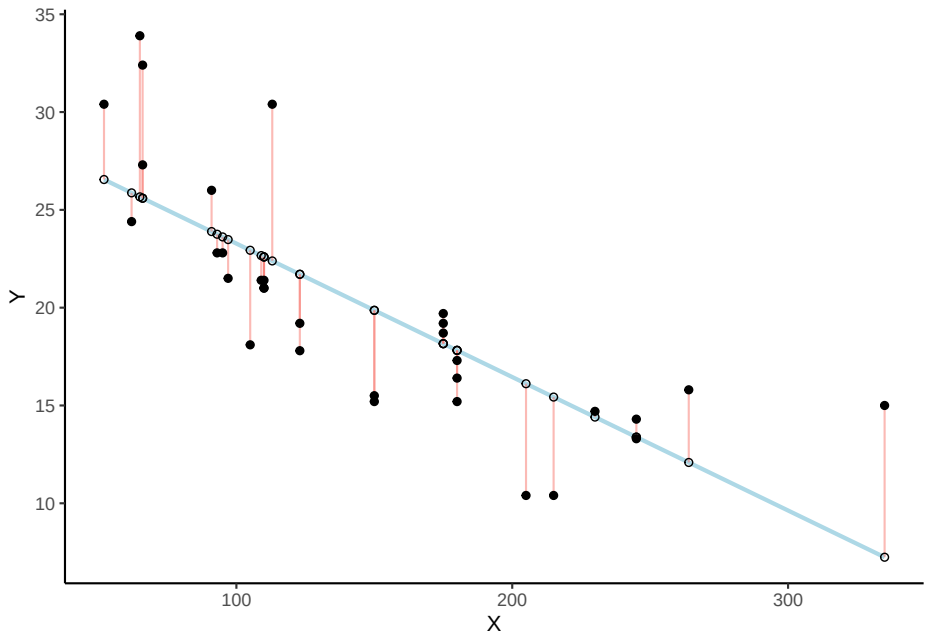


Figure 3.6: Black dots represent data points. The blue line is the best fit regression line. The white dots are repesent the predicted location of each black dot. The red lines show the error between each black dot and the regression line. The blue line is the best fit line because it minimizes the error shown by the red lines

There's a lot going on in this graph. 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