

Answering questions with data

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

2018 Last Compiled 2018-10-13

Contents

Preface	8
0.1 Important notes	9
0.2 Copying the textbook	10
1 Why Statistics?	13
1.1 On the psychology of statistics	13
1.2 The cautionary tale of Simpson's paradox	15
1.3 Statistics in psychology	18
1.4 Statistics in everyday life	19
1.5 There's more to research methods than statistics	19
1.6 A brief introduction to research design	20
1.7 Introduction to psychological measurement	20
1.8 Scales of measurement	22
1.9 Assessing the reliability of a measurement	27
1.10 The role of variables: predictors and outcomes	27
1.11 Experimental and non-experimental research	28
1.12 Assessing the validity of a study	30
1.13 Confounds, artifacts and other threats to validity	33
1.14 Summary	40
2 Describing Data	41
2.1 This is what too many numbers looks like	41
2.2 Look at the data	43
2.3 Important Ideas: Distribution, Central Tendency, and Variance	45
2.4 Measures of Central Tendency (Sameness)	46
2.5 Measures of Variation (Differentness)	50
2.6 Using Descriptive Statistics with data	54
2.7 Rolling your own descriptive statistics	55
2.8 Remember to look at your data	56
3 Correlation	59
3.1 If something caused something else to change, what would that look like?	60
3.2 Pearson's r	63
3.3 Turning the numbers into a measure of co-variance	64
3.4 Examples with Data	68
3.5 Regression: A mini intro	71
3.6 Interpreting Correlations	76
3.7 Summary	86
4 Probability, Sampling, and Estimation	89
4.1 How are probability and statistics different?	90
4.2 What does probability mean?	91

4.3	Basic probability theory	95
4.4	The binomial distribution	99
4.5	The normal distribution	103
4.6	Other useful distributions	107
4.7	Summary of Probability	107
4.8	Samples, populations and sampling	107
4.9	The law of large numbers	113
4.10	Sampling distributions and the central limit theorem	114
4.11	The central limit theorem	117
4.12	z-scores	123
4.13	Estimating population parameters	126
4.14	Estimating a confidence interval	134
4.15	Summary	135
5	Foundations for inference	137
5.1	Brief review of Experiments	137
5.2	The data came from a distribution	138
5.3	Is there a difference?	148
5.4	Chance makes some differences more likely than others	150
5.5	The Crump Test	155
5.6	The randomization test (permutation test)	168
6	t-Tests	175
6.1	Check your confidence in your mean	176
6.2	One-sample t-test: A new t-test	177
6.3	Paired-samples <i>t</i> -test	183
6.4	The paired samples t-test strikes back	194
6.5	Independent samples t-test: The return of the t-test?	197
6.6	Simulating data for t-tests	198
7	ANOVA	209
7.1	ANOVA is Analysis of Variance	209
7.2	One-factor ANOVA	209
7.3	What does F mean?	215
7.4	ANOVA on Real Data	223
7.5	ANOVA Summary	226
8	Repeated Measures ANOVA	227
8.1	Repeated measures design	227
8.2	Partitioning the Sums of Squares	228
8.3	Calculating the RM ANOVA	229
8.4	Things worth knowing	235
8.5	Real Data	238
8.6	Summary	240
9	Factorial ANOVA	241
9.1	Factorial basics	242
9.2	Purpose of Factorial Designs	243
9.3	Graphing the means	249
9.4	Knowing what you want to find out	249
9.5	Simple analysis of 2x2 repeated measures design	252
9.6	2x2 Between-subjects ANOVA	259
9.7	Fireside chat	263
9.8	Real Data	264
9.9	Factorial summary	268

10 More On Factorial Designs	269
10.1 Looking at main effects and interactions	269
10.2 Interpreting main effects and interactions	272
10.3 Mixed Designs	275
10.4 More complicated designs	275
11 Simulating Data	283
11.1 Reasons to simulate	283
11.2 Simulation overview	284
11.3 Simulating t-tests	286
11.4 Simulating one-factor ANOVAs	288
11.5 Other resources	289
12 Thinking about answering questions with data	291
12.1 Effect-size and power	291
12.2 Power	295
12.3 Planning your design	298
12.4 Some considerations	299

Preface

Answering questions with data

Introductory Statistics for
Psychology Students



First Draft (version 0.9 = August 7th, 2018) Last Compiled: 2018-10-13

0.1 Important notes

This is a free textbook teaching introductory statistics for undergraduates in Psychology. There is a companion lab manual for this textbook here. Both books are released under a creative commons licence CC BY-SA 4.0. Click the link to read more about the license, or read more below in the license section.

This textbook is part of a larger OER course package for teaching undergraduate statistics in Psychology. Team members include, Matthew Crump, Alla Chavarga, Anjali Krishnan, Jeffrey Suzuki, and Stephen Volz. Jeffrey contributed the YouTube videos peppered throughout the textbook. Alla, Anjali, and Stephen wrote the lab manual exercises for SPSS, JAMOVI, and Excel. Matt Crump wrote the lab manual exercises for R. As this OER comes together, we will be providing a course website, written in R Markdown, as well as slide decks for the lectures (these will be more fully available by the end of fall 2018). As a result, this textbook, the lab manual, the course website, and the slide decks will all be free, under a creative commons license. The source code for all material is contained in the GitHub repositories for each, and they are written in R-markdown, so they are relatively easy to copy and edit. Have Fun!

0.1.1 Attributions

This textbook was primarily written by Matthew J. C. Crump.

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.

0.1.2 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.
2. The source code for this textbook is available in my GitHub repo statsforpsych

- a. Edit the .Rmd files, and push them back
- b. The edit link in the top bar of the textbook should automatically take you to the source .Rmd file

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

Adapted nearly verbatim from Chapters 1 and 2 in Navarro, D. “Learning Statistics with R.” <https://compcogscisydney.org/learning-statistics-with-r/>

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, Evans et al. (1983) ran an experiment looking at exactly this. What they found is that when pre-existing biases (i.e., beliefs) were in agreement with the structure of the data, everything went the way you'd hope:

	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 postgraduate courses. Specifically, the thing that caused the problem was that the gender breakdown of their admissions, which looked like this:

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

and they were worried about being sued. Given that there were nearly 13,000 applicants, a difference of 9% in admission rates between males and females is just way too big to be a coincidence. Pretty compelling data, right? And if I were to say to you that these data *actually* reflect a weak bias in 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 (Bickel et al., 1975) they told a rather different story. Specifically, when they looked at it on a department by department basis, it turned out that most of the departments actually had a slightly *higher* success rate for female applicants than for male applicants. 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 1.1 we see that this trend is systematic, and quite striking. This effect is known as **Simpson's paradox**. It's not common, but it does happen in real life, and most people are very surprised by it when they first encounter it, and many people refuse to even believe that it's real. It is very real. And while there are lots of very subtle statistical lessons buried in there, I want to use it to make a much more important point ...doing research is hard, and there are *lots* of subtle, counterintuitive traps lying in wait for the unwary. That's reason #2 why scientists love statistics, and why we teach research methods. Because science is hard, and the truth is sometimes cunningly hidden in the nooks and crannies of complicated data.

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

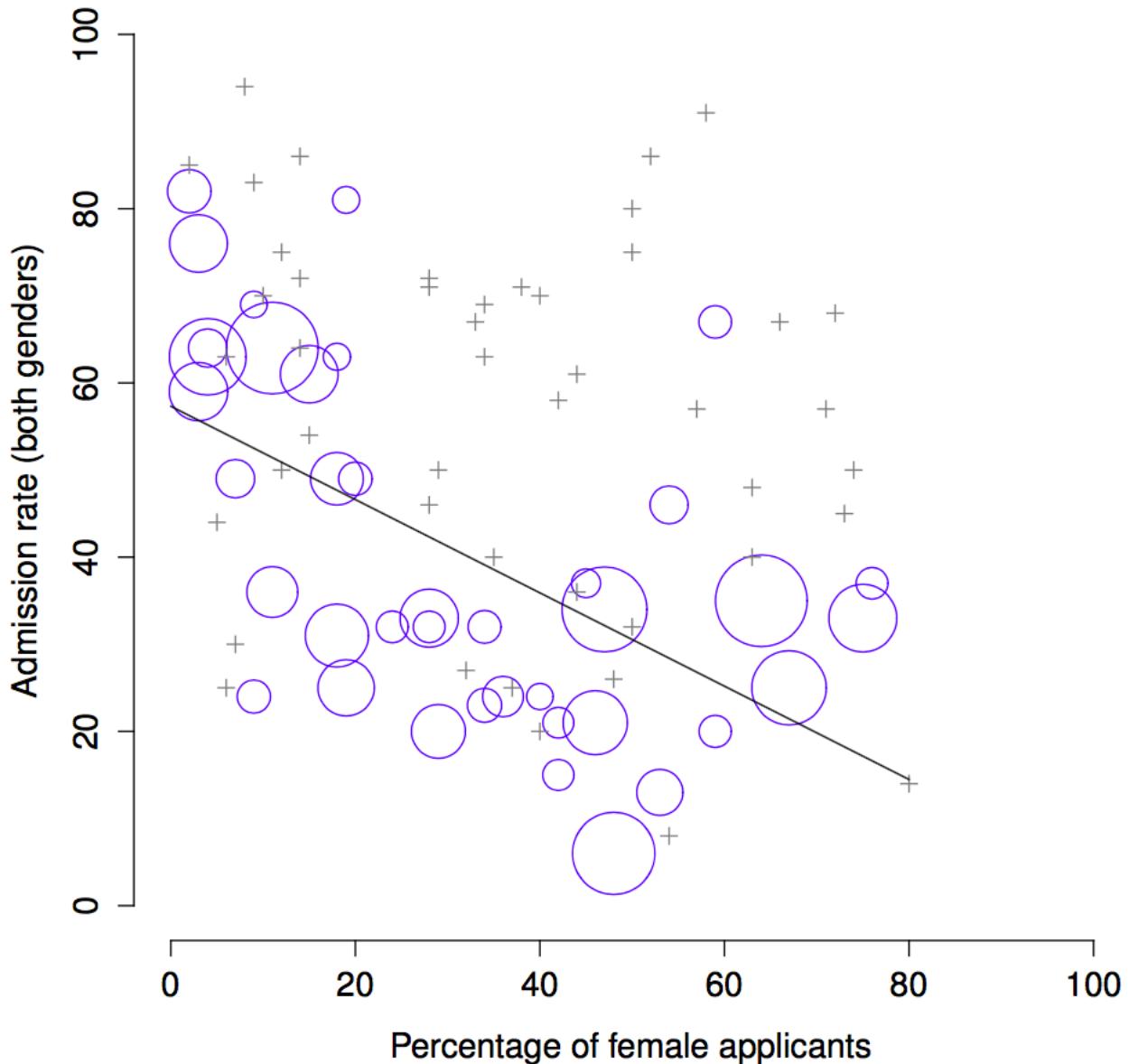


Figure 1.1: The Berkeley 1973 college admissions data. This figure plots the admission rate for the 85 departments that had at least one female applicant, as a function of the percentage of applicants that were female. The plot is a redrawing of Figure 1 from Bickel et al. (1975). Circles plot departments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

choose to apply to, and the decision making takes place at the level of the department it can hardly be held accountable for any biases that those choices produce.

That was the basis for my somewhat glib remarks earlier, but that's not exactly the whole story, is it? After all, if we're interested in this from a more sociological and psychological perspective, we might want to ask *why* there are such strong gender differences in applications. Why do males tend to apply to engineering more often than females, and why is this reversed for the English department? And why is it the case that the departments that tend to have a female-application bias tend to have lower overall admission rates than those departments that have a male-application bias? Might this not still reflect a gender bias, even though every single department is itself unbiased? It might. Suppose, hypothetically, that males preferred to apply to "hard sciences" and females prefer "humanities". And suppose further that the reason for why the humanities departments have low admission rates is because the government doesn't want to fund the humanities (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 Campbell and Stanley (1963) for the discussion of study design, and Stevens (1946) for the discussion of scales of measurement. Later versions will attempt to be more precise in the citations.

1.7 Introduction to 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 *Klinefelter's syndrome (XXY)*, which is more similar to male than to female. And I imagine there are other possibilities too.
- My **self-identified gender** is also very likely to be *male* or *female*, but it doesn't have to agree with my chromosomal gender. I may also choose to identify with *neither*, or to explicitly call myself *transgender*.

As you can see, for some things (like age) it seems fairly obvious what the set of possible measurements should be, whereas for other things it gets a bit tricky. 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 do we allow them to choose between? Is it enough to allow only “male” or “female”? Do you need an “other” option? Or should we not give people any specific options, and let them answer in their own words? And if you open up the set of possible values to include all verbal 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 are a lot of issues that are specific to each individual research project, there are some aspects to it that are pretty general.

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

- **A theoretical construct.** This is the thing that you’re trying to take a measurement of, like “age”, “gender” or an “opinion”. A theoretical construct can’t be directly observed, and often they’re actually a bit vague.
- **A measure.** The measure refers to the method or the tool that you use to make your observations. A question in a survey, a behavioural observation or a brain scan could all count as a measure.
- **An 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 $1\text{st} > 2\text{nd}$, and we know that $2\text{nd} > 3\text{rd}$, 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.

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

	continuous	discrete
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 its 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’re 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 its 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 introduce problems? As always, the answer is maybe. If the question that wasn't answered was on the last page of the questionnaire, and those 80 surveys were returned with the last page missing, there's a good chance that the missing data isn't a big deal: probably the pages just fell off. However, if the question that 80 people didn't answer was the most confrontational or invasive personal question in the questionnaire, then almost certainly you've got a problem. In essence, what you're dealing with here is what's called the problem of **missing data**. If the data that is missing was "lost" randomly, then it's not a big problem. If it's missing systematically, then it can be a big problem.

1.13.7 Regression to the mean

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

Here's an example. Suppose 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 (Kahneman and Tversky, 1973)

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, Pfungst (1911; Hothersall, 2004). Clever Hans was a horse that apparently was able to read and count, and perform other human like feats of intelligence. After Clever Hans became famous, psychologists started examining his 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 prophecy.

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 Rosenthal (1966). However, even if you manage to stop this from happening, it’s almost impossible to stop people from knowing that they’re part of a psychological study. And the mere fact of knowing that someone is watching/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 (Adair, 1984). A study done in the 1920s looking at the effects of lighting on worker productivity at the factory turned out to be an effect of the fact that the workers knew they were being studied, rather than the lighting.

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 “misdemeanor”.** 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 (artificially) 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 Ioannidis (2005) with the depressing title “Why most published research findings are false”. I’d also suggest taking a look at work by Kühberger et al. (2014) presenting statistical evidence that this actually happens in psychology.

There’s probably a lot more issues like this to think about, but that’ll do to start with. What I really want to point out is the blindingly obvious truth that real world science is conducted by actual humans, and only the most gullible of people automatically 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 Campbell and Stanley (1963), but there are of course a large number of textbooks out there on research design. Spend a few minutes with your favourite search engine and you’ll find dozens.

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:

659	607	736	-331	569	-591	-74	-614	558	393
-605	-395	1332	-1223	500	393	20	137	-257	291
-353	-88	335	686	-120	-394	182	527	691	12
492	130	341	1147	196	-251	233	627	695	-375
-269	92	207	355	-350	4	747	280	27	-550
152	-286	-80	86	1134	141	-246	347	531	180
194	638	-309	435	509	190	229	-274	-289	-91
17	-261	490	227	-57	-492	664	149	743	440
568	-234	26	135	63	917	598	770	-518	-214
353	-986	510	777	702	-167	360	-720	126	689
458	-475	-279	-692	-603	882	-107	-449	-1133	-149
-889	-45	213	768	-214	807	849	87	46	747
-545	-20	216	640	319	-677	-11	-101	107	51
-425	-167	74	307	1668	762	-389	714	769	-24
759	63	244	243	17	29	-217	110	-205	-160
-371	89	440	607	-337	-939	376	82	851	485
400	139	-341	35	-91	-233	523	-75	515	164
-284	60	284	384	387	681	83	536	269	-20
-1160	-557	304	77	302	-123	1291	851	-133	-520
712	763	-118	205	-177	-933	3	321	-255	-185
660	575	-301	467	-757	169	564	925	924	618
-165	352	492	-288	-49	217	643	127	433	532
-438	299	-280	298	626	-79	-627	18	-518	-170
391	-340	4	1035	601	-156	527	-285	235	683
-318	-53	-1140	59	396	-351	84	450	332	-487
404	-51	218	266	866	468	317	1070	5	470
1278	539	-131	218	720	43	441	372	716	690
-268	961	411	-170	332	-361	-515	167	-538	541
702	220	157	-519	560	-33	197	83	-182	-31
355	-715	906	983	158	-42	29	222	-98	1014
1235	533	-90	1479	676	424	1021	218	545	1067
-902	257	193	-161	-85	-209	157	-168	658	115
158	-317	-548	-81	-47	144	-235	783	-176	-318
-364	-655	128	70	-90	-402	410	214	-341	292
-151	76	-388	-927	543	-529	217	-212	911	103
212	252	283	-403	-672	410	100	-104	448	217
168	120	408	105	-533	-670	191	76	-85	-143
513	377	94	-1024	523	-286	695	246	342	1019
-356	-473	-952	302	799	75	38	-42	207	398
-521	-264	182	652	625	456	70	-395	889	67
-255	-333	254	766	371	-502	200	303	193	178
98	-264	-547	452	224	1028	-220	3	-601	83
-119	-77	498	330	767	-119	279	849	-558	35
400	644	357	162	405	141	-177	872	43	-54
54	111	1258	-427	-137	797	758	-54	-297	1140
-299	510	61	-580	363	42	96	-279	-987	1200
959	295	-311	188	-99	-198	-191	510	586	181
224	266	-22	1264	-39	1220	-450	149	264	222
773	-214	-9	65	-568	-966	-193	751	273	202
-352	375	-154	205	839	-221	-424	945	187	-389

Now, what are you going to do 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.

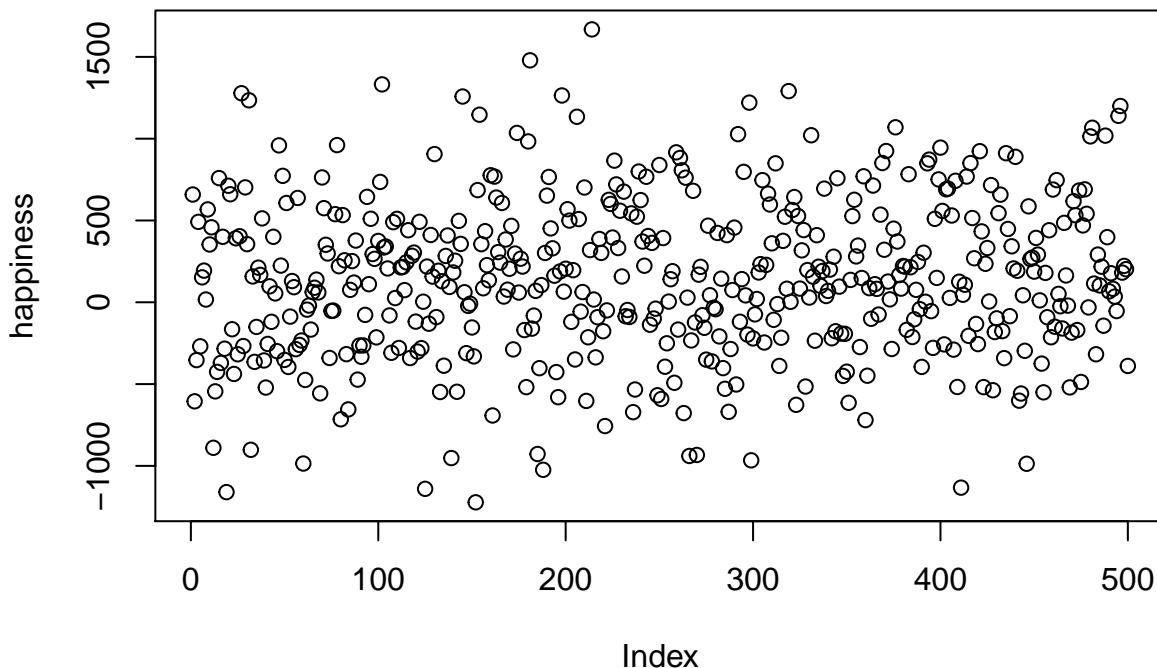


Figure 2.1: Pretend happiness ratings from 500 people

Figure 2.1 shows 500 measurements of happiness. The graph has two axes. The horizontal **x-axis**, going from left to right is labeled “Index”. The vertical **y-axis**, going up and down, is labelled “happiness”. Each dot represents one measurement of every person’s happiness from our pretend study. Before we talk about what we can and cannot see about the data, it is worth mentioning that the way you plot the data will make some things easier to see and some things harder to see. So, what can we now see about the data?

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

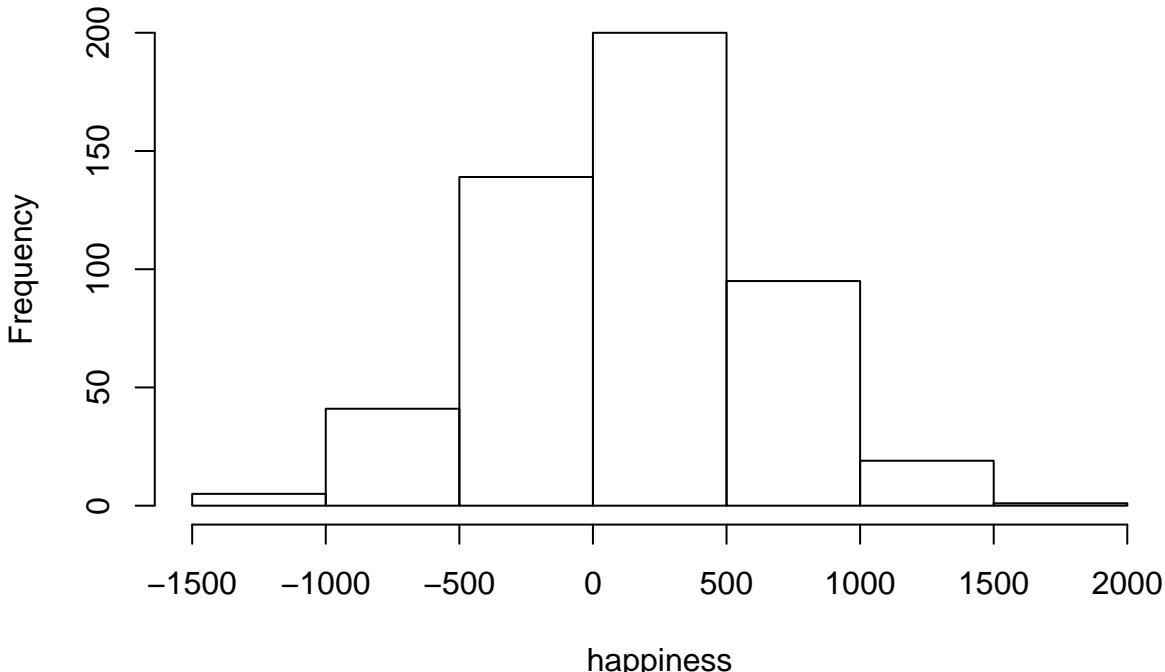
Take home: we can see all the numbers at once by putting them in a plot, and that is much easier and more helpful than looking at the raw numbers.

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

2.2.2 Histograms

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

Histogram of happiness



The dots have disappeared, and now we see 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.

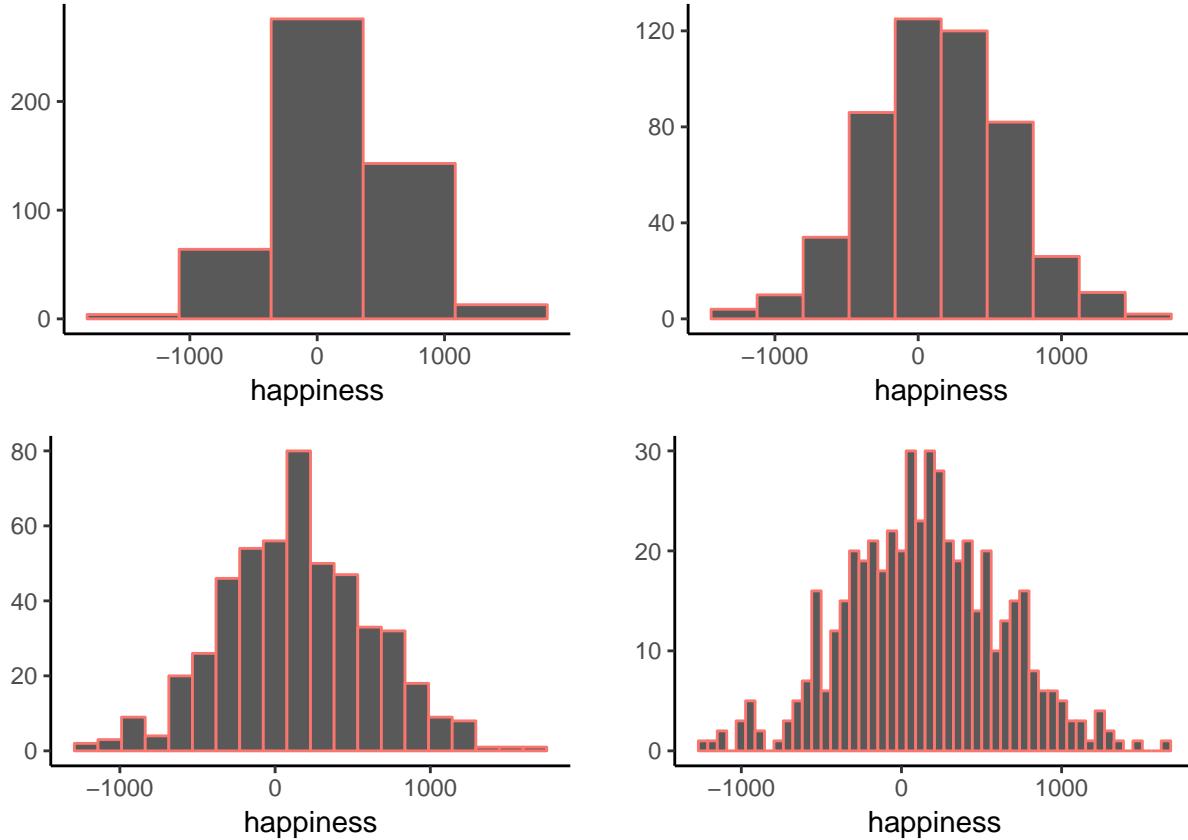


Figure 2.2: Four histograms of the same data using different bin widths

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 sensitutes 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 numbers 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 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 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 know what the average of a bunch of numbers is? It's the sum of the numbers, divided by the number of numbers right? How do we express that idea in a formula? Just like this:

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

"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:

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

"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:

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

Divide em:

$$\text{mean} = 27 / 5 = 5.4$$

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

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

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.

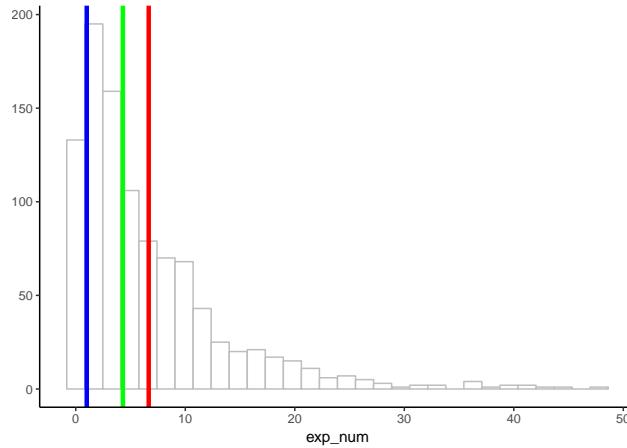


Figure 2.3: A histogram with the mean (red), the median (green), and the mode (blue)

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.

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 $500 \times 500 = 250,000$ 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	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:

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

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:

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

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:

$\text{variance} = \frac{\text{SS}}{N}$, 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:

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

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.

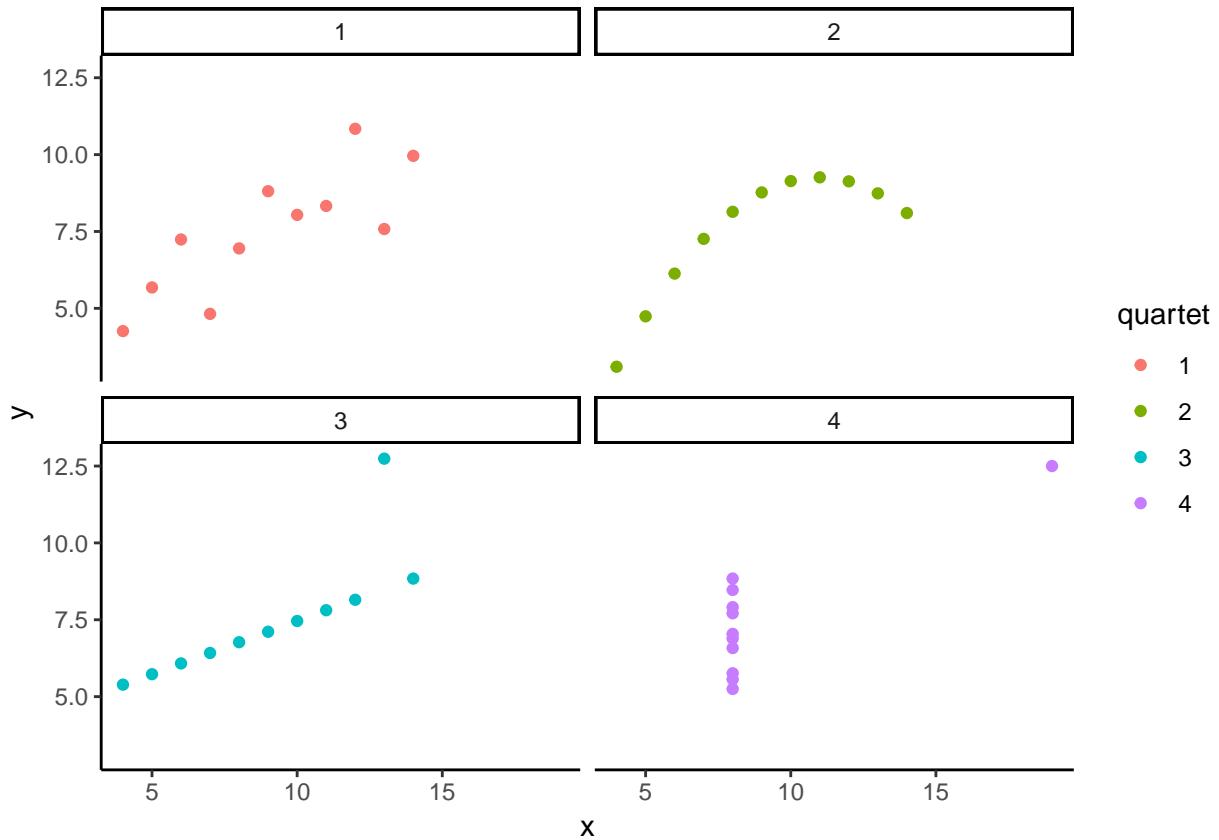
2.8 Remember to look at your data

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 short-comings. This is very true. They are compressed summaries of large piles of numbers. They will almost always be unable to represent all of the numbers fairly. There are also different kinds of descriptive statistics that you could use, and it sometimes not clear which one's you should use.

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

2.8.1 Anscombe's Quartet

To hit this point home, and to get you thinking about the issues we discuss in the next chapter, check this out. It's called Anscombe's Quartet, because these interesting graphs and numbers and numbers were produced by Anscombe (1973). 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 (Matejka and Fitzmaurice, 2017). Scroll down to see the examples. You will be looking at dot plots. The dot plots show many different patterns, including dinosaurs! What's amazing is that all of the dots have very nearly the same descriptive statistics. Just another reminder to look at your data, it might look like a dinosaur!

Chapter 3

Correlation

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

Chapter by Matthew Crump

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 sidestep those very important questions, and assume that, happiness is a thing, and our measure of happiness measures something about how happy people are.

OK then, after we have measured some happiness, I bet you can think of some more pressing questions. For example, what causes happiness to go up or down. If you knew the causes of happiness what could you do? How about increase your own happiness; or, help people who are unhappy; or, better appreciate why Eeyore

from Winnie the Pooh is unhappy; or, present valid scientific arguments that argue against incorrect claims about what causes happiness. A causal theory and understanding of happiness could be used for all of those things. How can we get there?

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

As a person who has happiness, you might already have some hunches about what causes changes in happiness. For example things like: weather, friends, music, money, education, drugs, books, movies, beliefs, personality, color of your shoes, eyebrow length, number of cat's you see per day, frequency of subway delay, a lifetime supply of chocolate, 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.

3.1. IF SOMETHING CAUSED SOMETHING ELSE TO CHANGE, WHAT WOULD THAT LOOK LIKE?61

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

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.

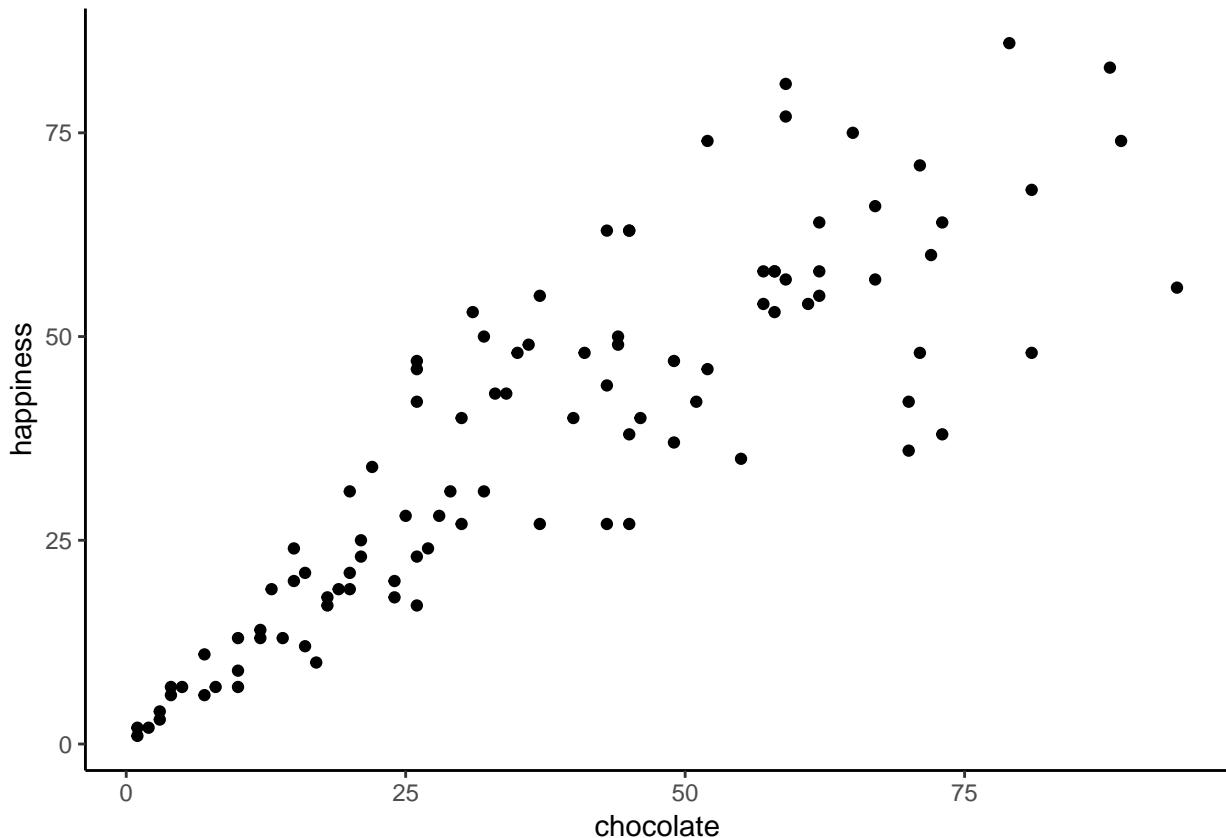


Figure 3.1: Imaginary data showing a positive correlation between amount of chocolate and amount happiness

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:

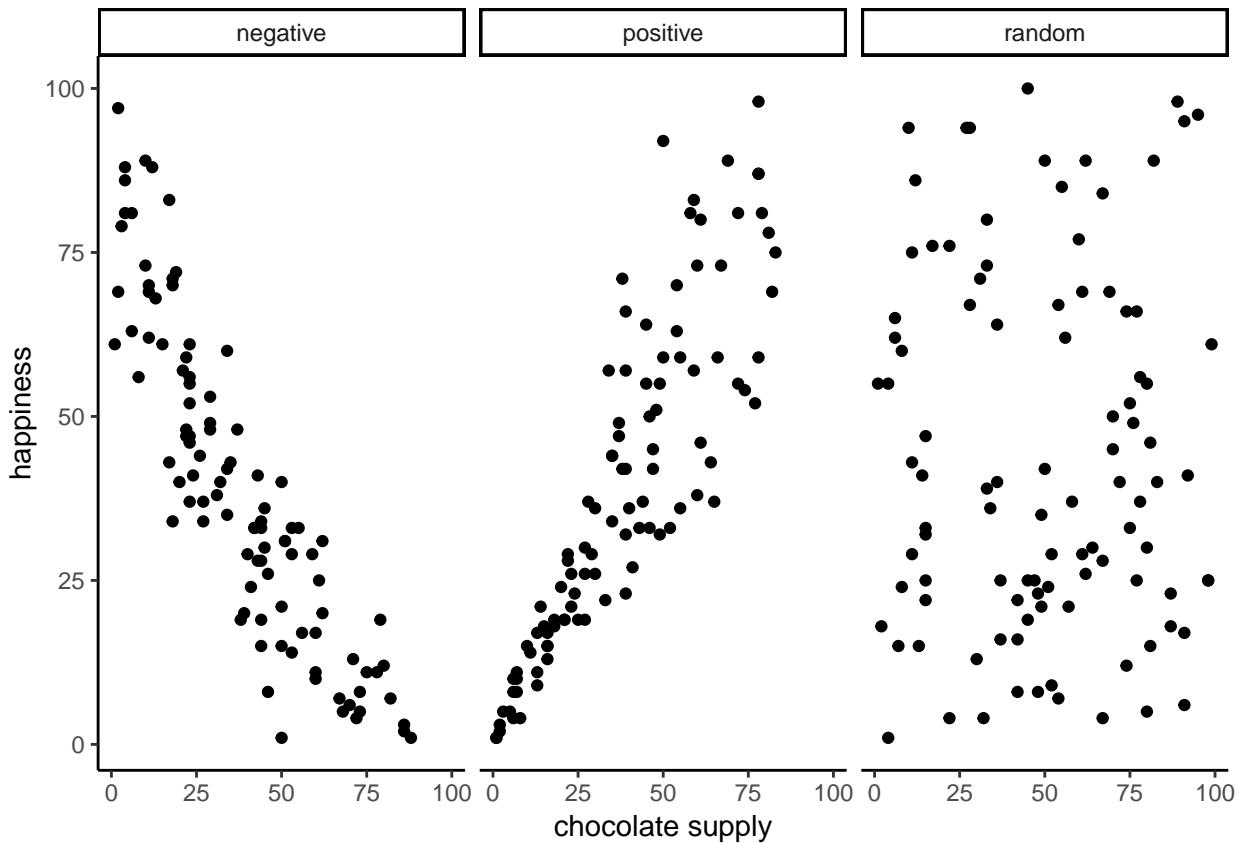


Figure 3.2: Three scatterplots showing negative, positive, and zero correlation

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” (Salsburg, 2001)

There's a statistic for that, and Karl Pearson invented it. Everyone now calls it, “Pearson's r ”. We will find out later that Karl Pearson was a big-wig editor at Biometrika in the 1930s. He took a hating to another big-wig statistician, Sir Ronald Fisher (who we learn about later), and they had some 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	1	2
3	2	2
4	3	4
5	3	3
6	4	6
7	4	7
8	5	7
9	8	7
10	7	6
Sums	38	45
Means	3.8	4.5

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: Successfully playing paddycake occurs when two people coordinate their actions so they have positively shared co-variance.

3.3 Turning the numbers into a measure of co-variance

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

subject	chocolate	happiness	Chocolate_X_Happiness
1	1	1	1
2	1	2	2
3	2	2	4
4	3	4	12
5	3	3	9
6	4	6	24
7	4	7	28
8	5	7	35
9	8	7	56
10	7	6	42
Sums	38	45	213
Means	3.8	4.5	21.3

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?.

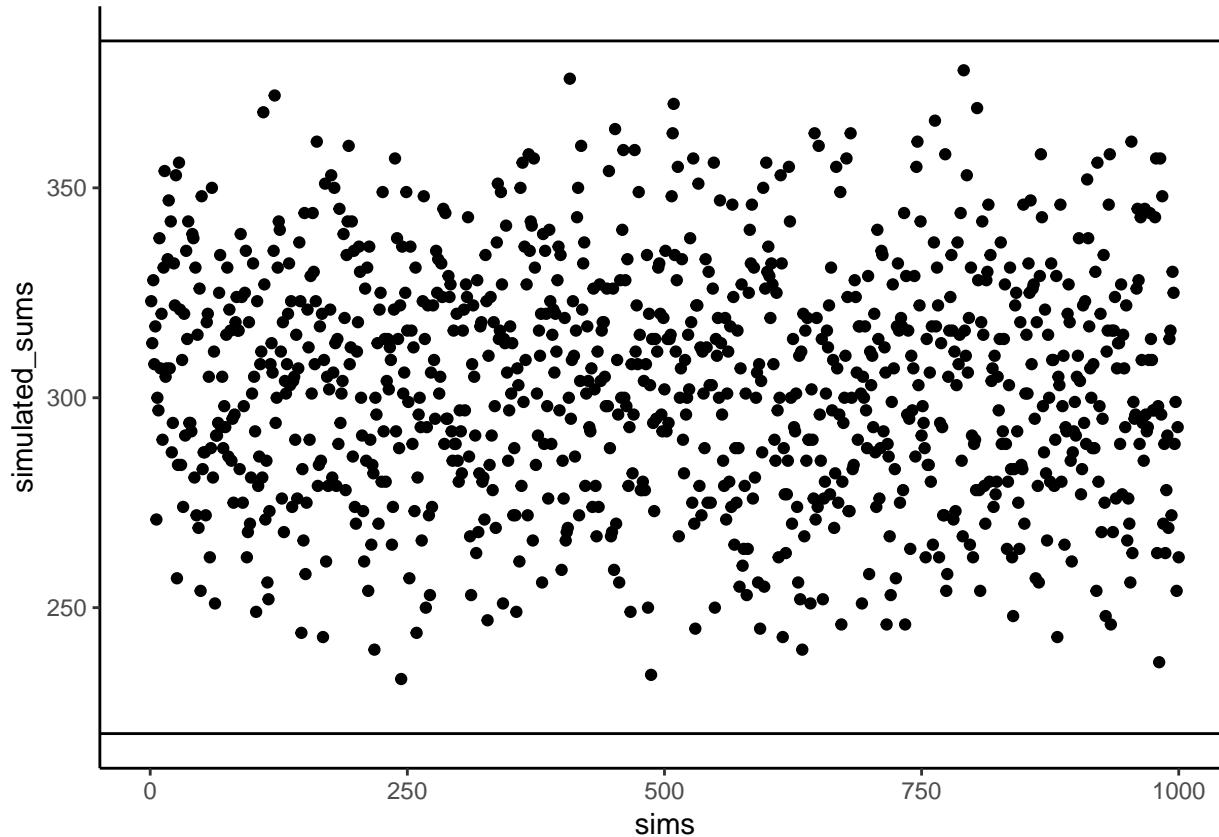
Then next step is to think about how the products of two measures sum together, depending on how they line up. Let's look at another table:

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 now, 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	-2.8	-3.5	9.8
2	1	2	-2.8	-2.5	7
3	2	2	-1.8	-2.5	4.5
4	3	4	-0.8	-0.5	0.4
5	3	3	-0.8	-1.5	1.2
6	4	6	0.2	1.5	0.3
7	4	7	0.2	2.5	0.5
8	5	7	1.2	2.5	3
9	8	7	4.2	2.5	10.5
10	7	6	3.2	1.5	4.8
Sums	38	45	0	0	42
Means	4	4	0	0	4

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:

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

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 :

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

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.

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.

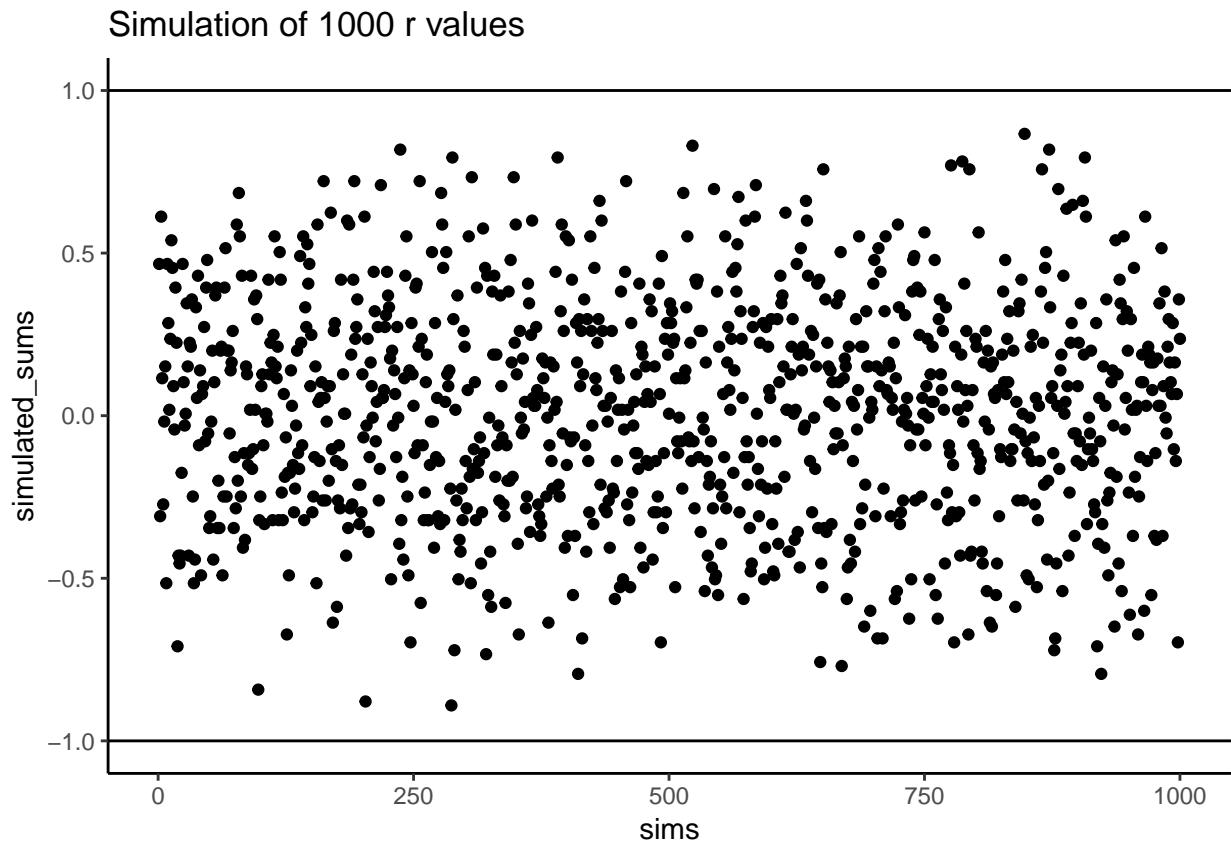


Figure 3.3: A simulation of correlations. Each dot represents the r-value for the correlation between an X and Y variable that each contain the numbers 1 to 10 in random orders. The figure illustrates that many r-values can be obtained by this random process

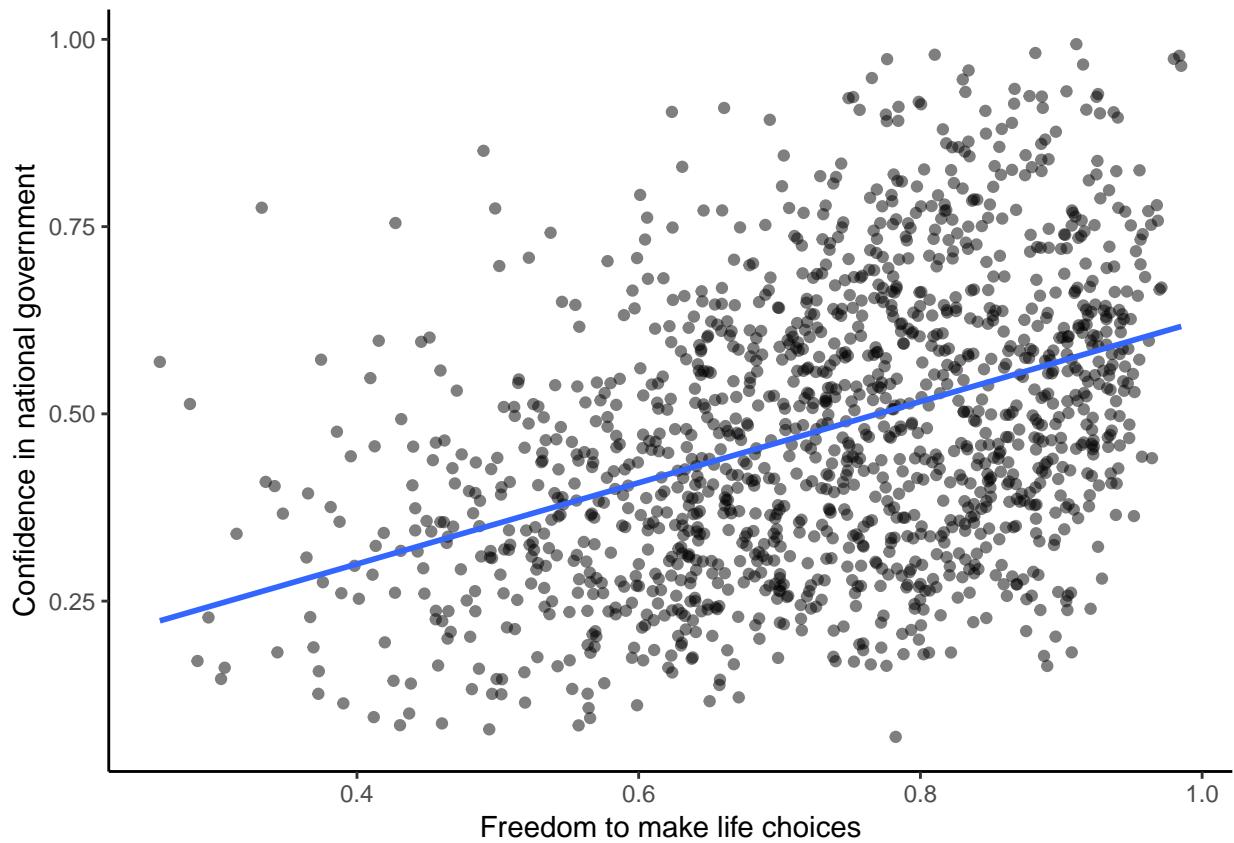


Figure 3.4: Relationship between freedom to make life choices and confidence in national government. Data from the world happiness report for 2018

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

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

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

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

3.5 Regression: A mini intro

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

First, let’s look at a linear regression. This way we can see what we’re trying to learn about. Here’s some scatter plots, same one’s you’ve already seen. But, we’ve added something new! Lines.

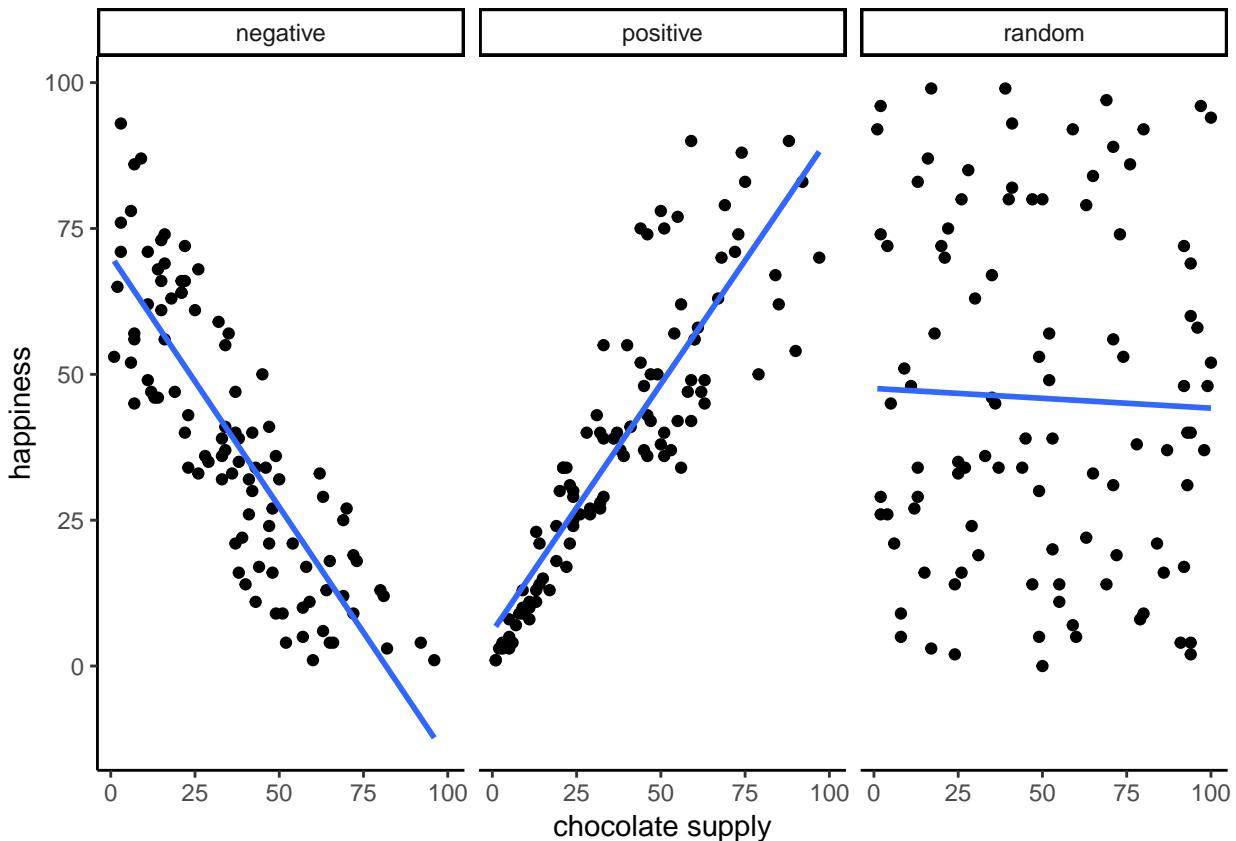


Figure 3.5: Three scatterplots showing negative, positive, and a random correlation (where the r -value is expected to be 0), along with the best fit regression line

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 through 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 its 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 at 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.

R code for plotting residuals thanks to Simon Jackson's blog post: <https://drsimonj.svble.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.

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 X increases, Y tends to decrease. We drew a regression line through the data, that's the blue line. There's these little white dots too. This is where the line thinks the black dots should be. The red lines are the important residuals we've been talking about. Each black dot has a red line that drops straight down, or straight up from the location of the black dot, and lands directly on the line. We can already see that many of the dots are not on the line, so we already know the line is "off" by some amount for each dot. The red line just makes it easier to see exactly how off the line is.

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

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

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

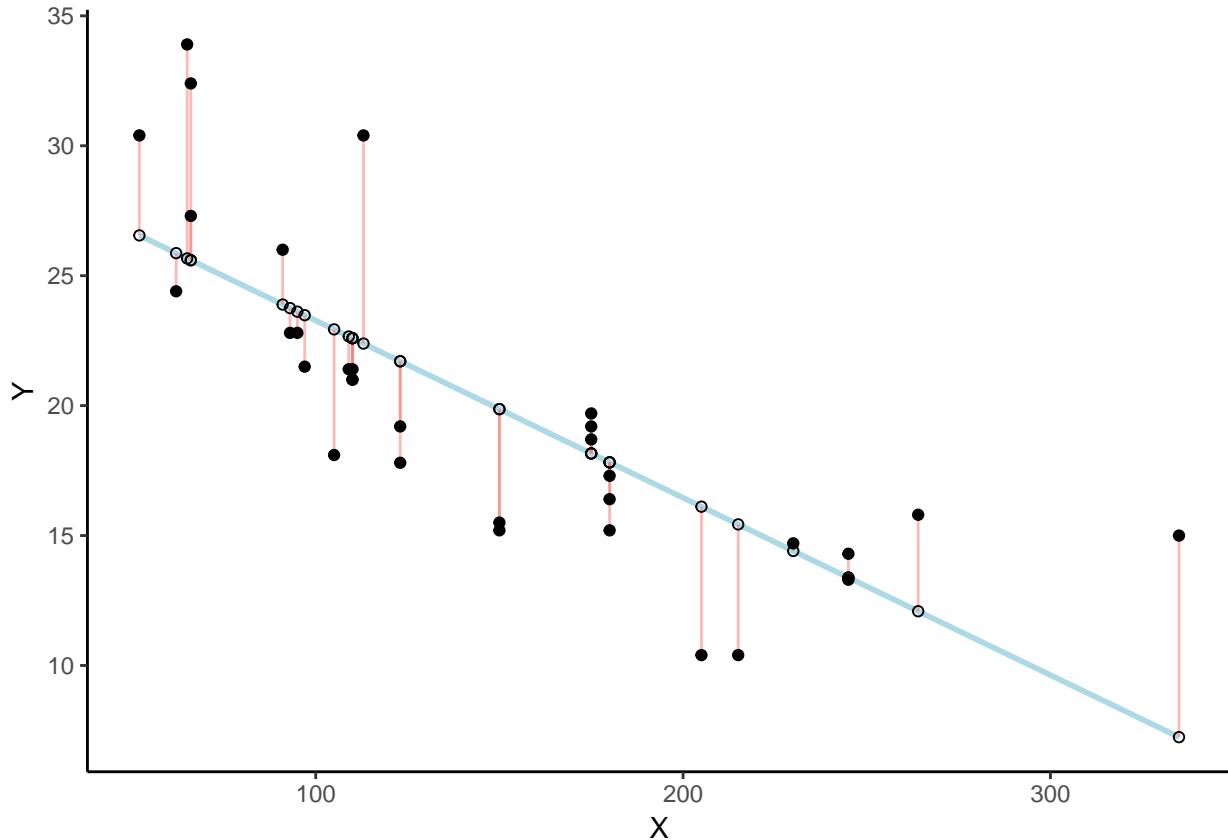


Figure 3.6: Black dots represent data points. The blue line is the best fit regression line. The white dots are represent the predicted location of each black dot. The red lines show the error between each black dot and the regression line. The blue line is the best fit line because it minimizes the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

Animation not available in .pdf version

Figure 3.7: The blue line is the best fit regression line explaining the co-variation among the black dots. The black line moves up and down showing alternative lines that could be drawn. The red lines show the amount of error between each data point and the black line. The total amount of error is depicted by the shaded grey area. The size of the grey area expands as the black line moves away from the best fit line, and shrinks to a minimum as the black line moves toward the best fit line

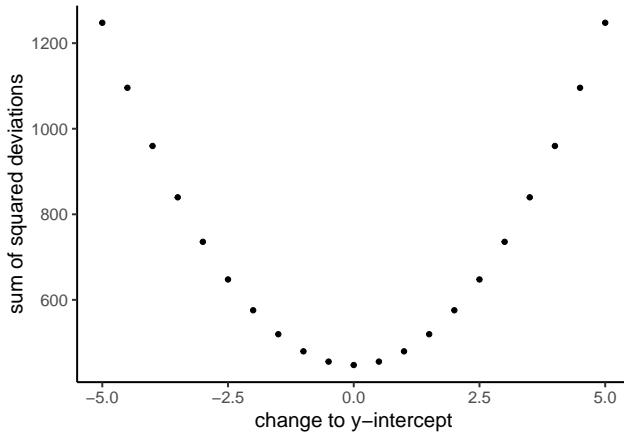


Figure 3.8: A plot of the sum of the squared deviations for different lines moving up and down, through the best fit line. The best fit line occurs at the position that minimizes the sum of the squared deviations

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

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

3.5.2 Lines

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

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

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

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

$$y = \text{slope} * x + \text{y-intercept}$$

The slope is the slant of the line, and the y-intercept is where the line crosses the y-axis. Let's look at some lines:

So there is two lines. The formula for the blue line is $y = 1 * x + 5$. Let's talk about that. When $x = 0$, where is the blue line on the y-axis? It's at five. That happens because 1 times 0 is 0, and then we just have

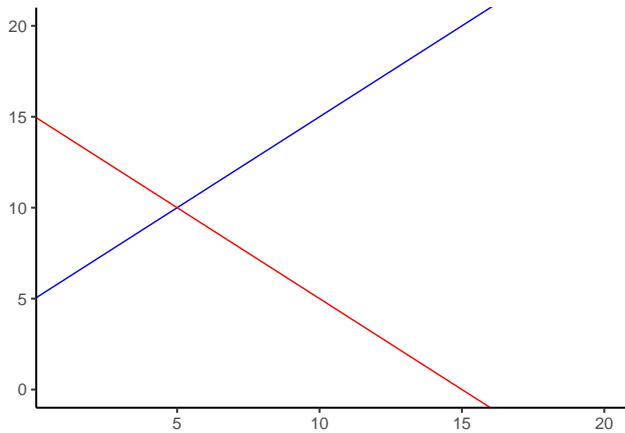


Figure 3.9: Two different lines with different y-intercepts (where the line crosses the y-axis), and different slopes. A positive slope makes the line go up from left to right. A negative slope makes the line go down from left to right.

the five left over. How about when $x = 5$? In that case $y = 10$. You just need the plug in the numbers to the formula, like this:

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

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

That's lines.

3.5.3 Computing the best fit line

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

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

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

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

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

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

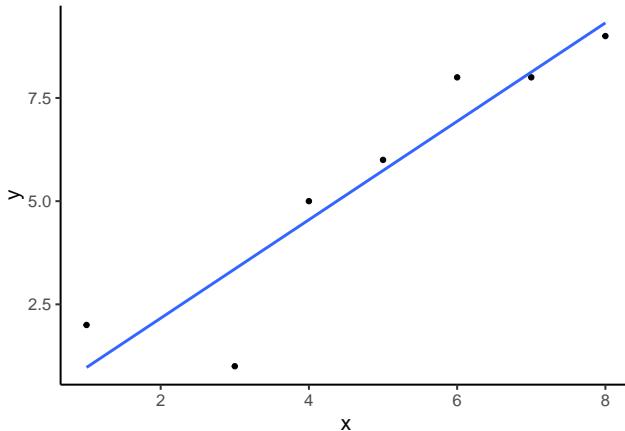


Figure 3.10: An example regression line going through a few data points in a scatterplot

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

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

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

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

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

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

```
## (Intercept)      x
## -0.2213115   1.1926230
```

3.6 Interpreting Correlations

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

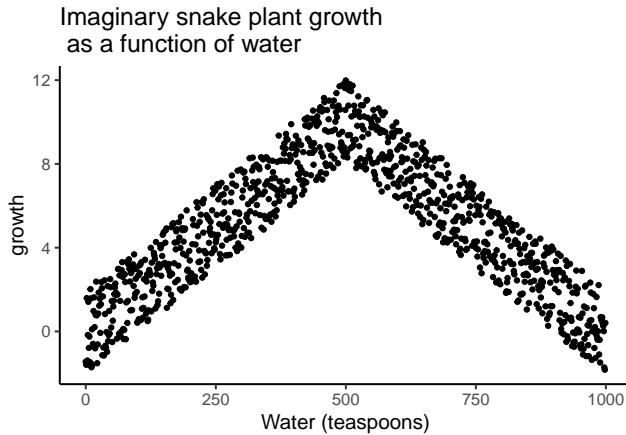


Figure 3.11: Illustration of a possible relationship between amount of water and snake plant growth. Growth goes up with water, but eventually goes back down as too much water makes snake plants die.

3.6.1 Correlation does not equal causation

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

Like most plants, snake plants need some water to stay alive. However, they also need just the right amount of water. Imagine an experiment where 1000 snake plants were grown in a house. Each snake plant is given a different amount of water per day, from zero teaspoons of water per day to 1000 teaspoons of water per day. We will assume that water is part of the causal process that allows snake plants to grow. The amount of water given to each snake plant per day can also be one of our measures. Imagine further that every week the experimenter measures snake plant growth, which will be the second measurement. Now, can you imagine for yourself what a scatter plot of weekly snake plant growth by tablespoons of water would look like?

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

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

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

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

```
## [1] -0.00389762
```

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

This example illustrates some conundrums in interpreting correlations. We already know that water is needed for plants to grow, so we are rightly expecting there to be a relationship between our measure of amount of water and plant growth. If we look at the first half of the data we see a positive correlation, if we look at the last half of the data we see a negative correlation, and if we look at all of the data we see no correlation. Yikes. So, even when there is a causal connection between two measures, we won't necessarily obtain clear evidence of the connection just by computing a correlation coefficient.

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

3.6.1.2 Confounding variable, or Third variable problem

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

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

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

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

3.6.2 Correlation and Random chance

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

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

causally influence the balls chosen by the second participant. They are on the other side of the world. We should assume that the balls will be chosen by chance alone.

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

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

In this one case, if we computed Pearson's r , we would find that $r = 0.3532628$. But, we already know that this value does not tell us anything about the relationship between the balls chosen in the north and south pole. We know that relationship should be completely random, because that is how we set up the game.

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

3.6.2.1 Monte-carlo simulation of random correlations

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

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

```

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

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

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

```

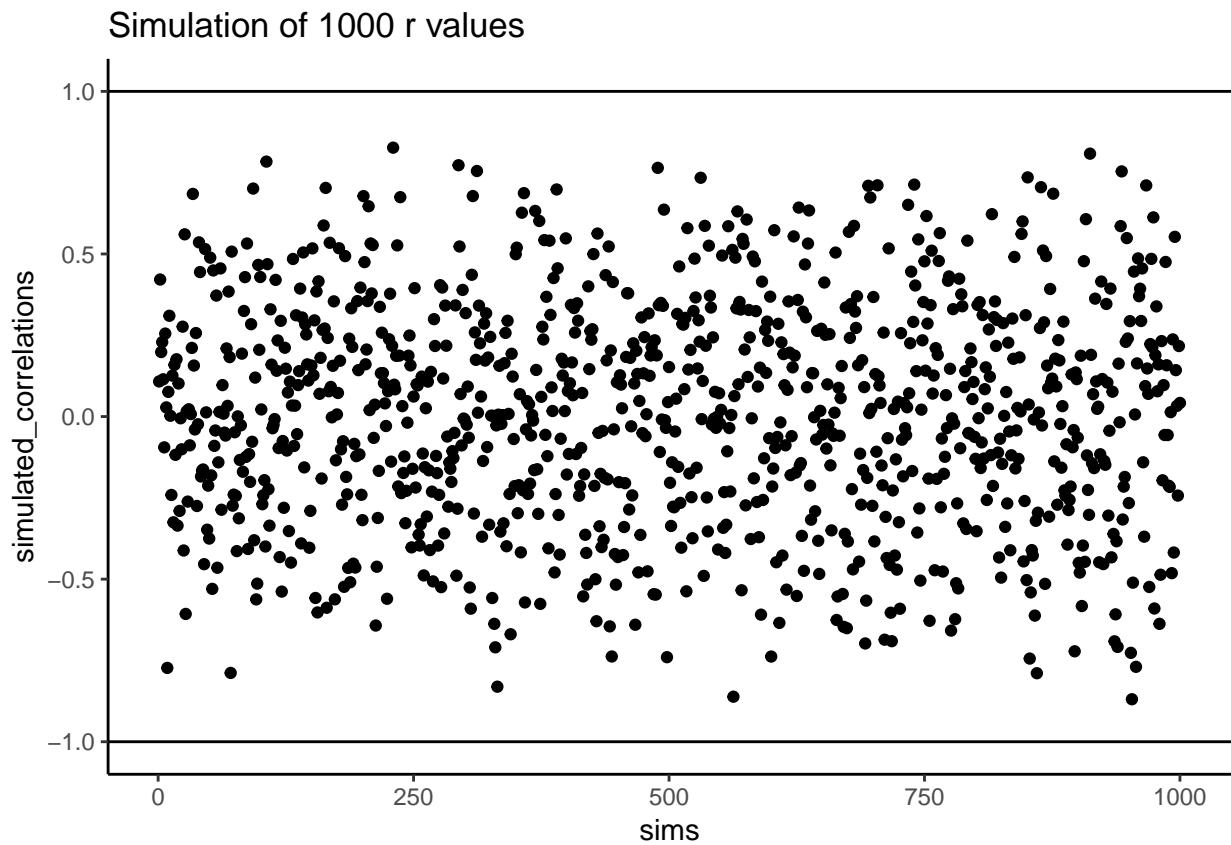


Figure 3.12: Another figure showing a range of r-values that can be obtained by chance

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

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

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

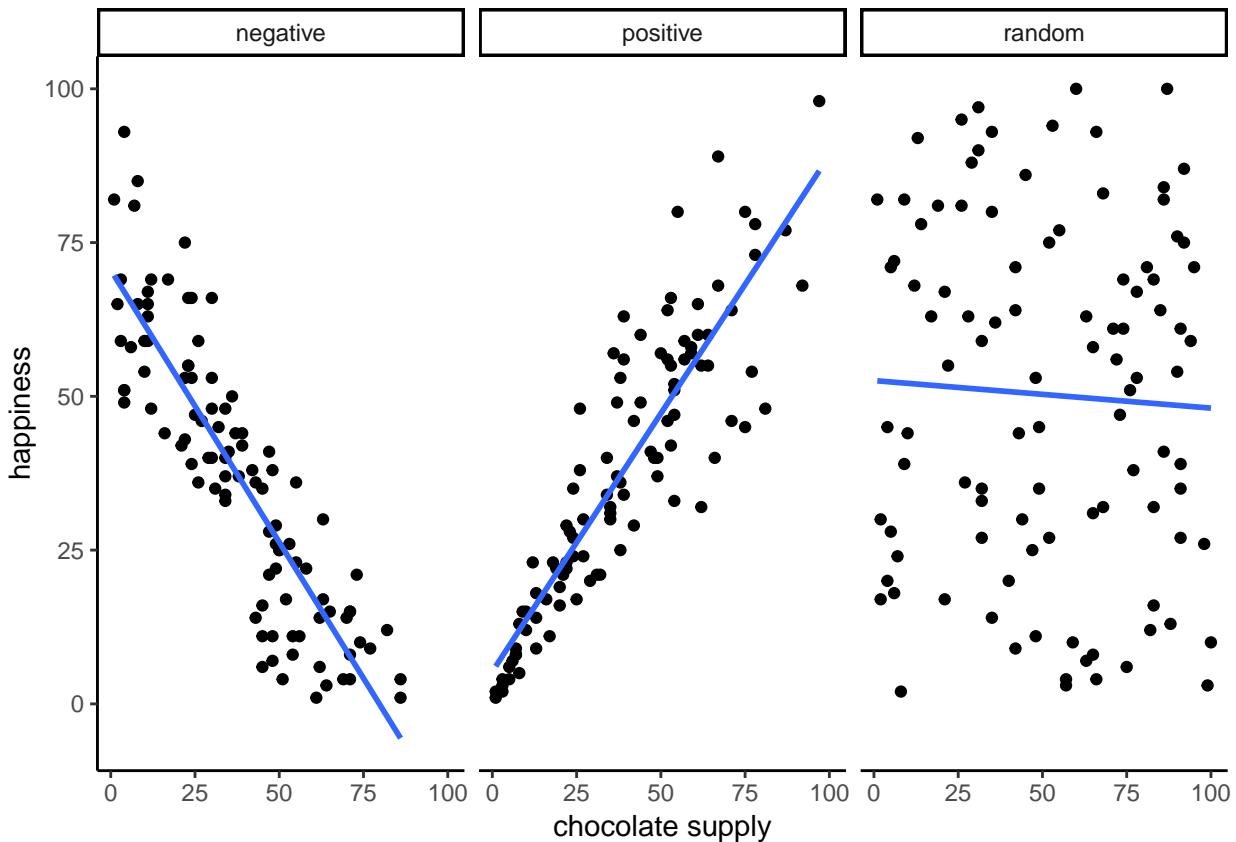


Figure 3.13: A reminder of what positive, negative, and zero correlation looks like

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

Animation not available in .pdf version

Figure 3.14: Completely random data points drawn from a uniform distribution with a small sample-size of 10. The blue line twirls around sometimes showing large correlations that are produced by chance

(line is more flat).

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

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

Histogram of simulated_correlations

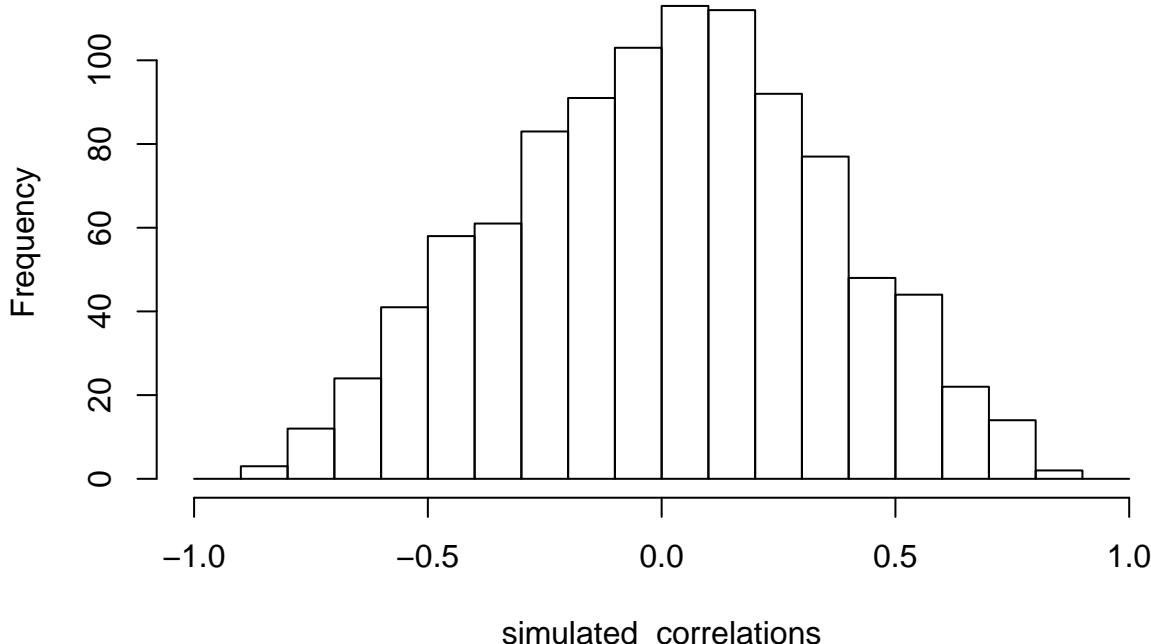


Figure 3.15: A histogram showing the frequency distribution of r -values for completely random values between an X and Y variable (sample-size=10). A full range of r -values can be obtained by chance alone. Larger r -values are less common than smaller r -values

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

occur as often as small correlations by chance alone.

You can think of this histogram as the window of chance. It shows what chance often does, and what it often does not do. If you found a correlation under these very same circumstances (e.g., measured the correlation between two sets of 10 random numbers), then you could consult this window. What should you ask the window? How about, could my observed correlation (the one that you found in your data) have come from this window. Let's say you found a correlation of $r = .1$. Could a $.1$ have come from the histogram? Well, look at the histogram around where the $.1$ mark on the x-axis is. Is there a big bar there? If so, this means that chance produces this value fairly often. You might be comfortable with the inference: Yes, this $.1$ could have been produced by chance, because it is well inside the window of chance. How about $r = .5$? The bar is much smaller here, you might think, "well, I can see that chance does produce $.5$ some times, so chance could have produced my $.5$. Did it? Maybe, maybe not, not sure". Here, your confidence in a strong inference about the role of chance might start getting a bit shakier.

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

3.6.2.2 Increasing sample-size decreases opportunity for spurious correlation

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

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

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

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

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

One take home here, is that increasing sample-size narrows the window of chance. So, for example, if you ran a study involving 1000 samples of two measures, and you found a correlation of $.5$, then you can clearly see in the bottom right histogram that $.5$ does not occur very often by chance alone. In fact, there is no bar, because it didn't happen even once in the simulation. As a result, when you have a large sample size like $n = 1000$, you might be more confident that your observed correlation (say of $.5$) was not a spurious correlation. If chance is not producing your result, then something else is.

Finally, notice how your confidence about whether or not chance is mucking about with your results depends on your sample size. If you only obtained 10 samples per measurement, and found $r = .5$, you should not

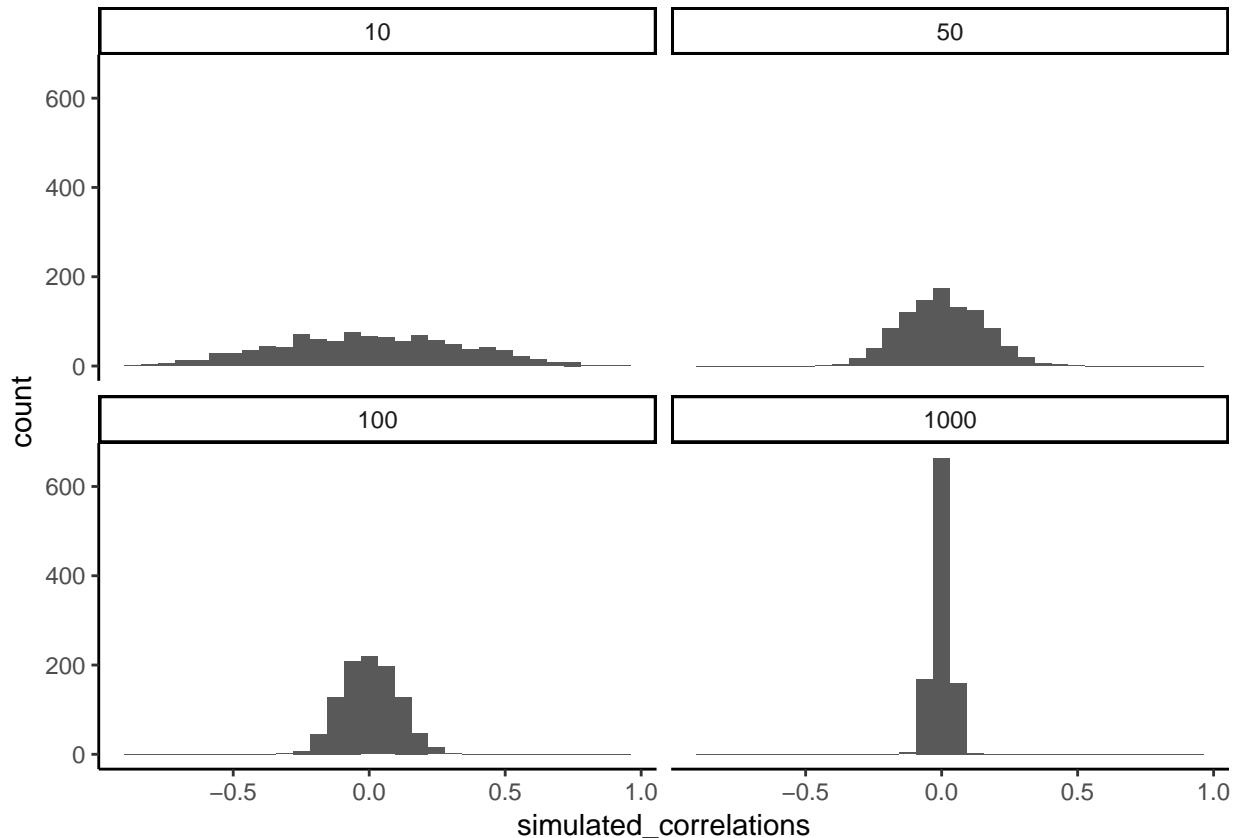


Figure 3.16: Four histograms showing the frequency distributions of r -values between completely random X and Y variables as a function of sample-size. The width of the distributions shrink as sample-size increases. Smaller sample-sizes are more likely to produce a wider range of r -values by chance. Larger sample-sizes always produce a narrow range of small r -values

Animation not available in .pdf version

Figure 3.17: Animation of how correlation behaves for completely random X and Y variables as a function of sample size. The best fit line is not very stable for small sample-sizes, but becomes more reliably flat as sample-size increases

be as confident that your correlation reflects a real relationship. Instead, you can see that r 's of .5 happen fairly often by chance alone.

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

3.6.3 Some more movies

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

3.6.3.1 Watching how correlation behaves when there is no correlation

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

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

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

In the above example, we sampled numbers random numbers from a uniform distribution. Many examples of real-world data will come from a normal or approximately normal distribution. We can repeat the above, but sample random numbers from the same normal distribution. There will still be zero actual correlation

Animation not available in .pdf version

Figure 3.18: Animation of correlation for random values sampled from a normal distribution, rather than a uniform distribution

Animation not available in .pdf version

Figure 3.19: How correlation behaves as a function of sample-size when there is a true correlation between X and Y variables

between the X and Y variables, because everything is sampled randomly. But, we still see the same behavior as above. The computed correlation for small sample-sizes fluctuate wildly, and large sample sizes do not.

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

3.6.3.2 Watching correlations behave when there really is a correlation

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

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

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

3.7 Summary

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

1. We can measure relationships in data using things like correlation
2. The correlations we measure can be produced by numerous things, so they are hard to interpret
3. Correlations can be produced by chance, so have the potential to be completely meaningless.

4. However, we can create a model of exactly what chance can do. The model tells us whether chance is more or less likely to produce correlations of different sizes
5. We can use the chance model to help us make decisions about our own data. We can compare the correlation we found in our data to the model, then ask whether or not chance could have or was likely to have produced our results.

Chapter 4

Probability, Sampling, and Estimation

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

Sections 4.1 & 4.9 - Adapted text by Danielle Navarro Section 4.10 - 4.11 & 4.13 - Mix of Matthew Crump & Danielle Navarro Section 4.12-4.13 - Adapted text by Danielle Navarro

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

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

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

This kind of remark is entirely unremarkable in the papers or in everyday life, but let's have a think about what it entails. A polling company has conducted a survey, usually a pretty big one because they can afford it. I'm too lazy to track down the original survey, so let's just imagine that they called 1000 voters at random, and 230 (23%) of those claimed that they intended to vote for the party. For the 2010 Federal election, the Australian Electoral Commission reported 4,610,795 enrolled voters in New South Wales; so the opinions of the remaining 4,609,795 voters (about 99.98% of voters) remain unknown to us. Even assuming that no-one lied to the polling company the only thing we can say with 100% confidence is that the true primary vote is somewhere between 230/4610795 (about 0.005%) and 4610025/4610795 (about 99.83%). So, on what basis is it legitimate for the polling company, the newspaper, and the readership to conclude that the ALP primary vote is only about 23%?

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

Inferential statistics provides the tools that we need to answer these sorts of questions, and since these kinds of questions lie at the heart of the scientific enterprise, they take up the lions share of every introductory course on statistics and research methods. However, our tools for making statistical inferences are 1) built on top of **probability theory**, and 2) require an understanding of how samples behave when you take them from distributions (defined by probability theory...). So, this chapter has two main parts. A brief introduction to probability theory, and an introduction to sampling from distributions.

4.1 How are probability and statistics different?

Before we start talking about probability theory, it's helpful to spend a moment thinking about the relationship between probability and statistics. The two disciplines are closely related but they're not identical. Probability theory is "the doctrine of chances". It's a branch of mathematics that tells you how often different kinds of events will happen. For example, all of these questions are things you can answer using probability theory:

- What are the chances of a fair coin coming up heads 10 times in a row?
- If I roll two six sided dice, how likely is it that I'll roll two sixes?
- How likely is it that five cards drawn from a perfectly shuffled deck will all be hearts?
- What are the chances that I'll win the lottery?

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

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

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

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

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

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

H H H H H H H H H H

and what I'm trying to do is work out which "model of the world" I should put my trust in. If the coin is fair, then the model I should adopt is one that says that the probability of heads is 0.5; that is, $P(\text{heads}) = 0.5$.

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

4.2 What does probability mean?

Let's start with the first of these questions. What is "probability"? It might seem surprising to you, but while statisticians and mathematicians (mostly) agree on what the **rules** of probability are, there's much less of a consensus on what the word really **means**. It seems weird because we're all very comfortable using words like "chance", "likely", "possible" and "probable", and it doesn't seem like it should be a very difficult question to answer. If you had to explain "probability" to a five year old, you could do a pretty good job. But if you've ever had that experience in real life, you might walk away from the conversation feeling like you didn't quite get it right, and that (like many everyday concepts) it turns out that you don't **really** know what it's all about.

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

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

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

4.2.1 The frequentist view

The first of the two major approaches to probability, and the more dominant one in statistics, is referred to as the *frequentist view*, and it defines probability as a *long-run frequency*. Suppose we were to try flipping a fair coin, over and over again. By definition, this is a coin that has $P(H) = 0.5$. What might we observe? One possibility is that the first 20 flips might look like this:

T, H, H, H, H, T, T, H, H, H, H, T, H, H, T, T, T, T, H

In this case 11 of these 20 coin flips (55%) came up heads. Now suppose that I'd been keeping a running tally of the number of heads (which I'll call N_H) that I've seen, across the first N flips, and calculate the proportion of heads N_H/N every time. Here's what I'd get (I did literally flip coins to produce this!):

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

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

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

The frequentist definition of probability has some desirable characteristics. First, it is objective: the probability of an event is **necessarily** grounded in the world. The only way that probability statements can make sense is if they refer to (a sequence of) events that occur in the physical universe. Second, it is unambiguous: any two people watching the same sequence of events unfold, trying to calculate the probability of an event, must inevitably come up with the same answer.

However, it also has undesirable characteristics. Infinite sequences don’t exist in the physical world. Suppose you picked up a coin from your pocket and started to flip it. Every time it lands, it impacts on the ground. Each impact wears the coin down a bit; eventually, the coin will be destroyed. So, one might ask whether it really makes sense to pretend that an “infinite” sequence of coin flips is even a meaningful concept, or an objective one. We can’t say that an “infinite sequence” of events is a real thing in the physical universe, because the physical universe doesn’t allow infinite anything.

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

4.2.2 The Bayesian view

The **Bayesian view** of probability is often called the subjectivist view, and it is a minority view among statisticians, but one that has been steadily gaining traction for the last several decades. There are many flavours of Bayesianism, making hard to say exactly what “the” Bayesian view is. The most common way of thinking about subjective probability is to define the probability of an event as the **degree of belief** that an intelligent and rational agent assigns to that truth of that event. From that perspective, probabilities don’t exist in the world, but rather in the thoughts and assumptions of people and other intelligent beings.

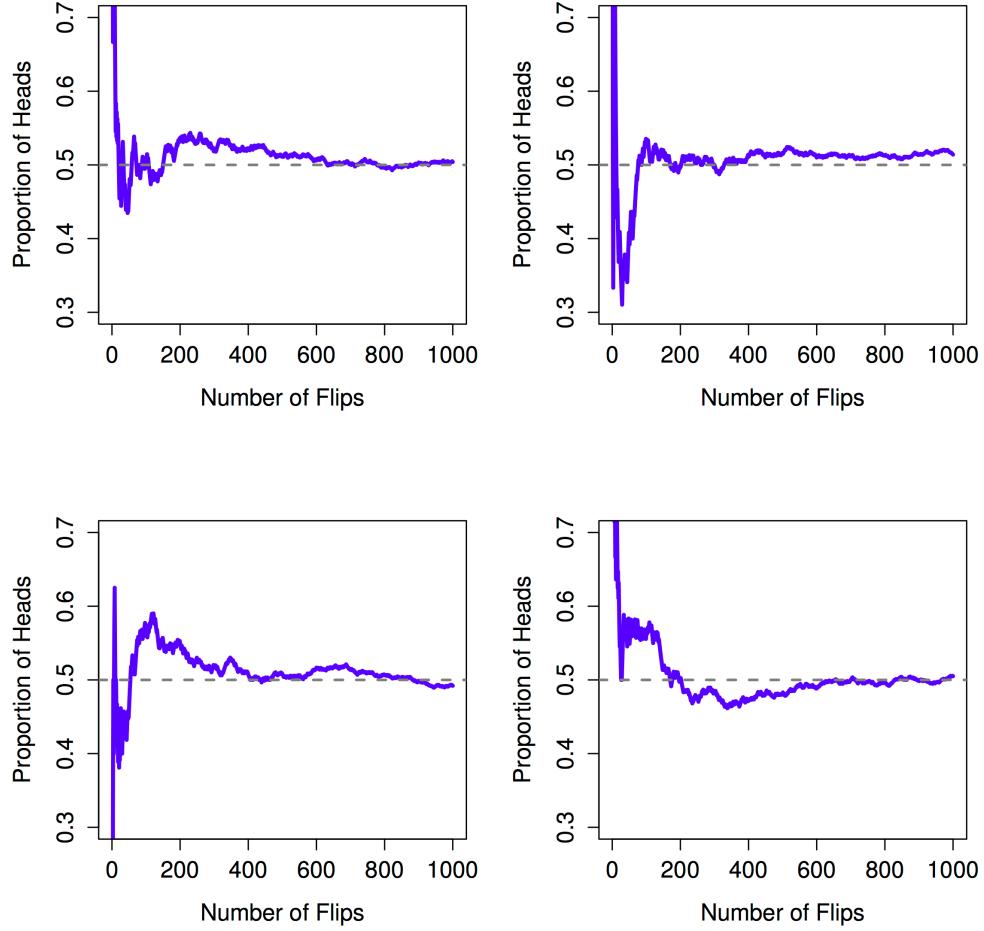


Figure 4.1: An illustration of how frequentist probability works. If you flip a fair coin over and over again, the proportion of heads that you've seen eventually settles down, and converges to the true probability of 0.5. Each panel shows four different simulated experiments: in each case, we pretend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

However, in order for this approach to work, we need some way of operationalising “degree of belief”. One way that you can do this is to formalise it in terms of “rational gambling”, though there are many other ways. Suppose that I believe that there’s a 60% probability of rain tomorrow. If someone offers me a bet: if it rains tomorrow, then I win \$5, but if it doesn’t rain then I lose \$5. Clearly, from my perspective, this is a pretty good bet. On the other hand, if I think that the probability of rain is only 40%, then it’s a bad bet to take. Thus, we can operationalise the notion of a “subjective probability” in terms of what bets I’m willing to accept.

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

4.2.3 What’s the difference? And who is right?

Now that you’ve seen each of these two views independently, it’s useful to make sure you can compare the two. Go back to the hypothetical robot soccer game at the start of the section. What do you think a frequentist and a Bayesian would say about these three statements? Which statement would a frequentist say is the correct definition of probability? Which one would a Bayesian do? Would some of these statements be meaningless to a frequentist or a Bayesian? If you’ve understood the two perspectives, you should have some sense of how to answer those questions.

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

For the most part, I’m a pragmatist so I’ll use any statistical method that I trust. As it turns out, that makes me prefer Bayesian methods, for reasons I’ll explain towards the end of the book, but I’m not fundamentally opposed to frequentist methods. Not everyone is quite so relaxed. For instance, consider Sir Ronald Fisher, one of the towering figures of 20th century statistics and a vehement opponent to all things Bayesian, whose paper on the mathematical foundations of statistics referred to Bayesian probability as “an impenetrable jungle [that] arrests progress towards precision of statistical concepts” Fisher (1922, p. 311). Or the psychologist Paul Meehl, who suggests that relying on frequentist methods could turn you into “a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring” Meehl (1967, p. 114). The history of statistics, as you might gather, is not devoid of entertainment.

4.3 Basic probability theory

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

4.3.1 Introducing probability distributions

One of the disturbing truths about my life is that I only own 5 pairs of pants: three pairs of jeans, the bottom half of a suit, and a pair of tracksuit pants. Even sadder, I've given them names: I call them X_1 , X_2 , X_3 , X_4 and X_5 . I really do: that's why they call me Mister Imaginative. Now, on any given day, I pick out exactly one of pair of pants to wear. Not even I'm so stupid as to try to wear two pairs of pants, and thanks to years of training I never go outside without wearing pants anymore. If I were to describe this situation using the language of probability theory, I would refer to each pair of pants (i.e., each X) as an *elementary event*. The key characteristic of elementary events is that every time we make an observation (e.g., every time I put on a pair of pants), then the outcome will be one and only one of these events. Like I said, these days I always wear exactly one pair of pants, so my pants satisfy this constraint. Similarly, the set of all possible events is called a *sample space*. Granted, some people would call it a "wardrobe", but that's because they're refusing to think about my pants in probabilistic terms. Sad.

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

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

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

Each of the events has a probability that lies between 0 and 1, and if we add up the probability of all events, they sum to 1. Awesome. We can even draw a nice bar graph to visualise this distribution, as shown in Figure 4.2. And at this point, we've all achieved something. You've learned what a probability distribution is, and I've finally managed to find a way to create a graph that focuses entirely on my pants. Everyone wins!

The only other thing that I need to point out is that probability theory allows you to talk about *non elementary events* as well as elementary ones. The easiest way to illustrate the concept is with an example. In the pants example, it's perfectly legitimate to refer to the probability that I wear jeans. In this scenario,

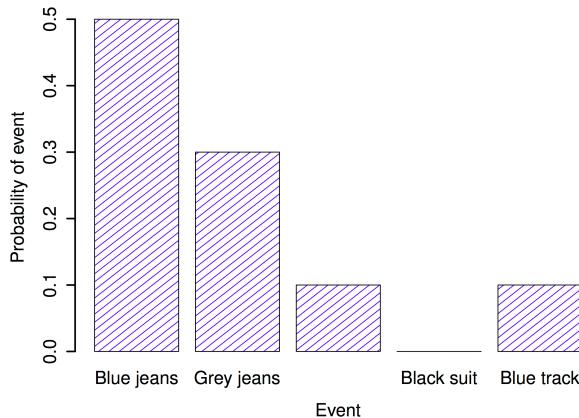


Figure 4.2: A visual depiction of the pants probability distribution. There are five elementary events, corresponding to the five pairs of pants that I own. Each event has some probability of occurring: this probability is a number between 0 to 1. The sum of these probabilities is 1

the “Dan wears jeans” event said to have happened as long as the elementary event that actually did occur is one of the appropriate ones; in this case “blue jeans”, “black jeans” or “grey jeans”. In mathematical terms, we defined the “jeans” event E to correspond to the set of elementary events (X_1, X_2, X_3) . If any of these elementary events occurs, then E is also said to have occurred. Having decided to write down the definition of the E this way, it’s pretty straightforward to state what the probability $P(E)$ is: we just add everything up. In this particular case

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

and, since the probabilities of blue, grey and black jeans respectively are .5, .3 and .1, the probability that I wear jeans is equal to .9.

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

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

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

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

$$X = (x_1, x_2, x_3, x_4, x_5)$$

$$A = (x_1, x_2, x_3)$$

mathematically denoted as $\neg A$) is to say that $\neg A$ consists of all elementary events that don't belong to A . In the case of the pants distribution it means that $\neg A = (x_4, x_5)$, or, to say it in English: “not jeans” consists of all pairs of pants that aren't jeans (i.e., the black suit and the blue tracksuit). Consequently, every single elementary event belongs to either A or $\neg A$, but not both. Okay, so now let's rearrange our statement above:

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

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

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

so therefore

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

Excellent. It all seems to work.

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

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

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

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

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

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

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

Oh yeah baby. Mathematics at its finest.

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

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

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

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

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

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

and therefore

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

Done.

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

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

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

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

In mathematical terms, we say that

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

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

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

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

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

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

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

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

4.4 The binomial distribution

As you might imagine, probability distributions vary enormously, and there's an enormous range of distributions out there. However, they aren't all equally important. In fact, the vast majority of the content in this book relies on one of five distributions: the binomial distribution, the normal distribution, the t distribution, the χ^2 ("chi-square") distribution and the F distribution. Given this, what I'll do over the next few sections is provide a brief introduction to all five of these, paying special attention to the binomial and the normal. I'll start with the binomial distribution, since it's the simplest of the five.

4.4.1 Introducing the binomial

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

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

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

and we're interested in the special case where $X = 4$, $\theta = .167$ and $N = 20$. There's only one more piece of notation I want to refer to before moving on to discuss the solution to the problem. If I want to say that X

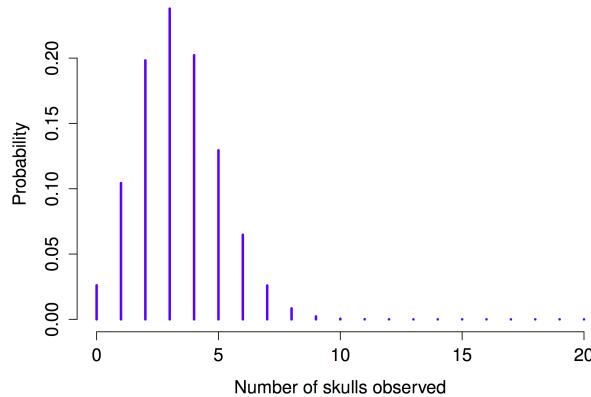


Figure 4.3: The binomial distribution with size parameter of $N = 20$ and an underlying success probability of $1/6$. Each vertical bar depicts the probability of one specific outcome (i.e., one possible value of X). Because this is a probability distribution, each of the probabilities must be a number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

is generated randomly from a binomial distribution with parameters θ and N , the notation I would use is as follows:

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

Yeah, yeah. I know what you’re thinking: notation, notation, notation. Really, who cares? Very few readers of this book are here for the notation, so I should probably move on and talk about how to use the binomial distribution. I’ve included the formula for the binomial distribution in Table [tab:distformulas], since some readers may want to play with it themselves, but since most people probably don’t care that much and because we don’t need the formula in this book, I won’t talk about it in any detail. Instead, I just want to show you what the binomial distribution looks like. To that end, Figure 4.3 plots the binomial probabilities for all possible values of X for our dice rolling experiment, from $X = 0$ (no skulls) all the way up to $X = 20$ (all skulls). Note that this is basically a bar chart, and is no different to the “pants probability” plot I drew in Figure 4.2. On the horizontal axis we have all the possible events, and on the vertical axis we can read off the probability of each of those events. So, the probability of rolling 4 skulls out of 20 times is about 0.20 (the actual answer is 0.2022036, as we’ll see in a moment). In other words, you’d expect that to happen about 20% of the times you repeated this experiment.

4.4.2 Working with the binomial distribution in R

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

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

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

```
dbinom( x = 4, size = 20, prob = 1/6 )
```

```
## [1] 0.2022036
```

To give you a feel for how the binomial distribution changes when we alter the values of θ and N , let's suppose that instead of rolling dice, I'm actually flipping coins. This time around, my experiment involves flipping a fair coin repeatedly, and the outcome that I'm interested in is the number of heads that I observe. In this scenario, the success probability is now $\theta = 1/2$. Suppose I were to flip the coin $N = 20$ times. In this example, I've changed the success probability, but kept the size of the experiment the same. What does this do to our binomial distribution?

Well, as Figure 4.4a shows, the main effect of this is to shift the whole distribution, as you'd expect. Okay, what if we flipped a coin $N = 100$ times? Well, in that case, we get Figure 4.4b. The distribution stays roughly in the middle, but there's a bit more variability in the possible outcomes.

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

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

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

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

```
pbinom( q= 4, size = 20, prob = 1/6)
```

```
## [1] 0.7687492
```

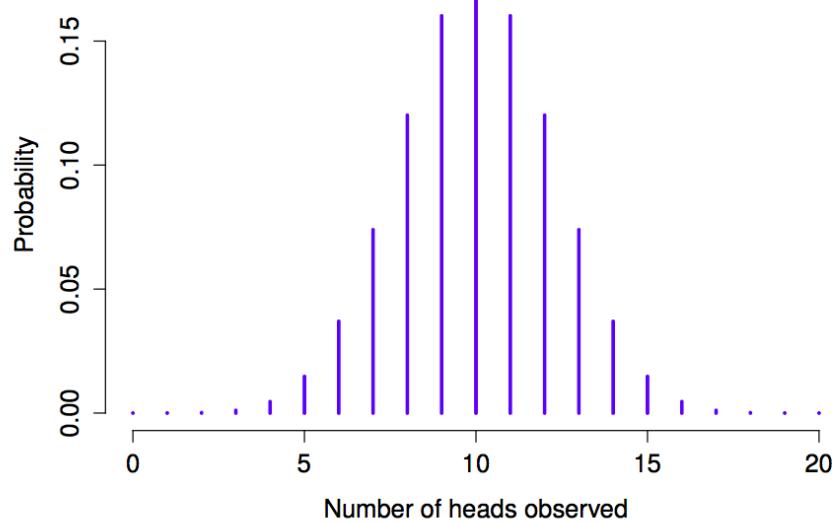
In other words, there is a 76.9% chance that I will roll 4 or fewer skulls. Or, to put it another way, R is telling us that a value of 4 is actually the 76.9th percentile of this binomial distribution.

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

```
qbinom( p = 0.75, size = 20, prob = 1/6 )
```

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

(a)



(b)

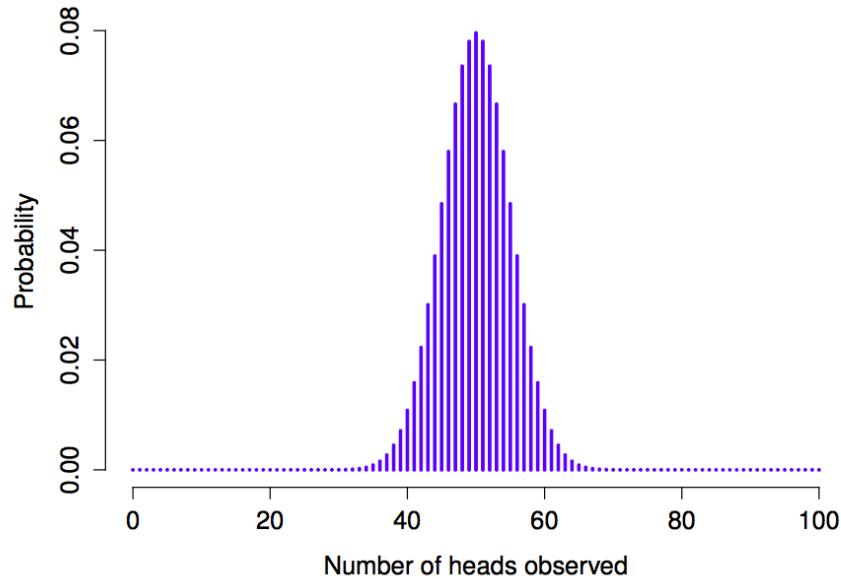


Figure 4.4: Two binomial distributions, involving a scenario in which I'm flipping a fair coin, so the underlying success probability is $1/2$. In panel (a), we assume I'm flipping the coin $N = 20$ times. In panel (b) we assume that the coin is flipped $N = 100$ times.

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

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

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

```
rbinom( n = 100, size = 20, prob = 1/6 )
```

```
##   [1] 5 5 3 3 5 2 6 0 3 5 4 3 3 3 3 4 3 4 2 3 1 2 3
## [24] 4 2 4 2 5 4 1 1 4 5 3 5 4 4 2 4 5 2 2 4 3 7 5
## [47] 3 0 3 5 5 4 3 4 3 3 2 3 4 2 4 3 5 4 5 1 2 3 1
## [70] 4 4 0 4 4 4 2 10 3 3 2 1 4 1 0 3 3 6 4 3 3 5 4
## [93] 2 3 3 2 3 4 3 4
```

As you can see, these numbers are pretty much what you'd expect given the distribution shown in Figure 4.3. Most of the time I roll somewhere between 1 to 5 skulls. There are a lot of subtleties associated with random number generation using a computer, but for the purposes of this book we don't need to worry too much about them.

4.5 The normal distribution

While the binomial distribution is conceptually the simplest distribution to understand, it's not the most important one. That particular honour goes to the *normal distribution*, which is also referred to as "the bell curve" or a "Gaussian distribution".

A normal distribution is described using two parameters, the mean of the distribution μ and the standard deviation of the distribution σ . The notation that we sometimes use to say that a variable X is normally distributed is as follows:

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

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

The formula is important enough that everyone who learns statistics should at least look at it, but since this is an introductory text I don't want to focus on it to much. Instead, we look at how R can be used to work with normal distributions. The R functions for the normal distribution are `dnorm()`, `pnorm()`, `qnorm()` and `rnorm()`. However, they behave in pretty much exactly the same way as the corresponding functions for the binomial distribution, so there's not a lot that you need to know. The only thing that I should point out is

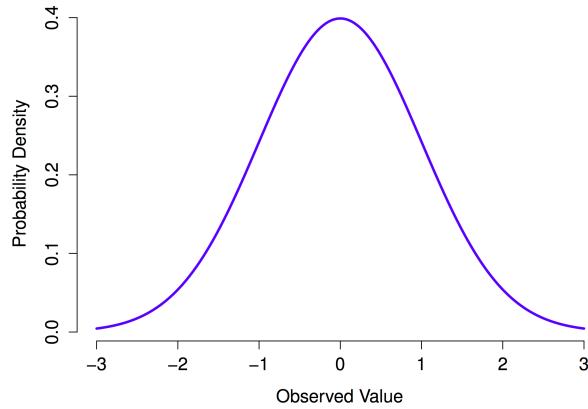


Figure 4.5: The normal distribution with mean = 0 and standard deviation = 1. The x-axis corresponds to the value of some variable, and the y-axis tells us something about how likely we are to observe that value. However, notice that the y-axis is labelled Probability Density and not Probability. There is a subtle and somewhat frustrating characteristic of continuous distributions that makes the y axis behave a bit oddly: the height of the curve here isn't actually the probability of observing a particular x value. On the other hand, it is true that the heights of the curve tells you which x values are more likely (the higher ones!).

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

Figure 4.6: Formula for the normal distribution

Animation not available in .pdf version

Figure 4.7: A normal distribution with a moving mean

Animation not available in .pdf version

Figure 4.8: A normal distribution with a shifting sd

that the argument names for the parameters are *mean* and *sd*. In pretty much every other respect, there's nothing else to add.

Instead of focusing on the maths, let's try to get a sense for what it means for a variable to be normally distributed. To that end, have a look at Figure 4.5, which plots a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. You can see where the name "bell curve" comes from: it looks a bit like a bell. Notice that, unlike the plots that I drew to illustrate the binomial distribution, the picture of the normal distribution in Figure 4.5 shows a smooth curve instead of "histogram-like" bars. This isn't an arbitrary choice: the normal distribution is continuous, whereas the binomial is discrete. For instance, in the die rolling example from the last section, it was possible to get 3 skulls or 4 skulls, but impossible to get 3.9 skulls.

With this in mind, let's see if we can't get an intuition for how the normal distribution works. Firstly, let's have a look at what happens when we play around with the parameters of the distribution. One parameter we can change is the mean. This will shift the distribution to the right or left. The animation below shows a normal distribution with mean = 0, moving up and down from mean = 0 to mean = 5. Note, when you change the mean the whole shape of the distribution does not change, it just shifts from left to right. In the animation the normal distribution bounces up and down a little, but that's just a quirk of the animation (plus it looks fund that way).

In contrast, if we increase the standard deviation while keeping the mean constant, the peak of the distribution stays in the same place, but the distribution gets wider. The next animation shows what happens when you start with a small standard deviation ($sd=0.5$), and move to larger and larger standard deviation (up to $sd=5$). As you can see, the distribution spreads out and becomes wider as the standard deviation increases.

Notice, though, that when we widen the distribution, the height of the peak shrinks. This has to happen: in the same way that the heights of the bars that we used to draw a discrete binomial distribution have to *sum* to 1, the total *area under the curve* for the normal distribution must equal 1. Before moving on, I want to point out one important characteristic of the normal distribution. Irrespective of what the actual mean and standard deviation are, 68.3% of the area falls within 1 standard deviation of the mean. Similarly, 95.4% of the distribution falls within 2 standard deviations of the mean, and 99.7% of the distribution is within 3 standard deviations.

4.5.1 Probability density

There's something I've been trying to hide throughout my discussion of the normal distribution, something that some introductory textbooks omit completely. They might be right to do so: this "thing" that I'm hiding is weird and counterintuitive even by the admittedly distorted standards that apply in statistics. Fortunately, it's not something that you need to understand at a deep level in order to do basic statistics:

rather, it's something that starts to become important later on when you move beyond the basics. So, if it doesn't make complete sense, don't worry: try to make sure that you follow the gist of it.

Throughout my discussion of the normal distribution, there's been one or two things that don't quite make sense. Perhaps you noticed that the y -axis in these figures is labelled "Probability Density" rather than density. Maybe you noticed that I used $p(X)$ instead of $P(X)$ when giving the formula for the normal distribution. Maybe you're wondering why R uses the "d" prefix for functions like `dnorm()`. And maybe, just maybe, you've been playing around with the `dnorm()` function, and you accidentally typed in a command like this:

```
dnorm( x = 1, mean = 1, sd = 0.1 )
```

```
## [1] 3.989423
```

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

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

Now think about what this implies when we talk about probabilities. Suppose that tomorrow's maximum temperature is sampled from a normal distribution with mean 23 and standard deviation 1. What's the probability that the temperature will be *exactly* 23 degrees? The answer is "zero", or possibly, "a number so close to zero that it might as well be zero". Why is this?

It's like trying to throw a dart at an infinitely small dart board: no matter how good your aim, you'll never hit it. In real life you'll never get a value of exactly 23. It'll always be something like 23.1 or 22.99998 or something. In other words, it's completely meaningless to talk about the probability that the temperature is exactly 23 degrees. However, in everyday language, if I told you that it was 23 degrees outside and it turned out to be 22.9998 degrees, you probably wouldn't call me a liar. Because in everyday language, "23 degrees" usually means something like "somewhere between 22.5 and 23.5 degrees". And while it doesn't feel very meaningful to ask about the probability that the temperature is exactly 23 degrees, it does seem sensible to ask about the probability that the temperature lies between 22.5 and 23.5, or between 20 and 30, or any other range of temperatures.

The point of this discussion is to make clear that, when we're talking about continuous distributions, it's not meaningful to talk about the probability of a specific value. However, what we *can* talk about is the **probability that the value lies within a particular range of values**. To find out the probability associated with a particular range, what you need to do is calculate the "area under the curve".

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

4.6 Other useful distributions

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

4.7 Summary of Probability

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

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

As you'd expect, this coverage is by no means exhaustive. Probability theory is a large branch of mathematics in its own right, entirely separate from its application to statistics and data analysis. As such, there are thousands of books written on the subject and universities generally offer multiple classes devoted entirely to probability theory. Even the "simpler" task of documenting standard probability distributions is a big topic. Fortunately for you, very little of this is necessary. You're unlikely to need to know dozens of statistical distributions when you go out and do real world data analysis, and you definitely won't need them for this book, but it never hurts to know that there's other possibilities out there.

Picking up on that last point, there's a sense in which this whole chapter is something of a digression. Many undergraduate psychology classes on statistics skim over this content very quickly (I know mine did), and even the more advanced classes will often "forget" to revisit the basic foundations of the field. Most academic psychologists would not know the difference between probability and density, and until recently very few would have been aware of the difference between Bayesian and frequentist probability. However, I think it's important to understand these things before moving onto the applications. For example, there are a lot of rules about what you're "allowed" to say when doing statistical inference, and many of these can seem arbitrary and weird. However, they start to make sense if you understand that there is this Bayesian/frequentist distinction.

4.8 Samples, populations and sampling

Remember, the role of descriptive statistics is to concisely summarize what we **do** know. In contrast, the purpose of inferential statistics is to "learn what we do not know from what we do". What kinds of things would we like to learn about? And how do we learn them? These are the questions that lie at the heart of inferential statistics, and they are traditionally divided into two "big ideas": estimation and hypothesis testing. The goal in this chapter is to introduce the first of these big ideas, estimation theory, but we'll talk about sampling theory first because estimation theory doesn't make sense until you understand sampling. So, this chapter divides into sampling theory, and how to make use of sampling theory to discuss how statisticians think about estimation. We have already done lots of sampling, so you are already familiar with some of the big ideas.

Sampling theory plays a huge role in specifying the assumptions upon which your statistical inferences rely. And in order to talk about "making inferences" the way statisticians think about it, we need to be a

bit more explicit about what it is that we're drawing inferences **from** (the sample) and what it is that we're drawing inferences **about** (the population).

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

4.8.1 Defining a population

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

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

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

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

4.8.2 Simple random samples

Irrespective of how we define the population, the critical point is that the sample is a subset of the population, and our goal is to use our knowledge of the sample to draw inferences about the properties of the population. The relationship between the two depends on the **procedure** by which the sample was selected. This procedure is referred to as a **sampling method**, and it is important to understand why it matters.

To keep things simple, imagine we have a bag containing 10 chips. Each chip has a unique letter printed on it, so we can distinguish between the 10 chips. The chips come in two colors, black and white.

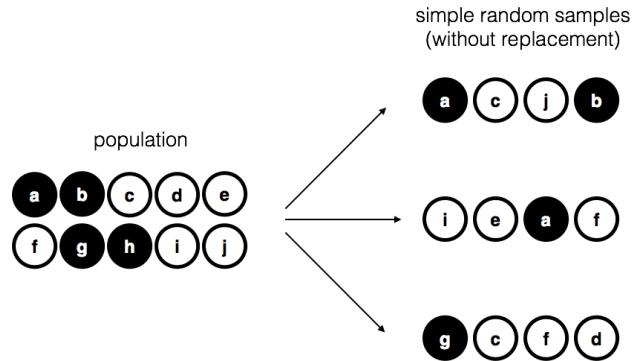


Figure 4.9: Simple random sampling without replacement from a finite population

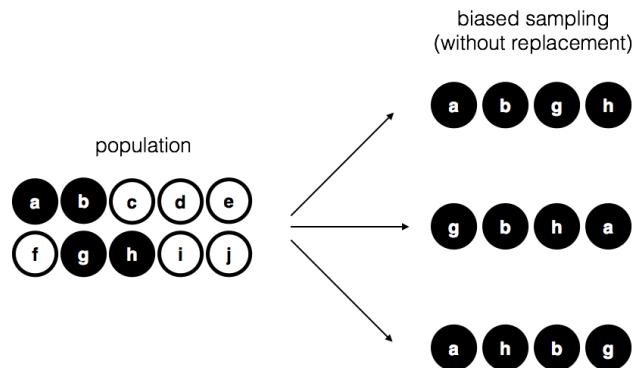


Figure 4.10: Biased sampling without replacement from a finite population

This set of chips is the population of interest, and it is depicted graphically on the left of Figure 4.9.

As you can see from looking at the picture, there are 4 black chips and 6 white chips, but of course in real life we wouldn't know that unless we looked in the bag. Now imagine you run the following "experiment": you shake up the bag, close your eyes, and pull out 4 chips without putting any of them back into the bag. First out comes the *a* chip (black), then the *c* chip (white), then *j* (white) and then finally *b* (black). If you wanted, you could then put all the chips back in the bag and repeat the experiment, as depicted on the right hand side of Figure 4.9. Each time you get different results, but the procedure is identical in each case. The fact that the same procedure can lead to different results each time, we refer to it as a **random** process. However, because we shook the bag before pulling any chips out, it seems reasonable to think that every chip has the same chance of being selected. A procedure in which every member of the population has the same chance of being selected is called a **simple random sample**. The fact that we did **not** put the chips back in the bag after pulling them out means that you can't observe the same thing twice, and in such cases the observations are said to have been sampled **without replacement**.

To help make sure you understand the importance of the sampling procedure, consider an alternative way in which the experiment could have been run. Suppose that my 5-year old son had opened the bag, and decided to pull out four black chips without putting any of them back in the bag. This **biased** sampling scheme is depicted in Figure 4.10.

Now consider the evidentiary value of seeing 4 black chips and 0 white chips. Clearly, it depends on the sampling scheme, does it not? If you know that the sampling scheme is biased to select only black chips, then a sample that consists of only black chips doesn't tell you very much about the population! For this reason, statisticians really like it when a data set can be considered a simple random sample, because it makes the data analysis **much** easier.

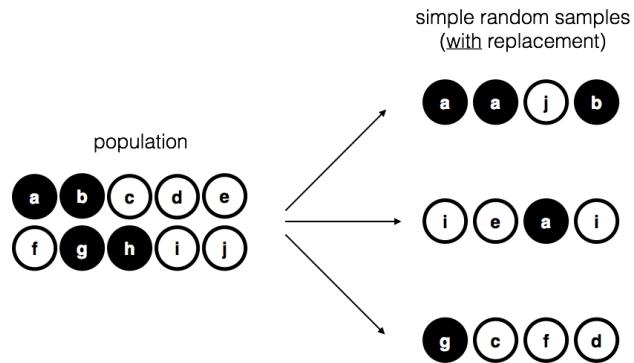


Figure 4.11: Simple random sampling with replacement from a finite population

A third procedure is worth mentioning. This time around we close our eyes, shake the bag, and pull out a chip. This time, however, we record the observation and then put the chip back in the bag. Again we close our eyes, shake the bag, and pull out a chip. We then repeat this procedure until we have 4 chips. Data sets generated in this way are still simple random samples, but because we put the chips back in the bag immediately after drawing them it is referred to as a sample **with replacement**. The difference between this situation and the first one is that it is possible to observe the same population member multiple times, as illustrated in Figure 4.11.

Most psychology experiments tend to be sampling without replacement, because the same person is not allowed to participate in the experiment twice. However, most statistical theory is based on the assumption that the data arise from a simple random sample **with** replacement. In real life, this very rarely matters. If the population of interest is large (e.g., has more than 10 entities!) the difference between sampling with- and without-replacement is too small to be concerned with. The difference between simple random samples and biased samples, on the other hand, is not such an easy thing to dismiss.

4.8.3 Most samples are not simple random samples

As you can see from looking at the list of possible populations that I showed above, it is almost impossible to obtain a simple random sample from most populations of interest. When I run experiments, I'd consider it a minor miracle if my participants turned out to be a random sampling of the undergraduate psychology students at Adelaide university, even though this is by far the narrowest population that I might want to generalize to. A thorough discussion of other types of sampling schemes is beyond the scope of this book, but to give you a sense of what's out there I'll list a few of the more important ones:

- **Stratified sampling.** Suppose your population is (or can be) divided into several different sub-populations, or **strata**. Perhaps you're running a study at several different sites, for example. Instead of trying to sample randomly from the population as a whole, you instead try to collect a separate random sample from each of the strata. Stratified sampling is sometimes easier to do than simple random sampling, especially when the population is already divided into the distinct strata. It can also be more efficient than simple random sampling, especially when some of the sub-populations are rare. For instance, when studying schizophrenia it would be much better to divide the population into two strata (schizophrenic and non-schizophrenic), and then sample an equal number of people from each group. If you selected people randomly, you would get so few schizophrenic people in the sample that your study would be useless. This specific kind of stratified sampling is referred to as **oversampling** because it makes a deliberate attempt to over-represent rare groups.
- **Snowball sampling** is a technique that is especially useful when sampling from a “hidden” or hard to access population, and is especially common in social sciences. For instance, suppose the researchers want to conduct an opinion poll among transgender people. The research team might only have contact details for a few trans folks, so the survey starts by asking them to participate (stage 1). At the end

of the survey, the participants are asked to provide contact details for other people who might want to participate. In stage 2, those new contacts are surveyed. The process continues until the researchers have sufficient data. The big advantage to snowball sampling is that it gets you data in situations that might otherwise be impossible to get any. On the statistical side, the main disadvantage is that the sample is highly non-random, and non-random in ways that are difficult to address. On the real life side, the disadvantage is that the procedure can be unethical if not handled well, because hidden populations are often hidden for a reason. I chose transgender people as an example here to highlight this: if you weren't careful you might end up outing people who don't want to be outed (very, very bad form), and even if you don't make that mistake it can still be intrusive to use people's social networks to study them. It's certainly very hard to get people's informed consent **before** contacting them, yet in many cases the simple act of contacting them and saying "hey we want to study you" can be hurtful. Social networks are complex things, and just because you can use them to get data doesn't always mean you should.

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

4.8.4 How much does it matter if you don't have a simple random sample?

Okay, so real world data collection tends not to involve nice simple random samples. Does that matter? A little thought should make it clear to you that it **can** matter if your data are not a simple random sample: just think about the difference between Figures 4.9 and 4.10. However, it's not quite as bad as it sounds. Some types of biased samples are entirely unproblematic. For instance, when using a stratified sampling technique you actually **know** what the bias is because you created it deliberately, often to **increase** the effectiveness of your study, and there are statistical techniques that you can use to adjust for the biases you've introduced (not covered in this book!). So in those situations it's not a problem.

More generally though, it's important to remember that random sampling is a means to an end, not the end in itself. Let's assume you've relied on a convenience sample, and as such you can assume it's biased. A bias in your sampling method is only a problem if it causes you to draw the wrong conclusions. When viewed from that perspective, I'd argue that we don't need the sample to be randomly generated in **every** respect: we only need it to be random with respect to the psychologically-relevant phenomenon of interest. Suppose I'm doing a study looking at working memory capacity. In study 1, I actually have the ability to sample randomly from all human beings currently alive, with one exception: I can only sample people born on a Monday. In study 2, I am able to sample randomly from the Australian population. I want to generalize my results to the population of all living humans. Which study is better? The answer, obviously, is study 1. Why? Because we have no reason to think that being "born on a Monday" has any interesting relationship to working memory capacity. In contrast, I can think of several reasons why "being Australian" might matter. Australia is a wealthy, industrialized country with a very well-developed education system. People growing up in that system will have had life experiences much more similar to the experiences of the people who designed the tests for working memory capacity. This shared experience might easily translate into similar beliefs about how to "take a test", a shared assumption about how psychological experimentation works, and so on. These things might actually matter. For instance, "test taking" style might have taught the Australian participants how to direct their attention exclusively on fairly abstract test materials relative to people that haven't grown up in a similar environment; leading to a misleading picture of what working memory capacity is.

There are two points hidden in this discussion. Firstly, when designing your own studies, it's important to

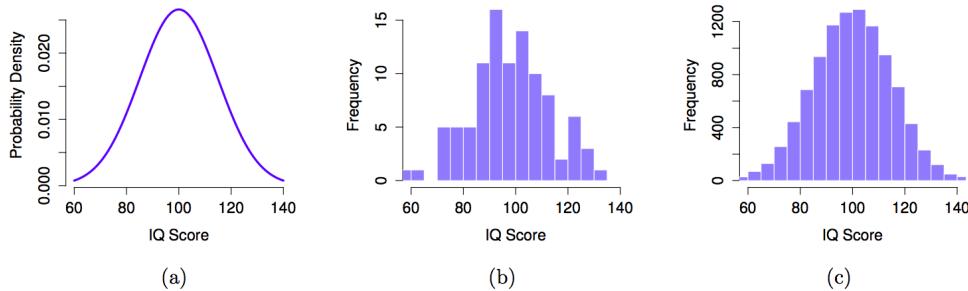


Figure 4.12: The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and panel c we have a sample of 10,000 observations.

think about what population you care about, and try hard to sample in a way that is appropriate to that population. In practice, you’re usually forced to put up with a “sample of convenience” (e.g., psychology lecturers sample psychology students because that’s the least expensive way to collect data, and our coffers aren’t exactly overflowing with gold), but if so you should at least spend some time thinking about what the dangers of this practice might be.

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

4.8.5 Population parameters and sample statistics

Okay. Setting aside the thorny methodological issues associated with obtaining a random sample, let’s consider a slightly different issue. Up to this point we have been talking about populations the way a scientist might. To a psychologist, a population might be a group of people. To an ecologist, a population might be a group of bears. In most cases the populations that scientists care about are concrete things that actually exist in the real world.

Statisticians, however, are a funny lot. On the one hand, they **are** interested in real world data and real science in the same way that scientists are. On the other hand, they also operate in the realm of pure abstraction in the way that mathematicians do. As a consequence, statistical theory tends to be a bit abstract in how a population is defined. In much the same way that psychological researchers operationalize our abstract theoretical ideas in terms of concrete measurements, statisticians operationalize the concept of a “population” in terms of mathematical objects that they know how to work with. You’ve already come across these objects they’re called probability distributions (remember, the place where data comes from).

The idea is quite simple. Let’s say we’re talking about IQ scores. To a psychologist, the population of interest is a group of actual humans who have IQ scores. A statistician “simplifies” this by operationally defining the population as the probability distribution depicted in Figure 4.12a.

IQ tests are designed so that the average IQ is 100, the standard deviation of IQ scores is 15, and the distribution of IQ scores is normal. These values are referred to as the **population parameters** because

they are characteristics of the entire population. That is, we say that the population mean μ is 100, and the population standard deviation σ is 15.

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

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

Each of these IQ scores is sampled from a normal distribution with mean 100 and standard deviation 15. So if I plot a histogram of the sample, I get something like the one shown in Figure 4.12b. As you can see, the histogram is **roughly** the right shape, but it's a very crude approximation to the true population distribution shown in Figure 4.12a. The mean of the sample is fairly close to the population mean 100 but not identical. In this case, it turns out that the people in the sample have a mean IQ of 98.5, and the standard deviation of their IQ scores is 15.9. These **sample statistics** are properties of the data set, and although they are fairly similar to the true population values, they are not the same. **In general, sample statistics are the things you can calculate from your data set, and the population parameters are the things you want to learn about.** Later on in this chapter we'll talk about how you can estimate population parameters using your sample statistics and how to work out how confident you are in your estimates but before we get to that there's a few more ideas in sampling theory that you need to know about.

4.9 The law of large numbers

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

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

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

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

```
print(IQ[1:100])
```

```
## [1] 115 99 109 87 98 94 100 74 93 91 82 86 89 136 98 102 97
## [18] 130 105 108 90 99 103 112 121 95 83 92 94 91 96 120 84 118
## [35] 115 106 116 108 85 83 110 76 116 101 91 100 94 81 84 103 100
## [52] 108 100 113 80 103 81 79 136 104 88 116 94 95 115 111 96 75
## [69] 106 104 92 81 117 86 95 90 80 99 119 130 77 112 83 107 91
## [86] 93 88 89 91 112 117 84 92 68 113 84 125 99 119 95
```

We can compute the mean IQ using the command **mean(IQ)** and the standard deviation using the command **sd(IQ)**, and draw a histogram using **hist()**. The histogram of this much larger sample is shown in Figure 4.12c. Even a moment's inspection makes clear that the larger sample is a much better approximation to the true population distribution than the smaller one. This is reflected in the sample statistics: the mean IQ for the larger sample turns out to be 99.9, and the standard deviation is 15.1. These values are now very close to the true population.

I feel a bit silly saying this, but the thing I want you to take away from this is that large samples generally give you better information. I feel silly saying it because it's so bloody obvious that it shouldn't need to be said. In fact, it's such an obvious point that when Jacob Bernoulli – one of the founders of probability theory – formalized this idea back in 1713, he was kind of a jerk about it. Here's how he described the fact that we all share this intuition:

For even the most stupid of men, by some instinct of nature, by himself and without any instruction (which is a remarkable thing), is convinced that the more observations have been made, the less danger there is of wandering from one's goal (see Stigler, 1986, p65).

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

I don't intend to subject you to a proof that the law of large numbers is true, but it's one of the most important tools for statistical theory. The law of large numbers is the thing we can use to justify our belief that collecting more and more data will eventually lead us to the truth. For any particular data set, the sample statistics that we calculate from it will be wrong, but the law of large numbers tells us that if we keep collecting more data those sample statistics will tend to get closer and closer to the true population parameters.

4.10 Sampling distributions and the central limit theorem

The law of large numbers is a very powerful tool, but it's not going to be good enough to answer all our questions. Among other things, all it gives us is a “long run guarantee”. In the long run, if we were somehow able to collect an infinite amount of data, then the law of large numbers guarantees that our sample statistics will be correct. But as John Maynard Keynes famously argued in economics, a long run guarantee is of little use in real life:

[The] long run is a misleading guide to current affairs. In the long run we are all dead. Economists set themselves too easy, too useless a task, if in tempestuous seasons they can only tell us, that when the storm is long past, the ocean is flat again. Keynes (1923, p.80)

As in economics, so too in psychology and statistics. It is not enough to know that we will **eventually** arrive at the right answer when calculating the sample mean. Knowing that an infinitely large data set will tell me the exact value of the population mean is cold comfort when my **actual** data set has a sample size of $N = 100$. In real life, then, we must know something about the behavior of the sample mean when it is calculated from a more modest data set!

4.10.1 Sampling distribution of the sample means

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

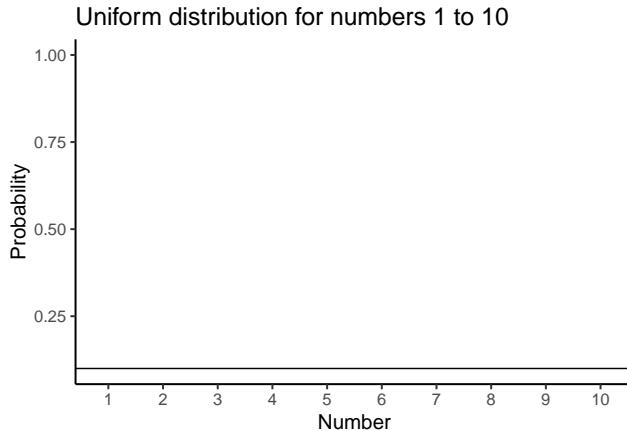


Figure 4.13: A uniform distribution illustrating the probabilities of sampling the numbers 1 to 10. In a uniform distribution, all numbers have an equal probability of being sampled, so the line is flat indicating all numbers have the same probability

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

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

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

4.10.2 Seeing the pieces

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

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

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

Let's do those four things. We will sample numbers from the uniform distribution, it looks like this if we are sampling from the set of integers from 1 to 10:

Animation not available in .pdf version

Figure 4.14: Animation showing histograms for different samples of size 20 from the uniform distribution. The red line shows the mean of each sample

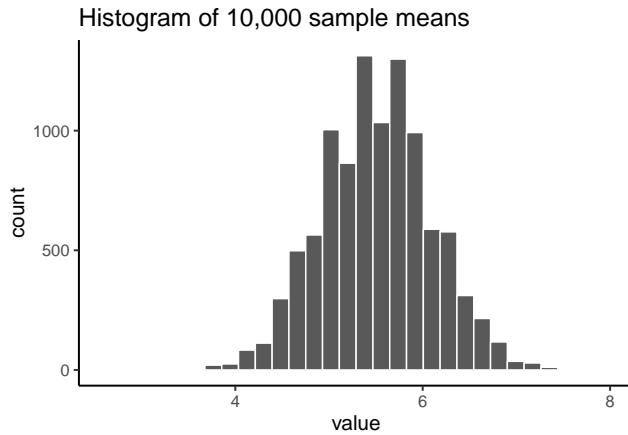


Figure 4.15: A histogram showing the sample means for 10,000 samples, each size 20, from the uniform distribution of numbers from 1 to 10. The expected mean is 5.5, and the histogram is centered on 5.5. The mean of each sample is not always 5.5 because of sampling error or chance

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

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

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

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

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

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

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

1 2 3 4 5 6 7 8 9 10

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

4.10.3 Sampling distributions exist for any sample statistic!

One thing to keep in mind when thinking about sampling distributions is that **any** sample statistic you might care to calculate has a sampling distribution. For example, suppose that each time you sampled some numbers from an experiment you wrote down the largest number in the experiment. Doing this over and over again would give you a very different sampling distribution, namely the **sampling distribution of the maximum**. You could calculate the smallest number, or the mode, or the median, or the variance, or the standard deviation, or anything else from your sample. Then, you could repeat many times, and produce the sampling distribution of those statistics. Neat!

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

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

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

4.11 The central limit theorem

OK, so now you've seen lots of sampling distributions, and you know what the sampling distribution of the mean is. Here, we'll focus on **how the sampling distribution of the mean changes as a function of**

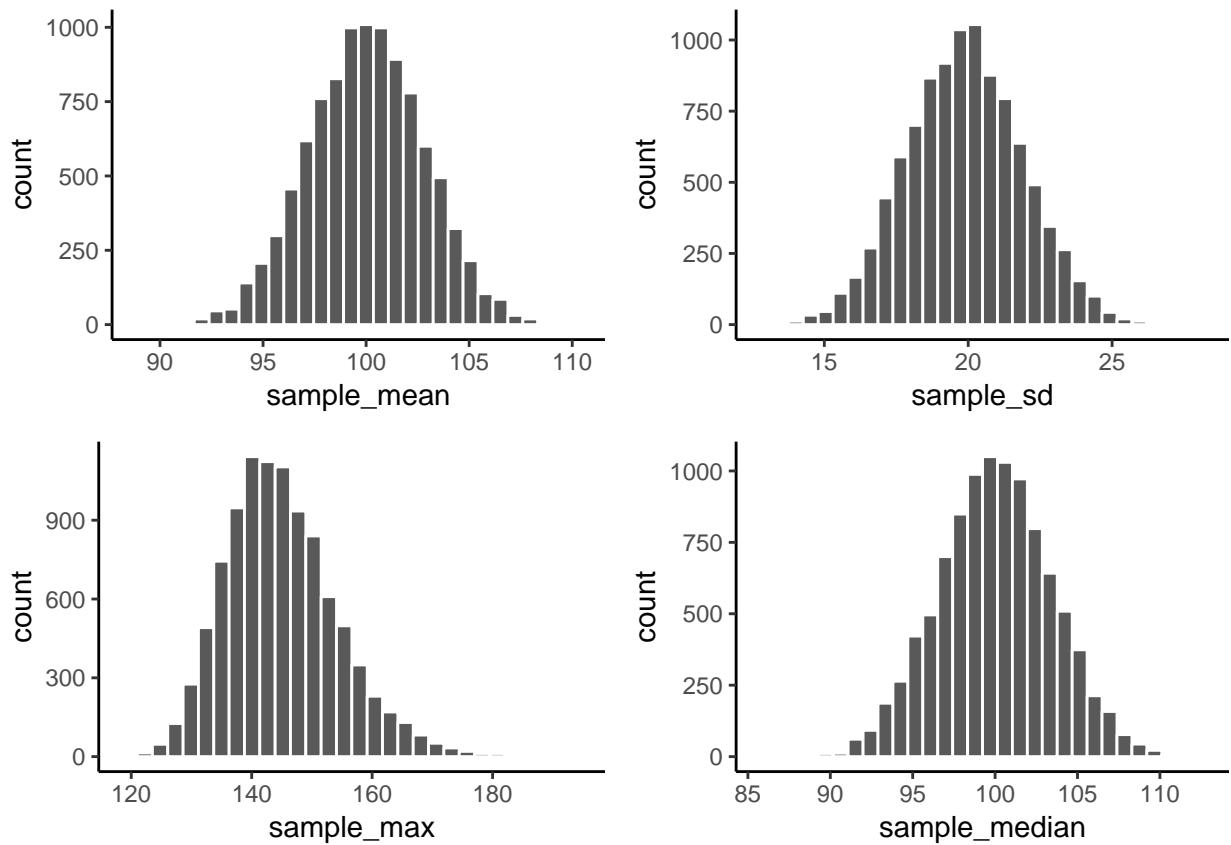


Figure 4.16: Each panel shows a histogram of a different sampling statistic

Animation not available in .pdf version

Figure 4.17: Animation of samples (grey histogram shows frequency counts of data in each sample), and the sampling distribution of the mean (histogram of the sampling means for many samples). Each sample is taken from the normal distribution shown in red. The moving red line is the mean of an individual sample. The blue line is the mean of the blue histogram, which represents the sampling distribution of the mean for many samples

sample size.

Intuitively, you already know part of the answer: if you only have a few observations, the sample mean is likely to be quite inaccurate (you've already seen it bounce around): if you replicate a small experiment and recalculate the mean you'll get a very different answer. In other words, the sampling distribution is quite wide. If you replicate a large experiment and recalculate the sample mean you'll probably get the same answer you got last time, so the sampling distribution will be very narrow.

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

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

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

So, the sampling distribution of the mean is another distribution, and it has some variance. It varies more when sample-size is small, and varies less when sample-size is large. We can quantify this effect by calculating the standard deviation of the sampling distribution, which is referred to as the **standard error**. The standard error of a statistic is often denoted SE , and since we're usually interested in the standard error of the sample **mean**, we often use the acronym SEM . As you can see just by looking at the movie, as the sample size N increases, the SEM decreases.

Okay, so that's one part of the story. However, there's something we've been glossing over a little bit. We've seen it already, but it's worth looking at it one more time. Here's the thing: **no matter what shape your population distribution is, as N increases the sampling distribution of the mean starts to look more like a normal distribution.** This is the central limit theorem.

To see the central limit theorem in action, we are going to look at some histograms of sample means different kinds of distributions. It is very important to recognize that you are looking at distributions of sample means, not distributions of individual samples! Here we go, starting with sampling from a normal distribution. The red line is the distribution, the blue bars are the histogram for the sample means. They both look normal!

Let's do it again. This time we sample from a flat uniform distribution. Again, we see that the distribution of the sample means is not flat, it looks like a normal distribution.

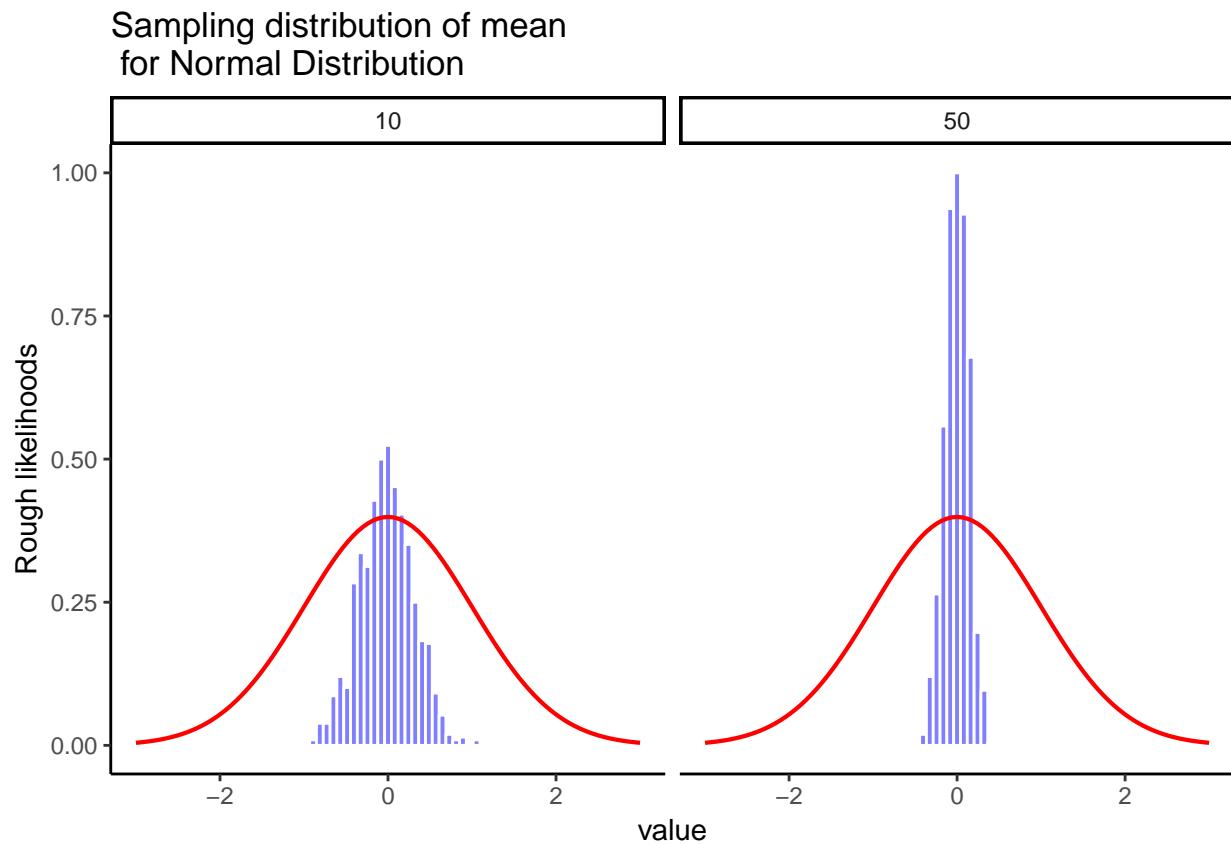


Figure 4.18: Comparison of two normal distributions, and histograms for the sampling distribution of the mean for different samples-sizes. The range of sampling distribution of the mean shrinks as sample-size increases

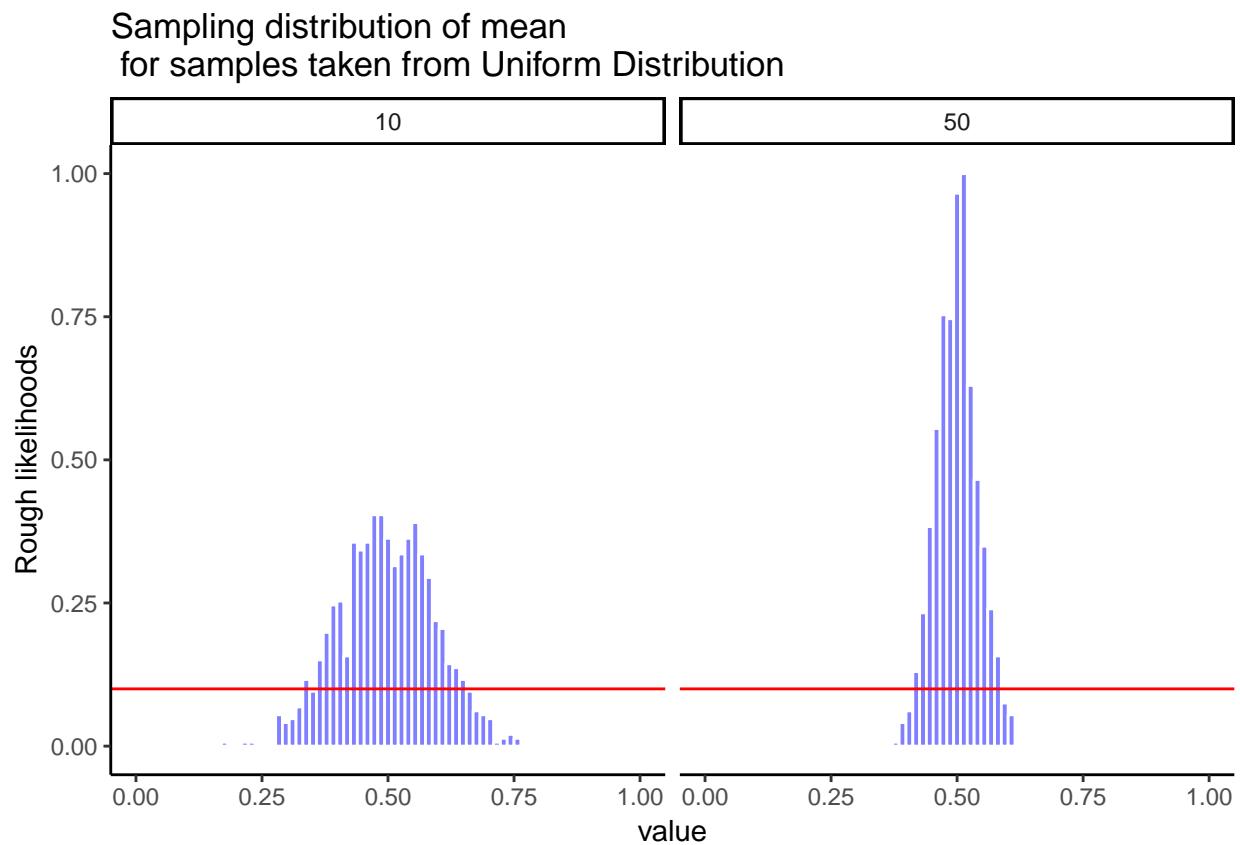


Figure 4.19: Illustration that the shape of the sampling distribution of the mean is normal, even when the samples come from a non-normal (uniform in this case) distribution

One more time with an exponential distribution. Even though way more of the numbers should be smaller than bigger, then sampling distribution of the mean again does not look the red line. Instead, it looks more normal-ish. That's the central limit theorem. It just works like that.

Sampling distribution of mean for samples from exponential Distribution

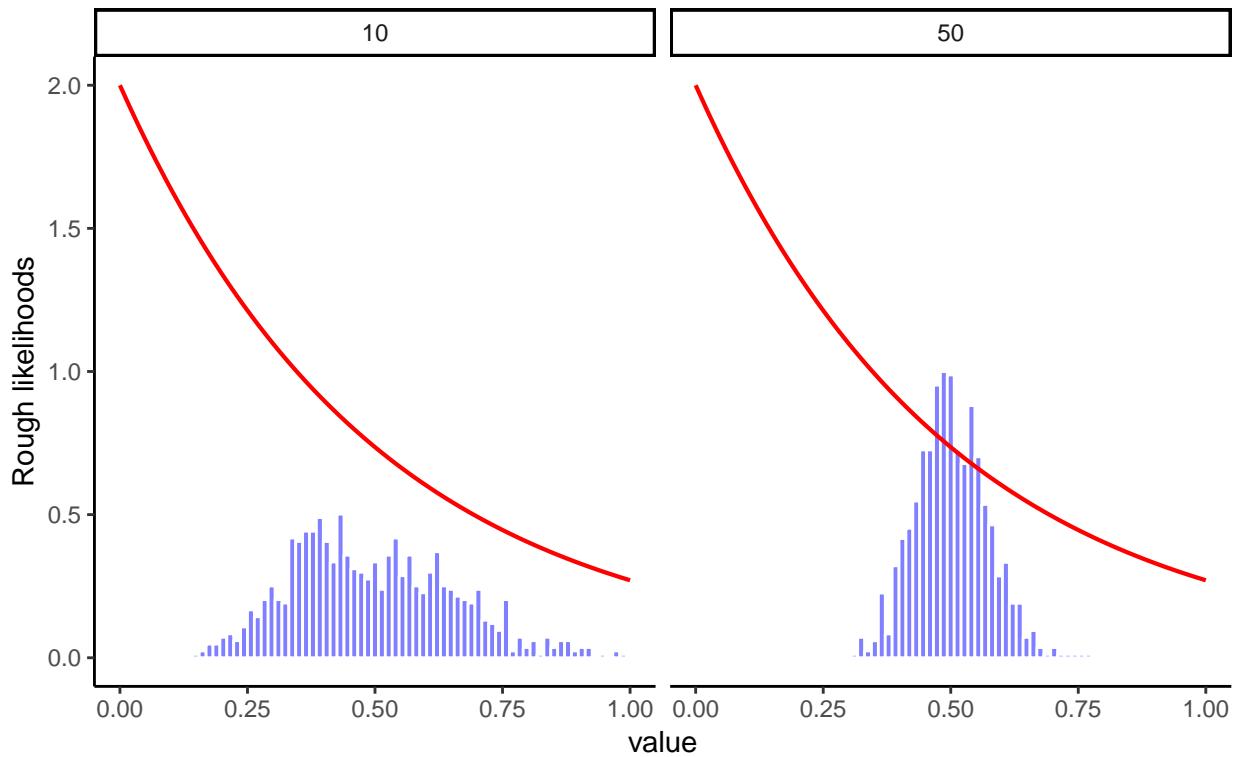


Figure 4.20: Illustration that the shape of the sampling distribution of the mean is normal, even when the samples come from a non-normal (exponential in this case) distribution

On the basis of these figures, it seems like we have evidence for all of the following claims about the sampling distribution of the mean:

- The mean of the sampling distribution is the same as the mean of the population
- The standard deviation of the sampling distribution (i.e., the standard error) gets smaller as the sample size increases
- The shape of the sampling distribution becomes normal as the sample size increases

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

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

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

This result is useful for all sorts of things. It tells us why large experiments are more reliable than small ones, and because it gives us an explicit formula for the standard error it tells us **how much** more reliable a large

experiment is. It tells us why the normal distribution is, well, **normal**. In real experiments, many of the things that we want to measure are actually averages of lots of different quantities (e.g., arguably, “general” intelligence as measured by IQ is an average of a large number of “specific” skills and abilities), and when that happens, the averaged quantity should follow a normal distribution. Because of this mathematical law, the normal distribution pops up over and over again in real data.

4.12 z-scores

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

The first thing we show you seems to be something that many students remember from their statistics class. This thing is probably remembered because instructors may test this knowledge many times, so students have to learn it for the test. Let’s look at this thing. We are going to look at a normal distribution, and we are going to draw lines through the distribution at 0 , $+\/-1$, $+\/-2$, and $+\/-3$ standard deviations from the mean:

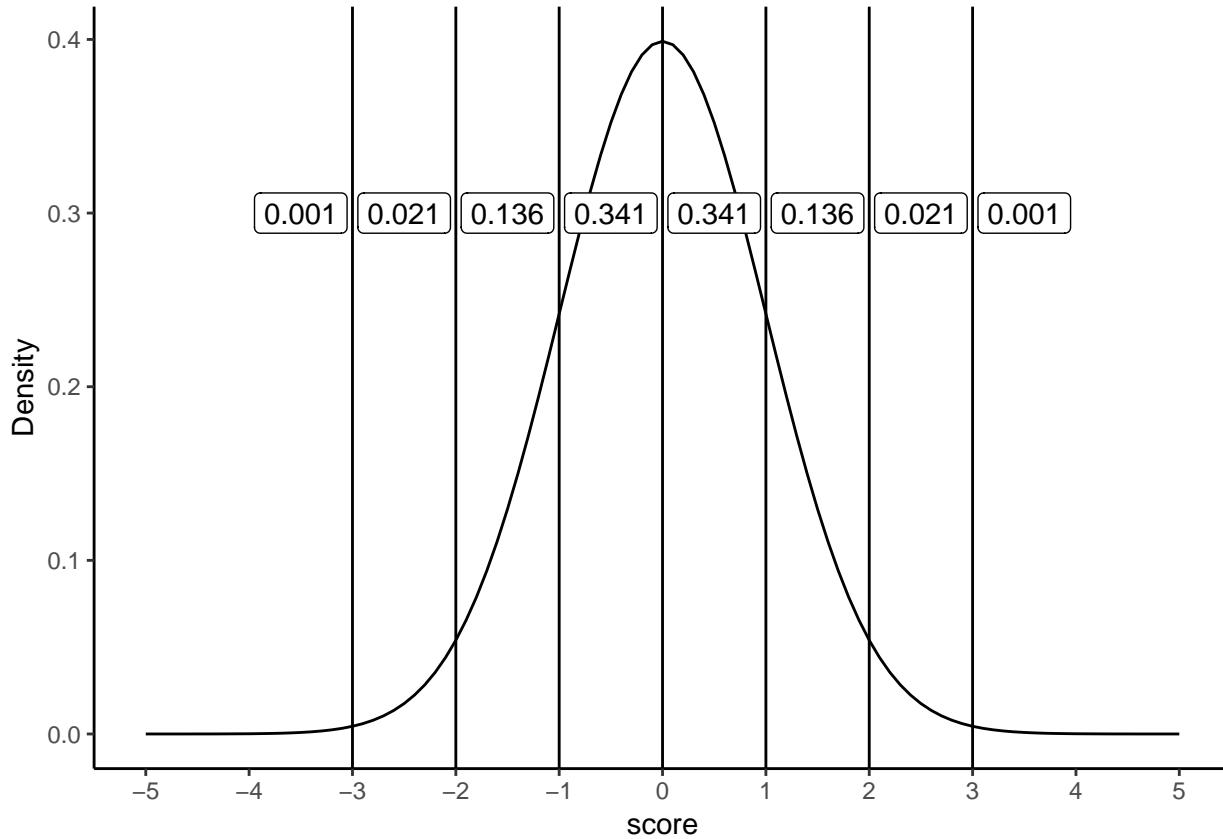


Figure 4.21: A normal distribution. Each line represents a standard deviation from the mean. The labels show the proportions of scores that fall between each bar.

The figure shows a normal distribution with $\text{mean} = 0$, and $\text{standard deviation} = 1$. We’ve drawn lines at each of the standard deviations: -3 , -2 , -1 , 0 , 1 , 2 , and 3 . We also show some numbers in the labels, in between each line. These numbers are proportions. For example, we see the proportion is $.341$ for scores that fall between the range 0 and 1 . Scores between 0 and 1 occur 34.1% of the time. Scores in between -1

and 1, occur 68.2% of the time, that's more than half of the scores. Scores between 1 and occur about 13.6% of the time, and scores between 2 and 3 occur even less, only 2.1% of the time.

Normal distributions always have these properties, even when they have different means and standard deviations. For example, take a look at this normal distribution, it has a mean =100, and standard deviation =25.

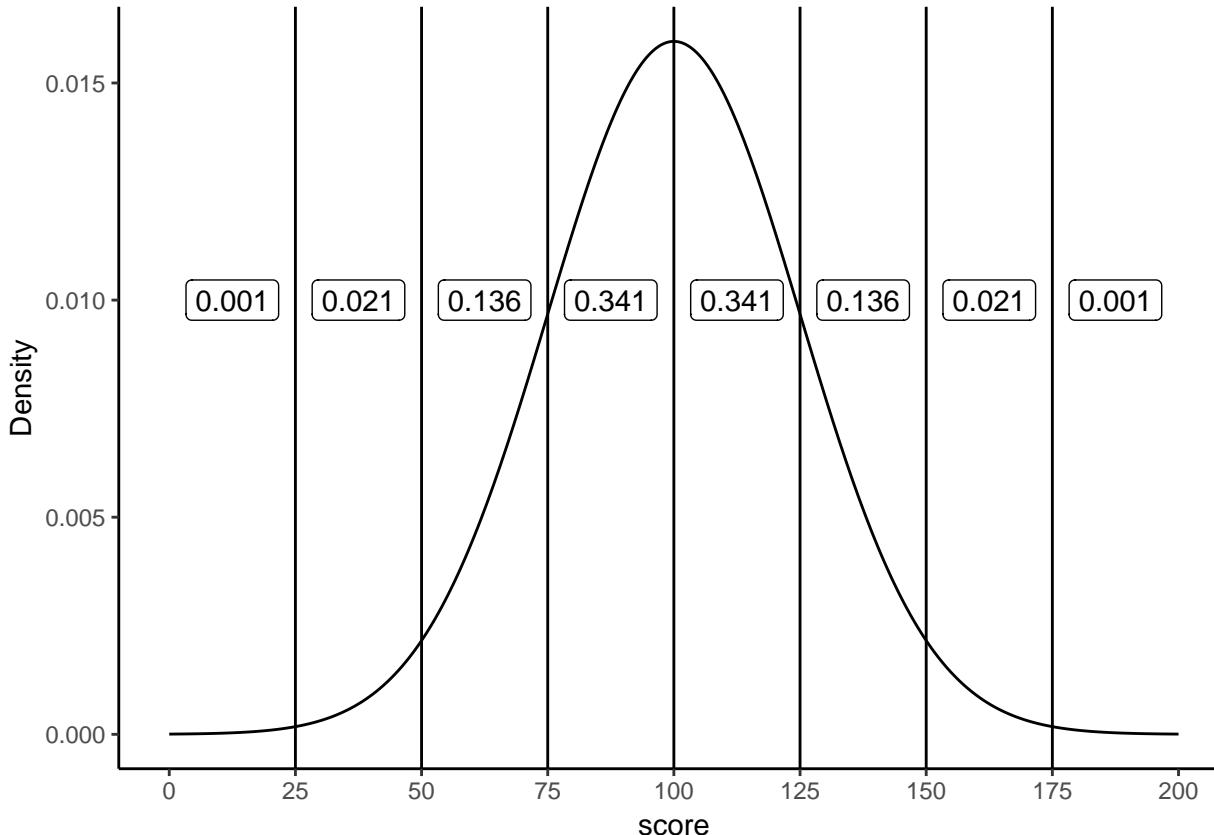


Figure 4.22: A normal distribution. Each line represents a standard deviation from the mean. The labels show the proportions of scores that fall between each bar.

Now we are looking at a normal distribution with mean = 100 and standard deviation = 25. Notice that the region between 100 and 125 contains 34.1% of the scores. This region is 1 standard deviation away from the mean (the standard deviation is 25, the mean is 100, so 25 is one whole standard deviation away from 100). As you can see, the very same proportions occur between each of the standard deviations, as they did when our standard deviation was set to 1 (with a mean of 0).

4.12.1 Idea behind z-scores

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

The idea behind z-scores is a similar kind of transformation. The idea is to express each raw score in terms of its standard deviation. For example, if I told you I got a 75% on test, you wouldn't know how well I did

compared to the rest of the class. But, if I told you that I scored 2 standard deviations above the mean, you'd know I did quite well compared to the rest of the class, because you know that most scores (if they are distributed normally) fall below 2 standard deviations of the mean.

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

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

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

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

4.12.2 Calculating z-scores

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

First compute the difference between the score and the mean:

$$97 - 100 = -3$$

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

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

Our z-score for 97 is -.12.

Here's the general formula:

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

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

```
## [1] 127.78 98.55 86.80 91.38 48.69 91.93 72.35 84.48 109.59 95.90
```

The z-scores would be:

```
## [1] 1.1112 -0.0580 -0.5280 -0.3448 -2.0524 -0.3228 -1.1060 -0.6208
## [9] 0.3836 -0.1640
```

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

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

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

4.13 Estimating population parameters

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

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

4.13.1 Concrete population parameters

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

Fortunately, it's pretty easy to get the population parameters without measuring the entire population. Who has time to measure every-bodies feet? Nobody, that's who. Instead, you would just need to randomly pick a bunch of people, measure their feet, and then measure the parameters of the sample. If you take a big enough sample, we have learned that the sample mean gives a very good estimate of the population mean. We will learn shortly that a version of the standard deviation of the sample also gives a good estimate of the standard deviation of the population. Perhaps shoe-sizes have a slightly different shape than a normal distribution. Here too, if you collect a big enough sample, the shape of the distribution of the sample will

be a good estimate of the shape of the populations. All of these are good reasons to care about estimating population parameters. But, do you run a shoe company? Probably not.

4.13.2 Abstract population parameters

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

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

1. = very unhappy
2. = unhappy
3. = sort of unhappy
4. = in the middle
5. = sort of happy
6. = happy
7. = very happy

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

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

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

4.13.2.1 Complications with inference

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

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

Problem 2: What do these questions measure?: If the whole point of doing the questionnaire is to estimate the population's happiness, we really need wonder if the sample measurements actually tell us anything about happiness in the first place. Some questions: Are people accurate in saying how happy they

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

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

4.13.3 Experiments and Population parameters

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

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

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

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

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

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

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

our samples, the more they will look the same, especially when we don't do anything to cause them to be different. In other words, we can use the parameters of one sample to estimate the parameters of a second sample, because they will tend to be the same, especially when they are large.

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

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

4.13.4 Interim summary

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

4.13.5 Estimating the population mean

Suppose we go to Brooklyn and 100 of the locals are kind enough to sit through an IQ test. The average IQ score among these people turns out to be $\bar{X} = 98.5$. So what is the true mean IQ for the entire population of Brooklyn? Obviously, we don't know the answer to that question. It could be 97.2, but it could also be 103.5. Our sampling isn't exhaustive so we cannot give a definitive answer. Nevertheless if forced to give a "best guess" I'd have to say 98.5. That's the essence of statistical estimation: giving a best guess. We're using the sample mean as the best guess of the population mean.

In this example, estimating the unknown population parameter is straightforward. I calculate the sample mean, and I use that as my **estimate of the population mean**. It's pretty simple, and in the next section we'll explain the statistical justification for this intuitive answer. However, for the moment let's make sure you recognize that the sample statistic and the estimate of the population parameter are conceptually different things. A sample statistic is a description of your data, whereas the estimate is a guess about the population. With that in mind, statisticians often use different notation to refer to them. For instance, if true population mean is denoted μ , then we would use $\hat{\mu}$ to refer to our estimate of the population mean. In contrast, the sample mean is denoted \bar{X} or sometimes m . However, in simple random samples, the estimate of the population mean is identical to the sample mean: if I observe a sample mean of $\bar{X} = 98.5$, then my estimate of the population mean is also $\hat{\mu} = 98.5$. To help keep the notation clear, here's a handy table:

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

4.13.6 Estimating the population standard deviation

So far, estimation seems pretty simple, and you might be wondering why I forced you to read through all that stuff about sampling theory. In the case of the mean, our estimate of the population parameter (i.e. $\hat{\mu}$) turned out to be identical to the corresponding sample statistic (i.e. \bar{X}). However, that's not always true. To see this, let's have a think about how to construct an **estimate of the population standard deviation**, which we'll denote $\hat{\sigma}$. What shall we use as our estimate in this case? Your first thought might be that we could do the same thing we did when estimating the mean, and just use the sample statistic as our estimate. That's almost the right thing to do, but not quite.

Here's why. Suppose I have a sample that contains a single observation. For this example, it helps to consider a sample where you have no intuitions at all about what the true population values might be, so let's use something completely fictitious. Suppose the observation in question measures the **cromulence** of my shoes. It turns out that my shoes have a cromulence of 20. So here's my sample:

20

This is a perfectly legitimate sample, even if it does have a sample size of $N = 1$. It has a sample mean of 20, and because every observation in this sample is equal to the sample mean (obviously!) it has a sample standard deviation of 0. As a description of the **sample** this seems quite right: the sample contains a single observation and therefore there is no variation observed within the sample. A sample standard deviation of $s = 0$ is the right answer here. But as an estimate of the **population** standard deviation, it feels completely insane, right? Admittedly, you and I don't know anything at all about what "cromulence" is, but we know something about data: the only reason that we don't see any variability in the **sample** is that the sample is too small to display any variation! So, if you have a sample size of $N = 1$, it **feels** like the right answer is just to say "no idea at all".

Notice that you **don't** have the same intuition when it comes to the sample mean and the population mean. If forced to make a best guess about the population mean, it doesn't feel completely insane to guess that the population mean is 20. Sure, you probably wouldn't feel very confident in that guess, because you have only the one observation to work with, but it's still the best guess you can make.

Let's extend this example a little. Suppose I now make a second observation. My data set now has $N = 2$ observations of the cromulence of shoes, and the complete sample now looks like this:

20, 22

This time around, our sample is **just** large enough for us to be able to observe some variability: two observations is the bare minimum number needed for any variability to be observed! For our new data set, the sample mean is $\bar{X} = 21$, and the sample standard deviation is $s = 1$. What intuitions do we have about the population? Again, as far as the population mean goes, the best guess we can possibly make is the

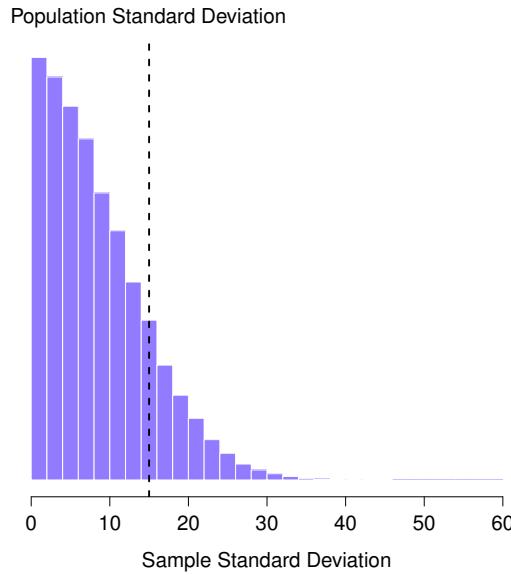


Figure 4.23: The sampling distribution of the sample standard deviation for a two IQ scores experiment. The true population standard deviation is 15 (dashed line), but as you can see from the histogram, the vast majority of experiments will produce a much smaller sample standard deviation than this. On average, this experiment would produce a sample standard deviation of only 8.5, well below the true value! In other words, the sample standard deviation is a biased estimate of the population standard deviation.

sample mean: if forced to guess, we'd probably guess that the population mean cromulence is 21. What about the standard deviation? This is a little more complicated. The sample standard deviation is only based on two observations, and if you're at all like me you probably have the intuition that, with only two observations, we haven't given the population "enough of a chance" to reveal its true variability to us. It's not just that we suspect that the estimate is **wrong**: after all, with only two observations we expect it to be wrong to some degree. The worry is that the error is **systematic**.

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

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

It would be nice to demonstrate this somehow. There are in fact mathematical proofs that confirm this intuition, but unless you have the right mathematical background they don't help very much. Instead, what I'll do is use R to simulate the results of some experiments. With that in mind, let's return to our IQ studies. Suppose the true population mean IQ is 100 and the standard deviation is 15. I can use the `rnorm()` function to generate the the results of an experiment in which I measure $N = 2$ IQ scores, and calculate the sample standard deviation. If I do this over and over again, and plot a histogram of these sample standard deviations, what I have is the **sampling distribution of the standard deviation**. I've plotted this distribution in Figure 4.23.

Even though the true population standard deviation is 15, the average of the **sample** standard deviations is only 8.5. Notice that this is a very different from when we were plotting sampling distributions of the sample mean, those were always centered around the mean of the population.

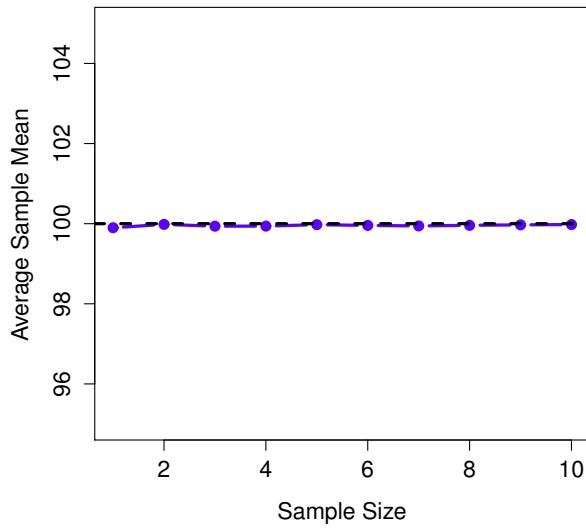


Figure 4.24: An illustration of the fact that the sample mean is an unbiased estimator of the population mean.

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

Figure 4.24 shows the sample mean as a function of sample size. Notice it's a flat line. The sample mean doesn't underestimate or overestimate the population mean. It is an unbiased estimate!

Figure 4.25 shows the sample standard deviation as a function of sample size. Notice it is not a flat line. The sample standard deviation systematically underestimates the population standard deviation!

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

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

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

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

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

This is an unbiased estimator of the population variance σ .

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

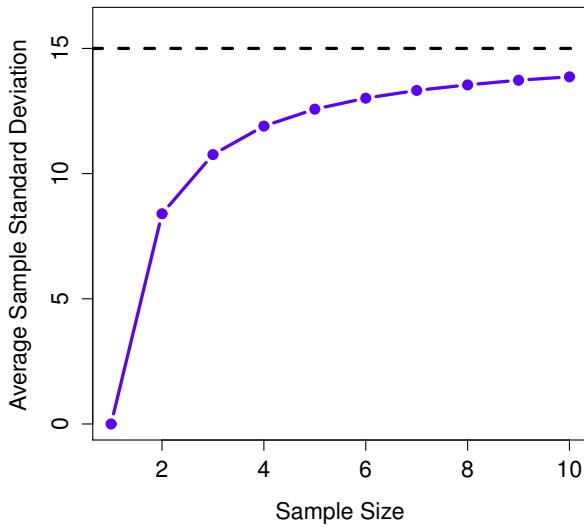


Figure 4.25: An illustration of the fact that the sample standard deviation is a biased estimator of the population standard deviation

population standard deviation becomes:

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

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

One final point: in practice, a lot of people tend to refer to $\hat{\sigma}$ (i.e., the formula where we divide by $N - 1$) as the **sample** standard deviation. Technically, this is incorrect: the **sample** standard deviation should be equal to s (i.e., the formula where we divide by N). These aren't the same thing, either conceptually or numerically. One is a property of the sample, the other is an estimated characteristic of the population. However, in almost every real life application, what we actually care about is the estimate of the population parameter, and so people always report $\hat{\sigma}$ rather than s .

Note, whether you should divide by N or $N-1$ also depends on your philosophy about what you are doing. For example, if you don't think that what you are doing is estimating a population parameter, then why would you divide by $N-1$? Also, when N is large, it doesn't matter too much. The difference between a big N , and a big $N-1$, is just -1 .

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

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

Symbol	What is it?	Do we know what it is?
s^2	Sample variance	Yes, calculated from the raw data
σ^2	Population variance	Almost never known for sure
$\hat{\sigma}^2$	Estimate of the population variance	Yes, but not the same as the sample variance

4.14 Estimating a confidence interval

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

Up to this point in this chapter, we've outlined the basics of sampling theory which statisticians rely on to make guesses about population parameters on the basis of a sample of data. As this discussion illustrates, one of the reasons we need all this sampling theory is that every data set leaves us with some of uncertainty, so our estimates are never going to be perfectly accurate. The thing that has been missing from this discussion is an attempt to **quantify** the amount of uncertainty in our estimate. It's not enough to be able guess that the mean IQ of undergraduate psychology students is 115 (yes, I just made that number up). We also want to be able to say something that expresses the degree of certainty that we have in our guess. For example, it would be nice to be able to say that there is a 95% chance that the true mean lies between 109 and 121. The name for this is a **confidence interval** for the mean.

Armed with an understanding of sampling distributions, constructing a confidence interval for the mean is actually pretty easy. Here's how it works. Suppose the true population mean is μ and the standard deviation is σ . I've just finished running my study that has N participants, and the mean IQ among those participants is \bar{X} . We know from our discussion of the central limit theorem that the sampling distribution of the mean is approximately normal. We also know from our discussion of the normal distribution that there is a 95% chance that a normally-distributed quantity will fall within two standard deviations of the true mean. To be more precise, we can use the **qnorm()** function to compute the 2.5th and 97.5th percentiles of the normal distribution

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

Okay, so I lied earlier on. The more correct answer is that a 95% chance that a normally-distributed quantity will fall within 1.96 standard deviations of the true mean.

Next, recall that the standard deviation of the sampling distribution is referred to as the standard error, and the standard error of the mean is written as SEM. When we put all these pieces together, we learn that there is a 95% probability that the sample mean \bar{X} that we have actually observed lies within 1.96 standard errors of the population mean. Oof, that is a lot of mathy talk there. We'll clear it up, don't worry.

Mathematically, we write this as:

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

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

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

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

What this is telling us is that the range of values has a 95% probability of containing the population mean μ . We refer to this range as a **95% confidence interval**, denoted CI_{95} . In short, as long as N is sufficiently large – large enough for us to believe that the sampling distribution of the mean is normal – then we can write this as our formula for the 95% confidence interval:

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

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

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

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

4.14.1 A slight mistake in the formula

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

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

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

4.15 Summary

In this chapter I've covered two main topics. The first half of the chapter talks about sampling theory, and the second half talks about how we can use sampling theory to construct estimates of the population parameters. The section breakdown looks like this:

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

As always, there's a lot of topics related to sampling and estimation that aren't covered in this chapter, but for an introductory psychology class this is fairly comprehensive I think. For most applied researchers you won't need much more theory than this. One big question that I haven't touched on in this chapter is what you do when you don't have a simple random sample. There is a lot of statistical theory you can draw on to handle this situation, but it's well beyond the scope of this book.

Chapter 5

Foundations for inference

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

Chapter by Matthew Crump

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

Now we begin our journey into **inferential statistics**. The tools we use to make inferences about where our data came from, and more importantly make inferences about what causes what.

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

5.1 Brief review of Experiments

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

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

and the other 25 subjects would be randomly assigned to watch 2 hours of cat videos. We would measure happiness for each subject at the end of the videos. Then we could look at the data. What would we want to look at? Well, if watching cat videos cause change in happiness, then we would expect the measures of happiness for people watching 1 hour of cat videos to be different from the measures of happiness for people watching 2 hours of cat videos. If watching cat videos does not change happiness, then we would expect no differences in measures of happiness between conditions. Causal forces cause change, and the experiment is set up to detect the change.

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

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

5.2 The data came from a distribution

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

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

5.2.1 Uniform distribution

A uniform distribution is completely flat, it looks like this:

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

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

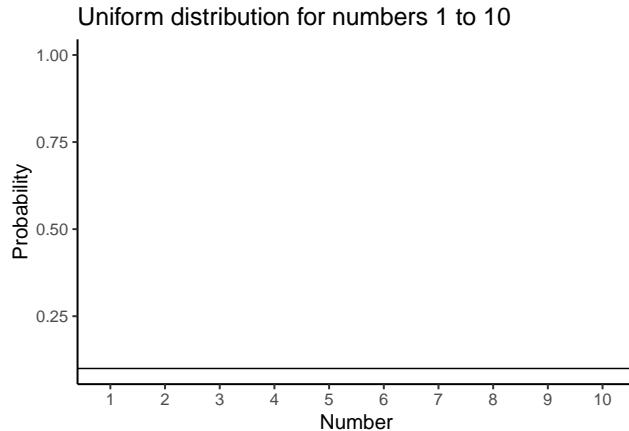


Figure 5.1: Uniform distribution showing that the numbers from 1 to 10 have an equal probability of being sampled

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

We used the uniform distribution to generate these numbers. Officially, we call this **sampling** from a **distribution**. Sampling is what you do at a grocery store when there is free food. You can keep taking more. However, if you take all of the samples, then what you have is called the **population**. We'll talk more about samples and populations as we go along.

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

We already know what a histogram is, so we can put our numbers into a histogram and see what the counts look like. If all of the numbers occur with equal frequency, then each number should occur 10 times, because we sampled a total of 100 numbers. The histogram looks like this:

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

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

Let's take a look at sampling error more closely. We will sample 20 numbers from the uniform. Here we should expect that each number between 1 and 10 occurs two times each. Let's take 20 samples and make a

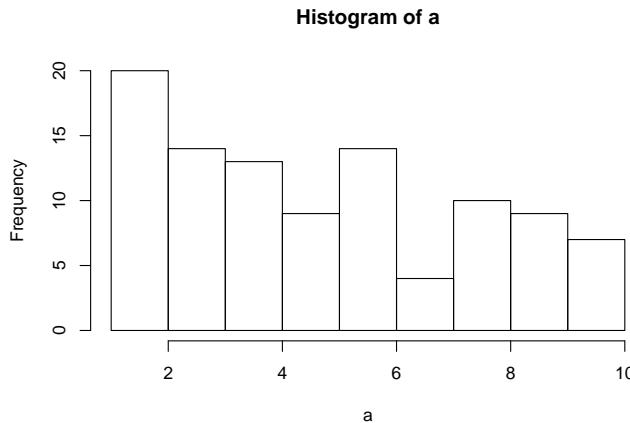


Figure 5.2: Histogram of a sample of 100 numbers from the uniform distribution containing the integers from 1 to 10

histogram. And then, let's do that 10 times. So we will be looking at 10 histograms, each showing us what the 10 different samples of twenty numbers looks like:

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

Here is movie version. You are watching a new histogram for each sample of 20 observations. The horizontal line shows the shape of the uniform distribution. It crosses the y-axis at 2, because we expect that each number (from 1 to 10) should occur about 2 times each in a sample of 20. However, as you can see, this does not happen. Instead, each sample bounces around quite a bit, due to random chance.

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

5.2.3 Large samples are more like the distribution they came from

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

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

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

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

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

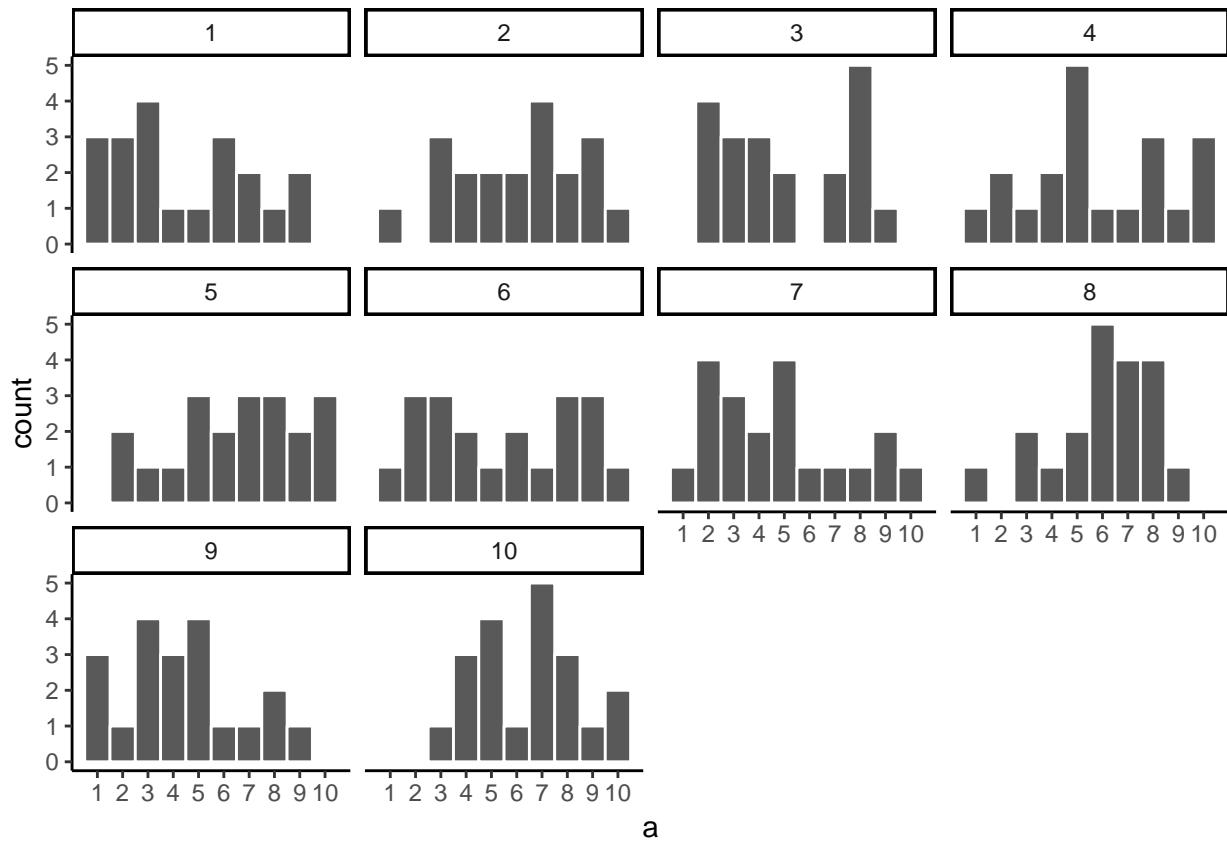


Figure 5.3: Histograms for 10 different samples from the uniform distribution. They all look quite different. The differences between the samples are due to sampling error

Animation not available in .pdf version

Figure 5.4: Animation of histograms for different samples of 20 from Uniform distribution (numbers 1 to 10). The black lines shows the expected number of times each number from 1 to 10 should occur. The fact that each number does not occur 2 times each illustrates the error associated with sampling

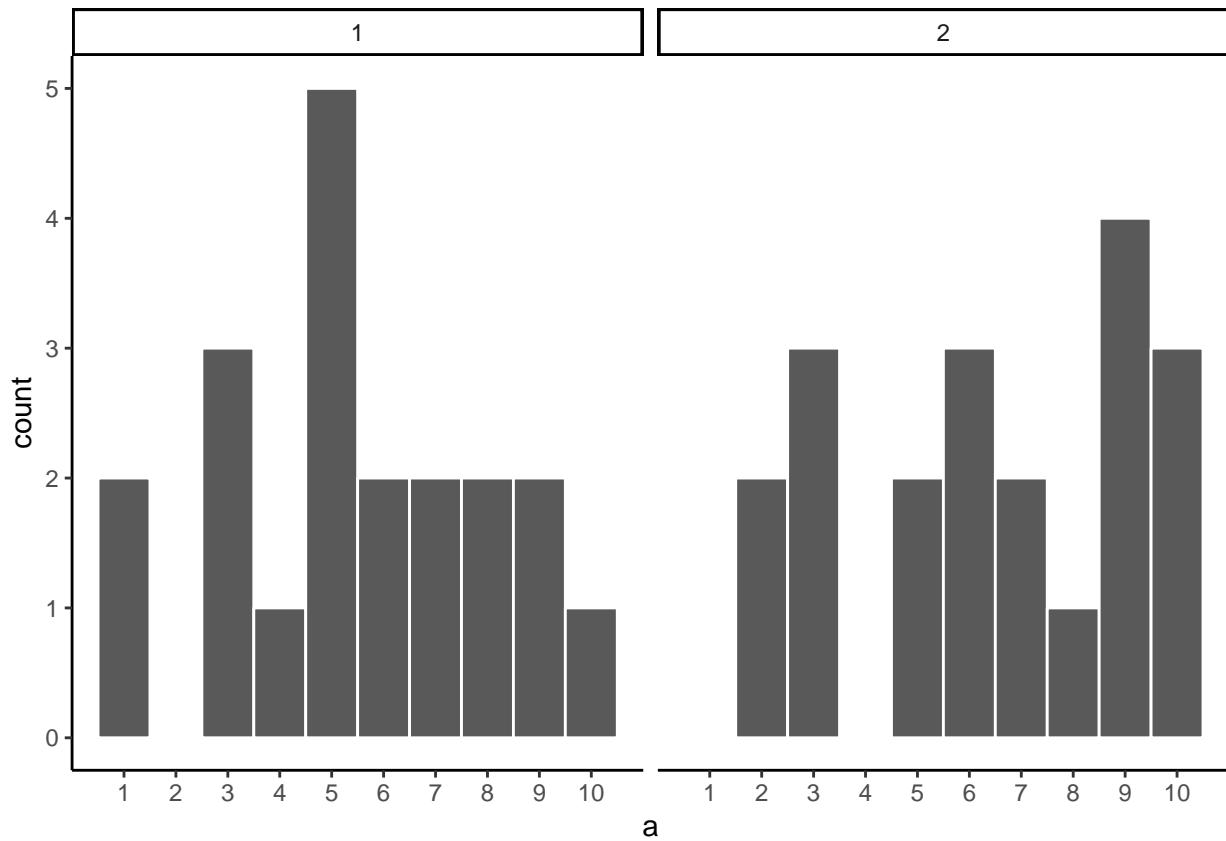


Figure 5.5: Which of these two samples came from a Uniformat distribution?

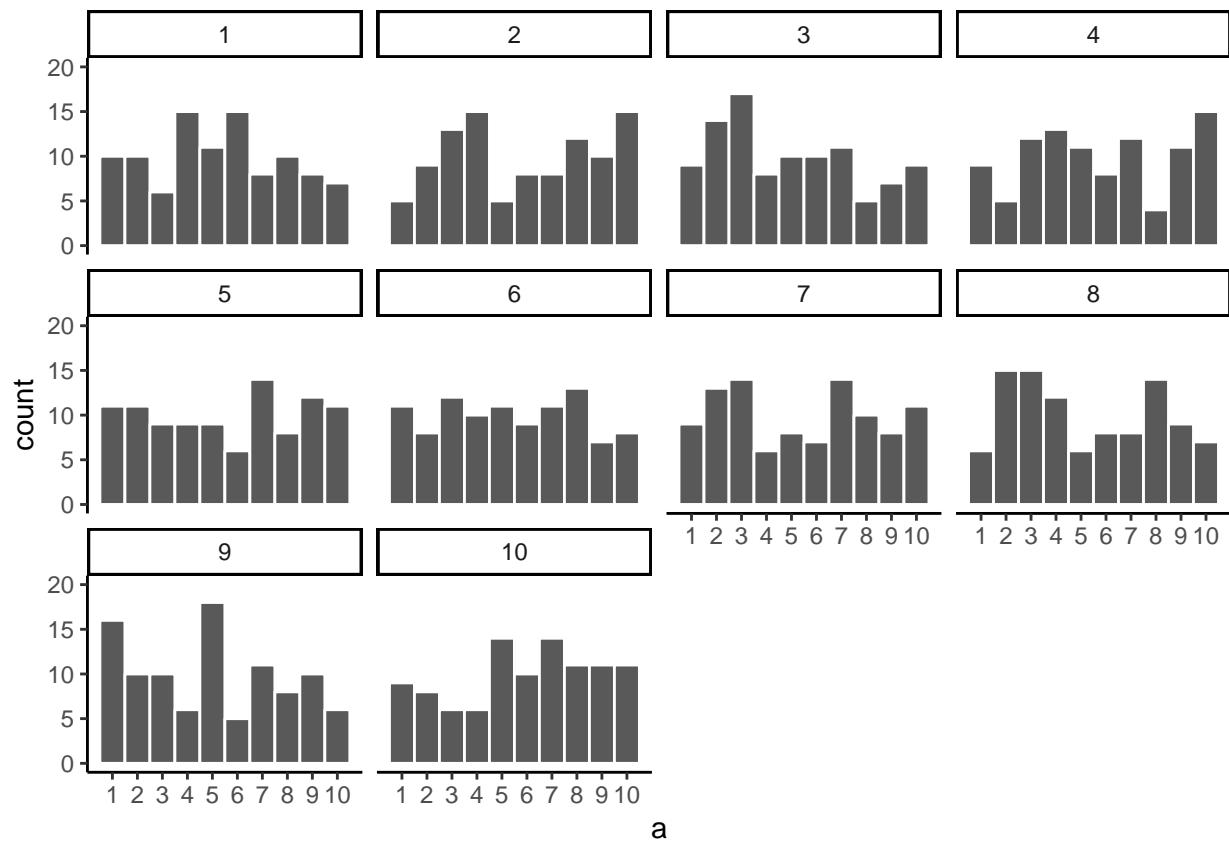


Figure 5.6: Histograms for different samples from a uniform distribution. Sample-size = 100 for each sample.

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

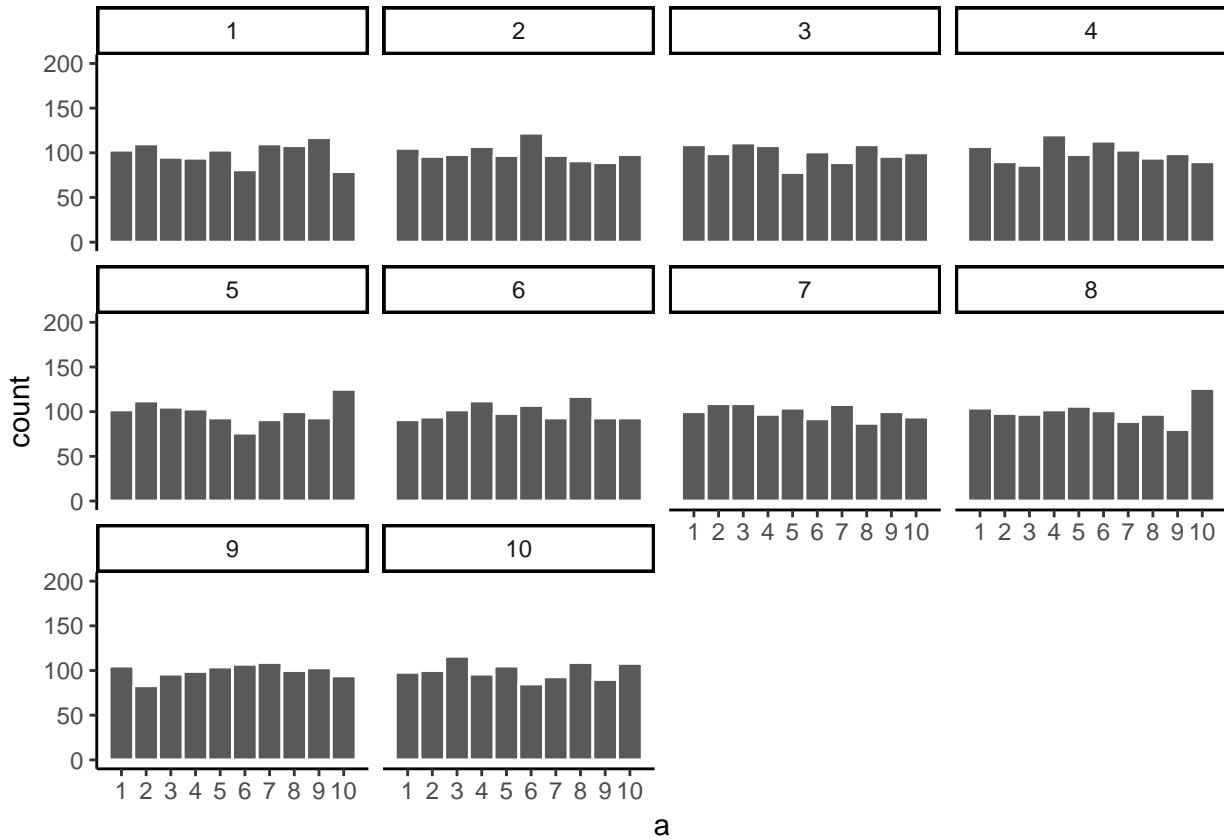


Figure 5.7: Histograms for different samples from a uniform distribution. Sample-size = 1000 for each sample.

Each of these histograms are starting to flatten out. The bars are still not perfectly at 100, because there is still sampling error (there always will be). But, if you found a histogram that looked flat and knew that the sample contained many observations, you might be more confident that those numbers came from a uniform distribution.

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

Now we see that all of our samples start to look the same. They all have 100,000 observations, and this gives chance enough opportunity to equally distribute the numbers, roughly making sure that they all occur very close to the same amount of times. As you can see, the bars are all very close to 10,000, where they should be if the sample came from a uniform distribution.

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

Before getting back to experiments, let's ask two more questions. First, which of these two samples do you think came from a uniform distribution? I will tell you that each of these samples had 20 observations each.

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

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

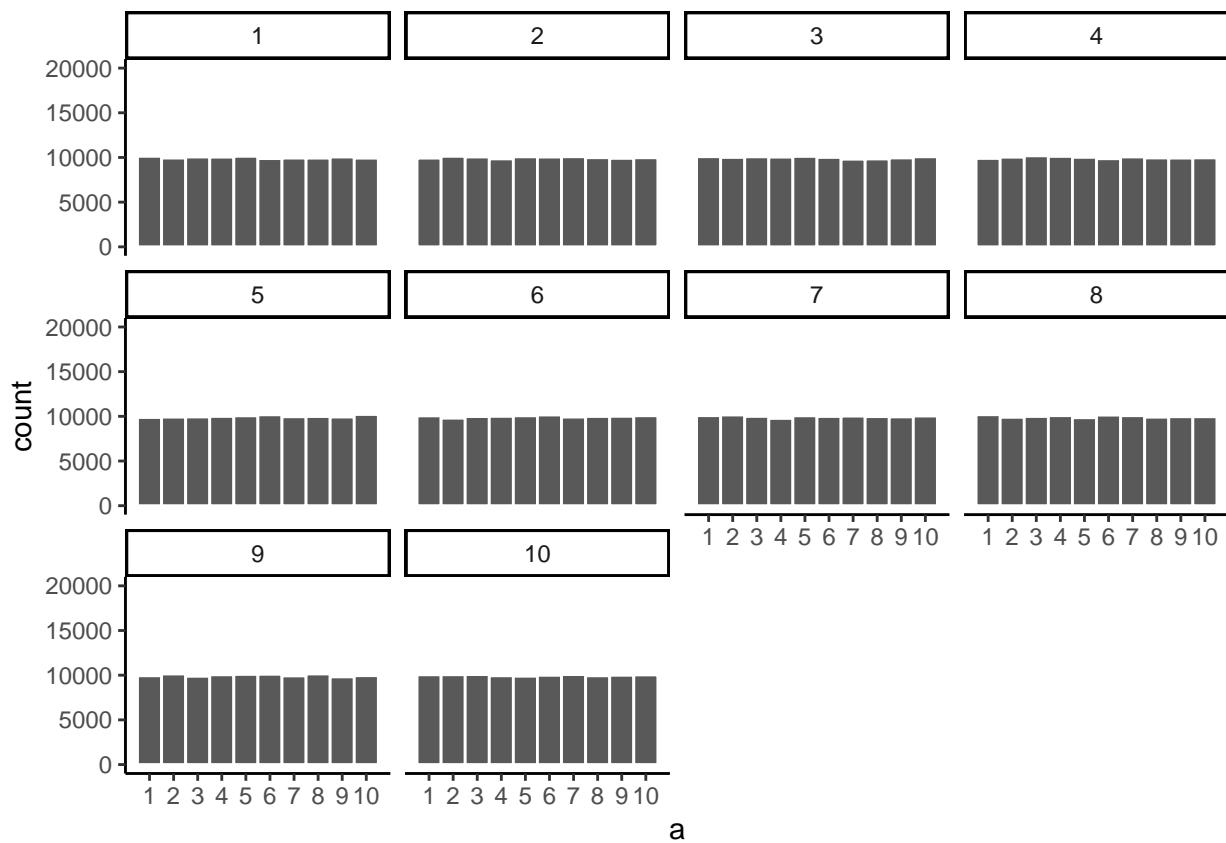


Figure 5.8: Histograms for different samples from a uniform distribution. Sample-size = 100,000 for each sample.

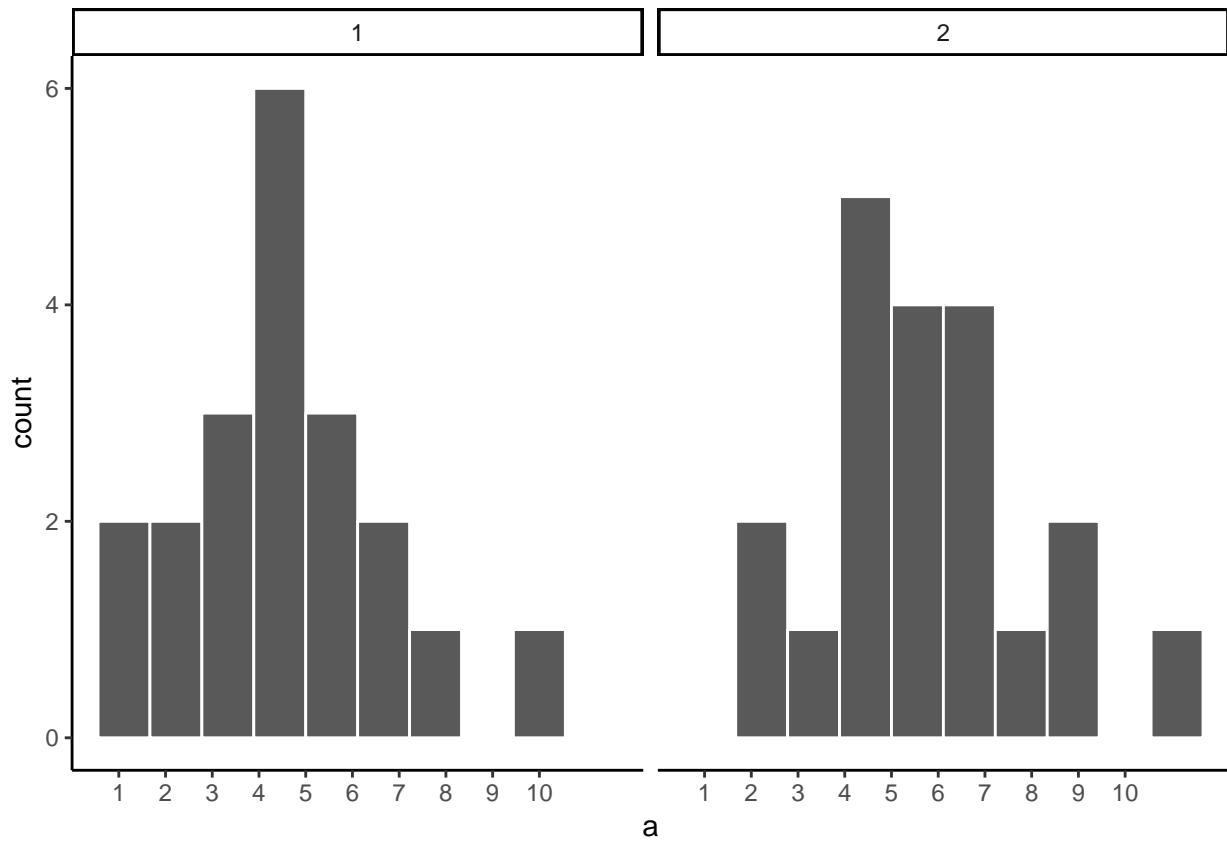


Figure 5.9: Which of these samples came from a uniform distribution?

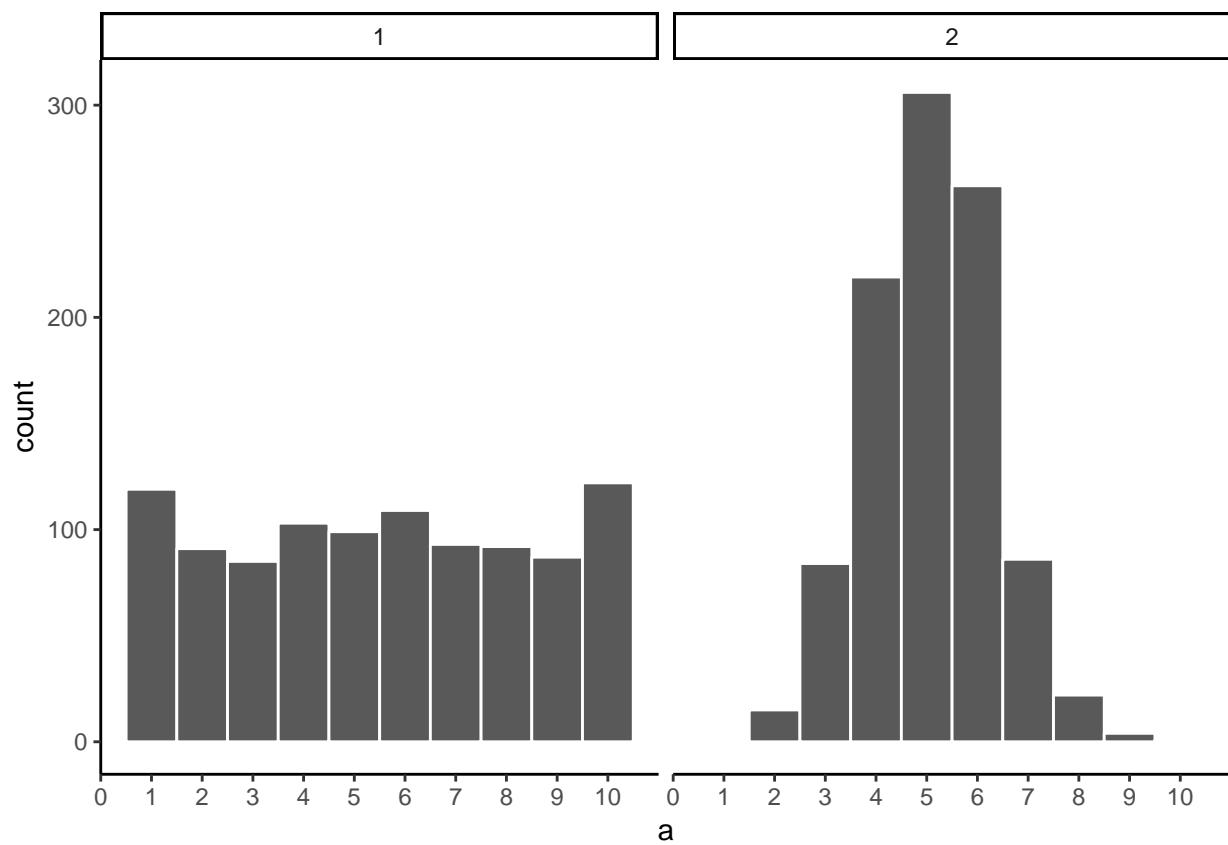


Figure 5.10: Which of these samples came from a uniform distribution?

Now that we have increased N, we can see the pattern in each sample becomes more obvious. The histogram for sample 1 has bars near 100, not perfectly flat, but it resembles a uniform distribution. The histogram for sample 2 does not look flat at all. Instead, there the number five appears most of the time, and numbers on either side of five happen less and less.

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

“We did?”. Yes, by looking at our two samples we have inferred that sample 2 did not come from a uniform distribution. We have also inferred that sample 1 could have come from a uniform distribution. Fantastic. This is really all we will be doing for the rest of the course. We will be looking at some numbers, wondering where they came from, then we will arrange the numbers in such a way so that we can make an inference about where they came from. That’s it.

5.3 Is there a difference?

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

Consider the light switch example:

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

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

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

5.3.1 Chance can produce differences

Do you think random chance can produce the appearance of differences, even when there really aren’t any? I hope so. We have already shown that the process of sampling numbers from a distribution is a chancy process that produces different samples. Different samples are different, so yes, chance can produce differences. This can muck up our interpretation of experiments.

Let’s conduct a fictitious experiment where we expect to find no differences, because we will manipulate something that shouldn’t do anything. Here’s the set-up:

You are the experimenter standing in front of a gumball machine. It is very big, has thousands of gumballs. 50% of the gumballs are green, and 50% are red. You want to find out if picking gumballs with your right hand vs. your left hand will cause you to pick more green gumballs. Plus, you will be blindfolded the entire time. The independent variable is Hand: right hand vs. left hand. The dependent variable is the measurement of the color of each gumball.

You run the experiment as follows. 1) put on blind fold. 2) pick 10 gumballs randomly with left hand, set them aside. 3) pick 10 gumballs randomly with right hand, set them aside. 4) count the number of green and red gumballs chosen by your left hand, and count the number of green and red gumballs chosen by your

right hand. Hopefully you will agree that your hands will not be able to tell the difference between the gumballs. If you don't agree, we will further stipulate the gumballs are completely identical in every way except their color, so it would be impossible to tell them apart using your hands. So, what should happen in this experiment?

"Umm, maybe you get 5 red gum balls and 5 green balls from your left hand, and also from your right hand?". Sort of yes, this is what you would usually get. But, it is not all that you can get. Here is some data showing what happened from one pretend experiment:

hand	gumball
left	1
left	0
left	0
left	1
left	0
left	1
left	1
left	1
left	0
left	1
right	0
right	1
right	1
right	0
right	0
right	0
right	1
right	1
right	0
right	0

"What am I looking at here". This is a long-format table. Each row is one gumball. The first column tells you what hand was used. The second column tells you what kind of gumball. We will say 1s stand for green gum balls, and 0s stand for red gumballs. So, did your left hand cause you to pick more green gumballs than your right hand?

It would be easier to look at the data using a bar graph. To keep things simple, we will only count green gumballs (the other gumballs must be red). So, all we need to do is sum up the 1s. The 0s won't add anything.

Oh look, the bars are not the same. One hand picked more green gum balls than the other. Does this mean that one of your hands secretly knows how to find green gumballs? No, it's just another case of sampling error, that thing we call luck or chance. The difference here is caused by chance, not by the manipulation (which hand you use). **Major problem for inference alert.** We run experiments to look for differences so we can make inferences about whether our manipulations cause change in our measures. Now we know that we can find differences by chance. How can we know if a difference is real, or just caused by chance?

5.3.2 Differences due to chance can be simulated

Remember when we showed that chance can produce correlations. We also showed that chance is restricted in its ability to produce correlations. For example, chance more often produces weak correlations than strong correlations. Remember the window of chance? We found out before that correlations falling outside the window of chance were very unlikely. We can do the same thing for differences. Let's find out just what chance can do in our experiment. Once we know what chance is capable of we will be in a better position to judge whether our manipulation caused a difference, or whether it could have been chance.

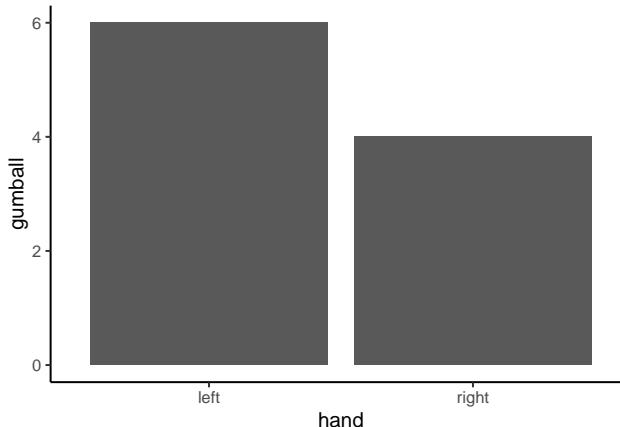


Figure 5.11: Counts of gumballs picked

The first thing to do is pretend you conduct the gumball experiment 10 times in a row. This will produce 10 different sets of results. For each of them we can make a bar graph, and look at whether the left hand chose more green gumballs than red gumballs. It looks like this:

These 10 experiments give us a better look at what chance can do. It should also mesh well with your expectations. If everything is left up to chance (as we have made it so), then sometimes your left hand will choose more green balls, sometimes your right hand will choose more green gumballs, and sometimes they will choose the same amount of gumballs. Right? Right.

5.4 Chance makes some differences more likely than others

OK, we have seen that chance can produce differences here. But, we still don't have a good idea about what chance usually does and doesn't do. For example, if we could find the window of opportunity here, we would be able find out that chance usually does not produce differences of a certain large size. If we knew what the size was, then if we ran experiment and our difference was bigger than what chance can do, we could be confident that chance did not produce our difference.

Let's use the word difference some more, because it will be helpful. In fact, let's think about our measure of green balls in terms of a difference. For example, in each experiment we counted the green balls for the left and right hand. What we really want to know is if there is a difference between them. So, we can calculate the **difference score**. Let's decide that difference score = # of green gumballs in left hand - # of green gumballs in right hand. Now, we can redraw the 10 bar graphs from above. But this time we will only see one bar for each experiment. This bar will show the difference in number of green gumballs.

Missing bars mean that there were an equal number of green gumballs chosen by the left and right hands (difference score is 0). A positive value means that more green gumballs were chosen by the left than right hand. A negative value means that more green gumballs were chosen by the right than left hand. Note that if we decided (and we get to decide) to calculate the difference in reverse (right hand - left hand), the signs of the differences scores would flip around.

We are starting to see more of the differences that chance can produce. The difference scores are mostly between -2 to +2. We could get an even better impression by running this pretend experiment 100 times instead of only 10 times. How about we do that.

Ooph, we just ran so many simulated experiments that the x-axis is unreadable, but it goes from 1 to 100. Each bar represents the difference of number of green balls chosen randomly by the left or right hand. Beginning to notice anything? Look at the y-axis, this shows the size of the difference. Yes, there are lots of bars of different sizes, this shows us that many kinds of differences do occur by chance. However, the y-axis

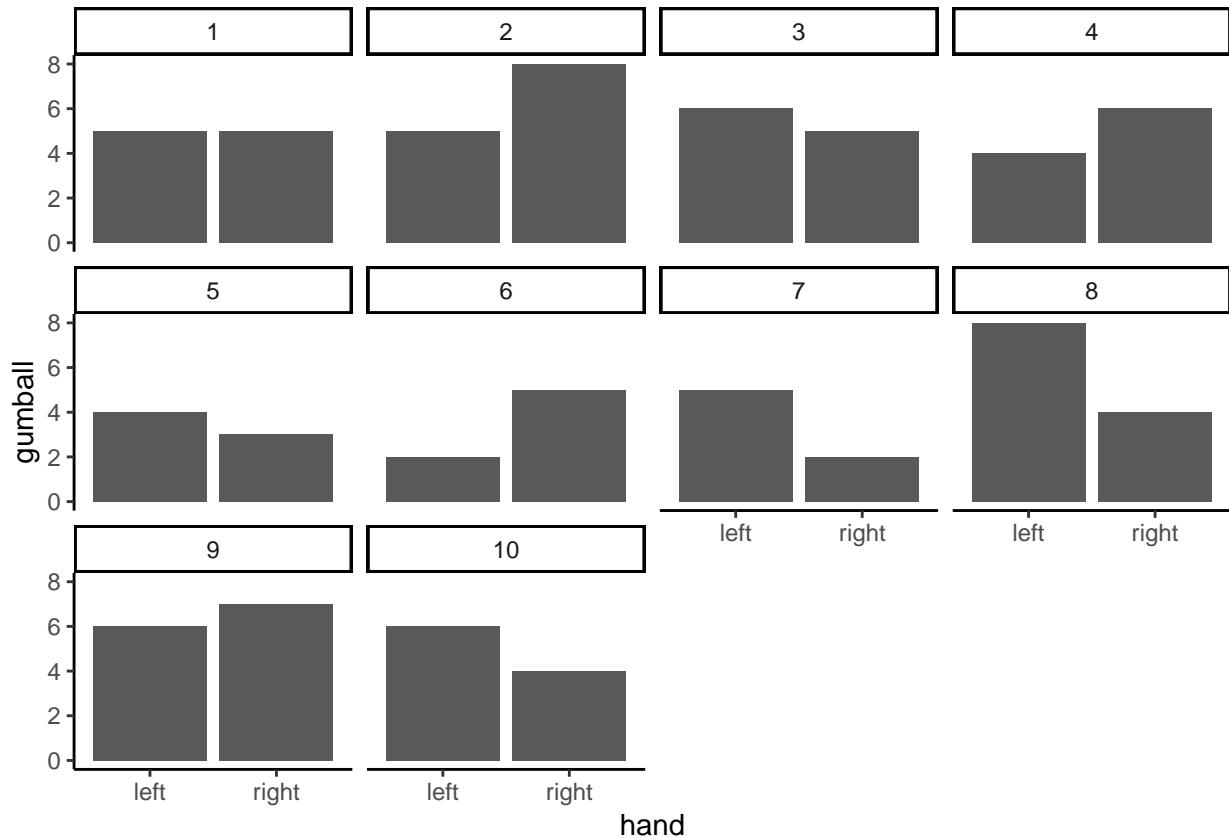


Figure 5.12: 10 simulated replications of picking gumballs. Each replication gives a slightly different answer. Any difference are all due to chance, or sampling error. This shows that chance alone can produce differences, just by the act of sampling

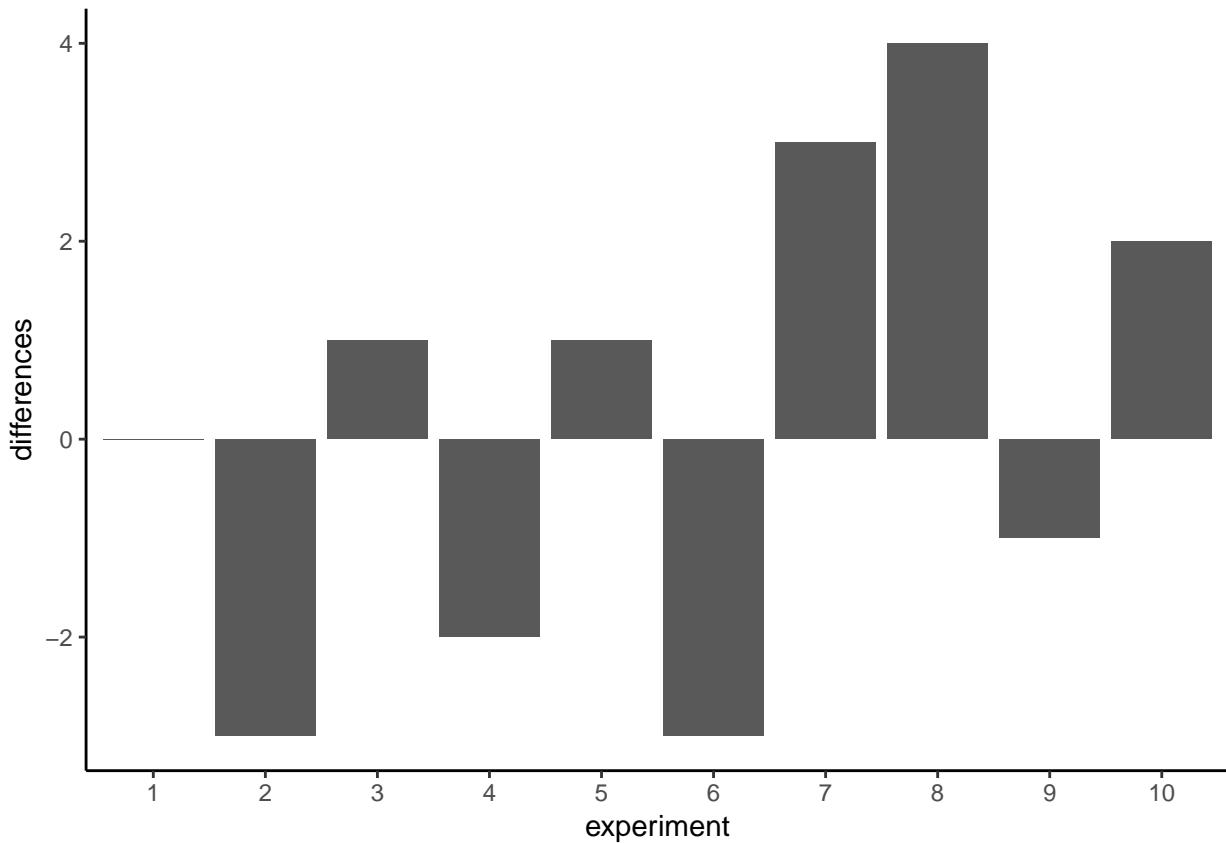


Figure 5.13: A look at the differences between number of each kind of gumball for the different replications. The difference should be zero, but sampling error produces non-zero differences

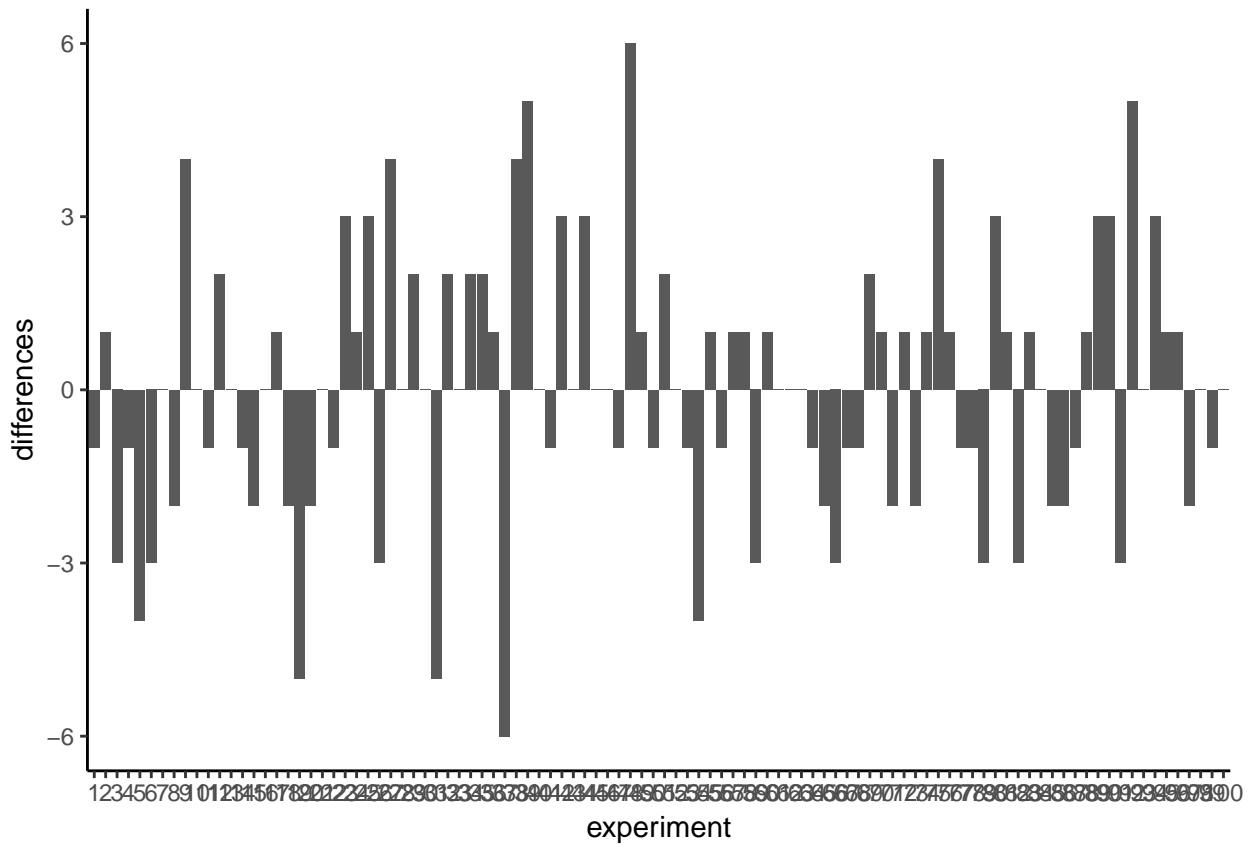


Figure 5.14: Replicating the sampling 100 times, and looking at the differences each time. There are many kinds of differences that chance alone can produce

is also restricted. It does not go from -10 to +10. Big differences greater than 5 or -5 don't happen very often.

Now that we have a method for simulating differences due to chance, let's run 10,000 simulated experiments. But, instead of plotting the differences in a bar graph for each experiment, how about we look at the histogram of difference scores. This will give us a clearer picture about which differences happen most often, and which ones do not. This will be another window into chance. The chance window of differences.

Histogram of differences

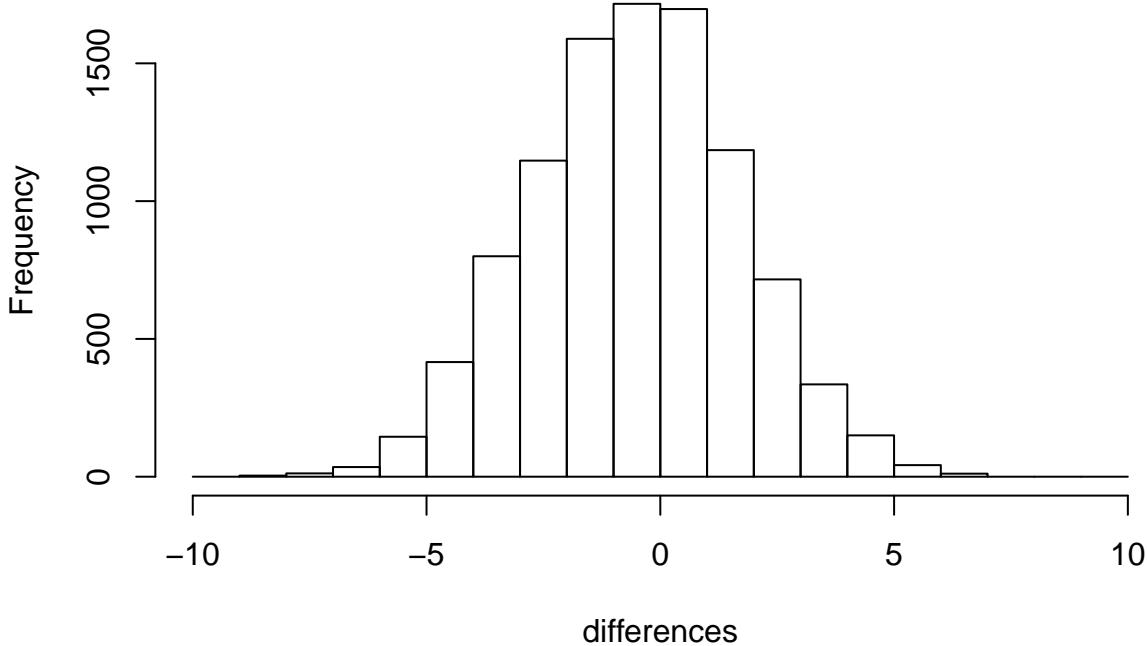


Figure 5.15: A histogram of the differences obtained by chance over 10,000 replications. The most frequent difference is 0, which is what we expect by chance. But the differences can be as large as -10 or +10. Larger differences occur less often by chance. Chance can't do everything

Our computer simulation allows us to force chance to operate hundreds of times, each time it produces a difference. We record the difference, then at the end of the simulation we plot the histogram of the differences. The histogram begins to show us where the differences came from. Remember the idea that numbers come from a distribution, and the distribution says how often each number occurs. We are looking at one of these distributions. It is showing us that chance produces some differences more often than others. First, chance usually produces 0 differences, that's the biggest bar in the middle. Chance also produces larger differences, but as the differences get larger (positive or negative), they occur less frequently. The shape of this histogram is your chance window, it tells you what chance can do, it tells you what chance usually does, and what it usually does not do.

You can use this chance window to help you make inferences. If you ran yourself in the gumball experiment and found that your left hand chose 2 more green gumballs than red gumballs, would you conclude that your left hand was special, and caused you to choose more green gumballs? Hopefully not. You could look at the chance window and see that differences of size +2 do happen fairly often by chance alone. You should not be surprised if you got a +2 difference. However, what if your left hand chose 5 more green gumballs than red gumballs. Well, chance doesn't do this very often, you might think something is up with your left hand. If you got a whopping 9 more green gumballs than red gumballs, you might really start to wonder. This is the kind of thing that could happen (it's possible), but virtually never happens by chance. When you get

things that almost never happen by chance, you can be more confident that the difference reflects a causal force that is not chance.

5.5 The Crump Test

We are going to be doing a lot of inference throughout the rest of this course. Pretty much all of it will come down to one question. Did chance produce the differences in my data? We will be talking about experiments mostly, and in experiments we want to know if our manipulation caused a difference in our measurement. But, we measure things that have natural variability, so every time we measure things we will always find a difference. We want to know if the difference we found (between our experimental conditions) could have been produced by chance. If chance is a very unlikely explanation of our observed difference, we will make the inference that chance did not produce the difference, and that something about our experimental manipulation did produce the difference. This is it (for this textbook).

Statistics is not only about determining whether chance could have produced a pattern in the observed data. The same tools we are talking about here can be generalized to ask whether any kind of distribution could have produced the differences. This allows comparisons between different models of the data, to see which one was the most likely, rather than just rejecting the unlikely ones (e.g., chance). But, we'll leave those advanced topics for another textbook.

This chapter is about building intuitions for making these kinds of inferences about the role of chance in your data. It's not clear to me what are the best things to say, to build up your intuitions for how to do statistical inference. So, this chapter tries different things, some of them standard, and some of them made up. What you are about to read, is a made up way of doing statistical inference, without using the jargon that we normally use to talk about it. The goal is to do things without formulas, and without probabilities, and just work with some ideas using simulations to see what happens. We will look at what chance can do, then we will talk about what needs to happen in your data in order for you to be confident that chance didn't do it.

5.5.1 Intuitive methods

Warning, this is an unofficial statistical test made up by Matt Crump. It makes sense to him (me), and if it turns out someone else already made this up, then Crump didn't do his homework, and we will change the name of this test to its original author. The point of this test is to show how simple operations that you already understand can be used to create a tool for inference. This test is not complicated, it uses

1. Sampling numbers randomly from a distribution
2. Adding, subtracting
3. Division, to find the mean
4. Counting
5. Graphing and drawing lines
6. NO FORMULAS

5.5.2 Part 1: Frequency based intuition about occurrence

Question: How many times does something need to happen, for it to happen a lot? Or, how many times does something need to happen for it to happen not very much, or even really not at all? Small enough for you to not worry about it at all happening to you?

Would you go outside everyday if you thought that you would get hit by lightning 1 out of 10 times? I wouldn't. You'd probably be hit by lightning more than once per month, you'd be dead pretty quickly. 1 out of 10 is a lot (to me, maybe not to you, there's no right answer here).

Would you go outside everyday if you thought that you would get hit by lightning 1 out of every 100 days? Jeez, that's a tough one. What would I even do? If I went out everyday, I'd probably be dead in a year! Maybe I would go out 2 or 3 times per year, I'm risky like that, but I'd probably live longer. It would massively suck.

Would you go outside everyday if you thought you would get hit by lightning 1 out of every 1000 days? Well, you'd probably be dead in 3-6 years if you did that. Are you a gambler? Maybe go out once per month, still sucks.

Would you go outside everyday if you thought lightning would get you 1 out every 10,000 days? 10,000 is a bigger number, harder to think about. It's about once every 27 years. Ya, I'd probably go out 150 days per year, and live a bit longer if I can.

Would you go outside everyday if you thought lightning would get you 1 out every 100,000 days? 100,000 is a bigger number, harder to think about. How many years is that? It's about 273 years. With those odds, I'd probably go out all the time and forget about being hit by lightning. It doesn't happen very often, and if it does, c'est la vie.

The point of considering these questions is to get a sense for yourself of what happens a lot, and what doesn't happen a lot, and how you would make important decisions based on what happens a lot and what doesn't.

5.5.3 Part 2: Simulating chance

This next part could happen a bunch of ways, I'll make loads of assumptions that I won't defend, and I won't claim the Crump test has problems. I will claim it helps us make an inference about whether chance could have produced some differences in data. We've already been introduced to simulating things, so we'll do that again. Here is what we will do. I am a cognitive psychologist who happens to be measuring X. Because of prior research in the field, I know that when I measure X, my samples will tend to have a particular mean and standard deviation. Let's say the mean is usually 100, and the standard deviation is usually 15. In this case, I don't care about using these numbers as estimates of the population parameters, I'm just thinking about what my samples usually look like. What I want to know is how they behave when I sample them. I want to see what kind of samples happen a lot, and what kind of samples don't happen a lot. Now, I also live in the real world, and in the real world when I run experiments to see what changes X, I usually only have access to some number of participants, who I am very grateful too, because they participate in my experiments. Let's say I usually can run 20 subjects in each condition in my experiments. Let's keep the experiment simple, with two conditions, so I will need 40 total subjects.

I would like to learn something to help me with inference. One thing I would like to learn is what the sampling distribution of the sample mean looks like. This distribution tells me what kinds of mean values happen a lot, and what kinds don't happen very often. But, I'm actually going to skip that bit. Because what I'm really interested in is what the **sampling distribution of the difference between my sample means** looks like. After all, I am going to run an experiment with 20 people in one condition, and 20 people in the other. Then I am going to calculate the mean for group A, and the mean for group B, and I'm going to look at the difference. I will probably find a difference, but my question is, did my manipulation cause this difference, or is this the kind of thing that happens a lot by chance. If I knew what chance can do, and how often it produces differences of particular sizes, I could look at the difference I observed, then look at what chance can do, and then I can make a decision! If my difference doesn't happen a lot (we'll get to how much not a lot is in a bit), then I might be willing to believe that my manipulation caused a difference. If my difference happens all the time by chance alone, then I wouldn't be inclined to think my manipulation caused the difference, because it could have been chance.

So, here's what we'll do, even before running the experiment. We'll do a simulation. We will sample numbers for group A and Group B, then compute the means for group A and group B, then we will find the difference in the means between group A and group B. But, we will do one very important thing. We will pretend that we haven't actually done a manipulation. If we do this (do nothing, no manipulation that could cause a difference), then we know that **only sampling error** could cause any differences between the mean of

group A and group B. We've eliminated all other causes, only chance is left. By doing this, we will be able to see exactly what chance can do. More importantly, we will see the kinds of differences that occur a lot, and the kinds that don't occur a lot.

Before we do the simulation, we need to answer one question. How much is a lot? We could pick any number for a lot. I'm going to pick 10,000. That is a lot. If something happens only 1 times out 10,000, I am willing to say that is not a lot.

OK, now we have our number, we are going to simulate the possible mean differences between group A and group B that could arise by chance. We do this 10,000 times. This gives chance a lot of opportunity to show us what it does do, and what it does not do.

This is what I did: I sampled 20 numbers into group A, and 20 into group B. The numbers both came from the same normal distribution, with mean = 100, and standard deviation = 15. Because the samples are coming from the same distribution, we expect that on average they will be similar (but we already know that samples differ from one another). Then, I compute the mean for each sample, and compute the difference between the means. I save the **mean difference score**, and end up with 10,000 of them. Then I draw a histogram. It looks like this:

Histogram of mean differences between two samples (n=10)

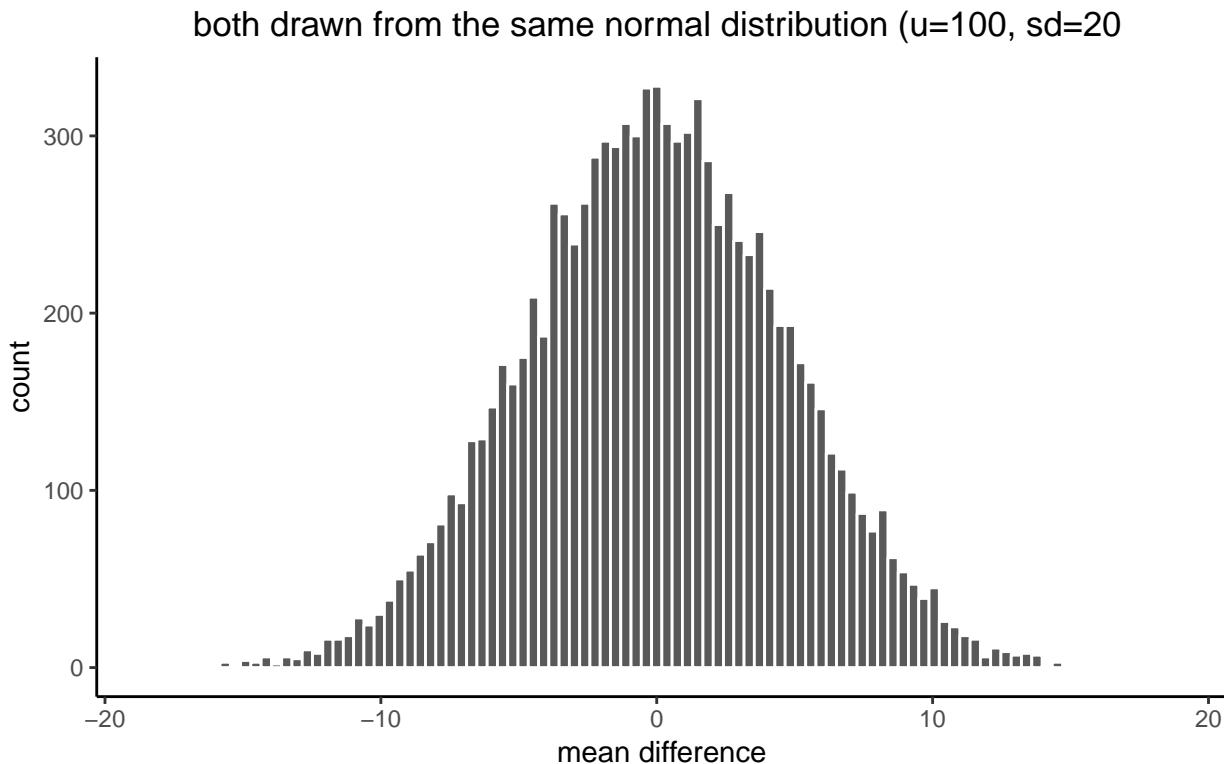


Figure 5.16: Histogram of mean differences arising by chance

Sidenote: Of course, we might recognize that chance could do a difference greater than 15. We just didn't give it the opportunity. We only ran the simulation 10,000 times. If we ran it on million times, maybe a difference greater than 20 would happen a couple times. If we ran it a bazillion gazillion times, maybe a difference greater than 30 would happen a couple times. If we go out to infinity, then chance might produce all sorts of bigger differences once in a while. But, we've already decided that 1/10,000 is not a lot. So things that happen 0/10,000 times, like differences greater than 15, just don't happen very much.

Now we can see what chance can do to the size of our mean difference. The x-axis shows the size of the

mean difference. We took our samples from the sample distribution, so the difference between them should usually be 0, and that's what we see in the histogram.

Pause for a second. Why should the mean differences usually be zero, wasn't the population mean = 100, shouldn't they be around 100? No. The mean of group A will tend to be around 100, and the mean of group B will tend to be around 100. So, the difference score will tend to be $100 - 100 = 0$. That is why we expect a mean difference of zero when the samples are drawn from the same population.

So, differences near zero happen the most, that's good, that's what we expect. Bigger or smaller differences happen increasingly less often. Differences greater than 15 or -15 never happen at all. For our purposes, it looks like chance only produces differences between -15 to 15.

OK, let's ask a couple simple questions. What was the biggest negative number that occurred in the simulation? We'll use R for this. All of the 10,000 difference scores are stored in a variable I made called **difference**. If we want to find the minimum value, we use the **min** function. Here's the result.

```
min(difference)
```

```
## [1] -18.0828
```

OK, so what was the biggest positive number that occurred? Let's use the **max** function to find out. It finds the biggest (maximum) value in the variable. FYI, we've just computed the range, the minimum and maximum numbers in the data. Remember we learned that before. Anyway, here's the max.

```
max(difference)
```

```
## [1] 18.78974
```

Both of these extreme values only occurred once. Those values were so rare we couldn't even see them on the histogram, the bar was so small. Also, these biggest negative and positive numbers are pretty much the same size if you ignore their sign, which makes sense because the distribution looks roughly symmetrical.

So, what can we say about these two numbers for the min and max? We can say the min happens 1 times out of 10,000. We can say the max happens 1 times out of 10,000. Is that a lot of times? Not to me. It's not a lot.

So, how often does a difference of 30 (much larger than the max) occur out of 10,000. We really can't say, 30s didn't occur in the simulation. Going with what we got, we say 0 out of 10,000. That's never.

We're about to move into part three, which involves drawing decision lines and talking about them. The really important part about part 3 is this. What would you say if you ran this experiment once, and found a mean difference of 30? I would say it happens 0 times out of 10,000 by chance. I would say chance did not produce my difference of 30. That's what I would say. We're going to expand upon this right now.

5.5.4 Part 3: Judgment and Decision-making

Remember, we haven't even conducted an experiment. We're just simulating what could happen if we did conduct an experiment. We made a histogram. We can see that chance produces some differences more than others, and that chance never produced really big differences. What should we do with this information?

What we are going to do is talk about judgment and decision making. What kind of judgment and decision making? Well, when you finally do run an experiment, you will get two means for group A and B, and then you will need to make some judgments, and perhaps even a decision, if you are so inclined. You will need to judge whether chance (sampling error) could have produced the difference you observed. If you judge that it did not, you might make the decision to tell people that your experimental manipulation actually works. If you judge that it could have been chance, you might make a different decision. These are important decisions for researchers. Their careers can depend on them. Also, their decisions matter for the public. Nobody wants to hear fake news from the media about scientific findings.

So, what we are doing is preparing to make those judgments. We are going to draw up a plan, before we even see the data, for how we will make judgments and decisions about what we find. This kind of planning is extremely important, because we discuss in part 4, that your planning can help you design an even better experiment than the one you might have been intending to run. This kind of planning can also be used to interpret other people's results, as a way of double-checking whether you believe those results are plausible.

The thing about judgement and decision making is that reasonable people disagree about how to do it, unreasonable people really disagree about it, and statisticians and researchers disagree about how to do it. I will propose some things that people will disagree with. That's OK, these things still make sense. And, the disagreeable things point to important problems that are very real for any "real" statistical inference test.

Let's talk about some objective facts from our simulation of 10,000 things that we definitely know to be true. For example, we can draw some lines on the graph, and label some different regions. We'll talk about two kinds of regions.

1. Region of chance. Chance did it. Chance could have done it
2. Region of not chance. Chance didn't do it. Chance couldn't have done it.

The regions are defined by the minimum value and the maximum value. Chance never produced a smaller or bigger number. The region inside the range is what chance did do, and the the region outside the range on both sides is what chance never did. It looks like this:

Histogram of mean differences between two samples (n=10)

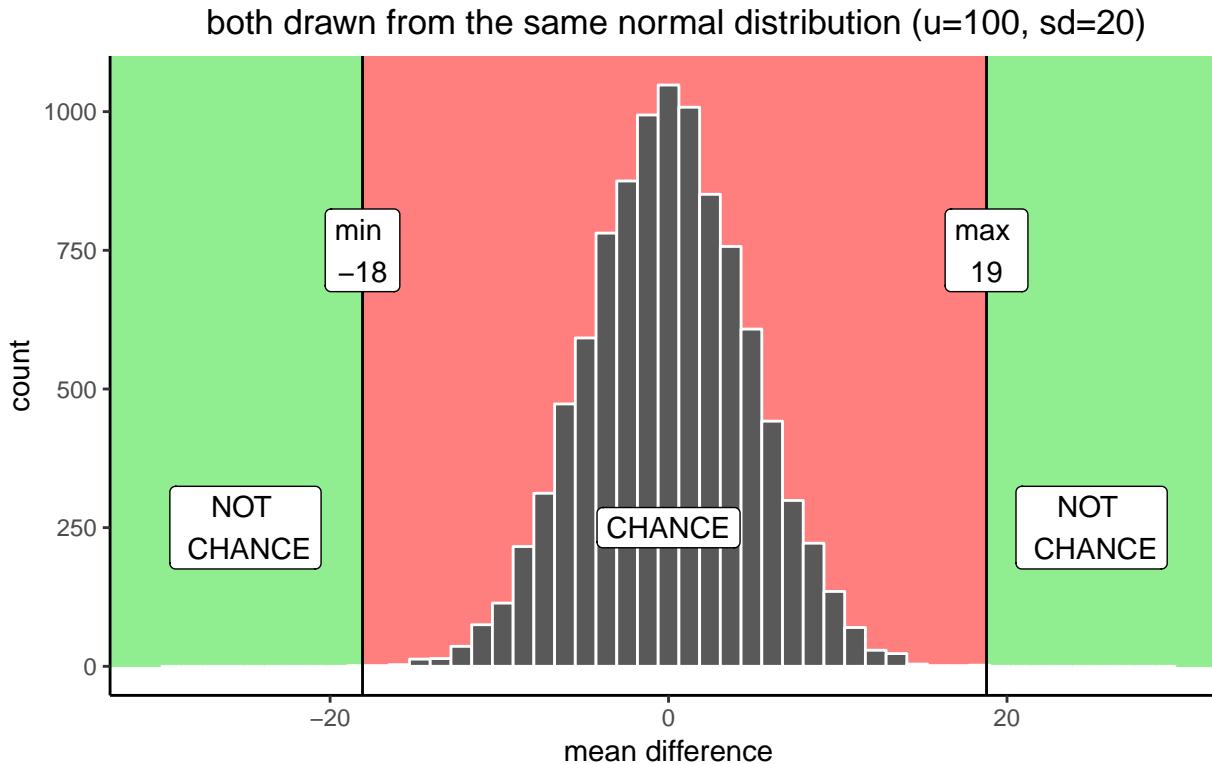


Figure 5.17: Applying decision boundaries to the histogram of mean differences. The boundaries identify what differences chance did or did not produce in the simulation

We have just drawn some lines, and shaded some regions, and made one plan we could use to make decisions. How would the decisions work. Let's say you ran the experiment and found a mean difference between groups A and B of 25. Where is 25 in the figure? It's in the green part. What does the green part say? NOT

CHANCE. What does this mean. It means chance never made a difference of 25. It did that 0 out of 10,000 times. If we found a difference of 25, perhaps we could confidently conclude that chance did not cause the difference. If I found a difference of 25 with this kind of data, I'd be pretty confident that my experimental manipulation caused the difference, because obviously chance never does.

What about a difference of +10? That's in the red part, where chance lives. Chance could have done a difference of +10 because we can see that it did do that. The red part is the window of what chance did in our simulation. Anything inside the window could have been a difference caused by chance. If I found a difference of +10, I'd say, coulda been chance. I would not be very confident that my experimental manipulation caused the difference.

Statistical inference could be this easy. The number you get from your experiment could be in the chance window (then you can't rule out chance as a cause), or it could be outside the chance window (then you can rule out chance). Case closed. Let's all go home.

5.5.4.1 Grey areas

So what's the problem? Depending on who you are, and what kinds of risks you're willing to take, there might not be a problem. But, if you are just even a little bit risky then there is a problem that makes clear judgments about the role of chance difficult. We would like to say chance did or did not cause our difference. But, we're really always in the position of admitting that it could have sometimes, or wouldn't have most times. These are wishy washy statements, they are in between yes or no. That's OK. Grey is a color too, let's give grey some respect.

"What grey areas are you talking about?, I only see red or green, am I grey blind?". Let's look at where some grey areas might be. I say might be, because people disagree about where the grey is. People have different comfort levels with grey. Here's my opinion on some clear grey areas.

I made two grey areas, and they are reddish grey, because we are still in the chance window. There are question marks (?) in the grey areas. Why? The question marks reflect some uncertainty that we have about those particular differences. For example, if you found a difference that was in a grey area, say a 15. 15 is less than the maximum, which means chance did create differences of around 15. But, differences of 15 don't happen very often.

What can you conclude or say about this 15 you found? Can you say without a doubt that chance did not produce the difference? Of course not, you know that chance could have. Still, it's one of those things that doesn't happen a lot. That makes chance an unlikely explanation. Instead of thinking that chance did it, you might be willing to take a risk and say that your experimental manipulation caused the difference. You'd be making a bet that it wasn't chance...but, could be a safe bet, since you know the odds are in your favor.

You might be thinking that your grey areas aren't the same as the ones I've drawn. Maybe you want to be more conservative, and make them smaller. Or, maybe you're more risky, and would make them bigger. Or, maybe you'd add some grey area going in a little bit to the green area (after all, chance could probably produce some bigger differences sometimes, and to avoid those you would have to make the grey area go a bit into the green area).

Another thing to think about is your decision policy. What will you do, when your observed difference is in your grey area? Will you always make the same decision about the role of chance? Or, will you sometimes flip-flop depending on how you feel. Perhaps, you think that there shouldn't be a strict policy, and that you should accept some level of uncertainty. The difference you found could be a real one, or it might not. There's uncertainty, hard to avoid that.

So let's illustrate one more kind of strategy for making decisions. We just talked about one that had some lines, and some regions. This makes it seem like we can either rule out, or not rule out the role of chance. Another way of looking at things is that everything is a different shade of grey. It looks like this:

OK, so I made it shades of blue (because it was easier in R). Now we can see two decision plans at the same time. Notice that as the bars get shorter, they also get become a darker stronger blue. The color can be

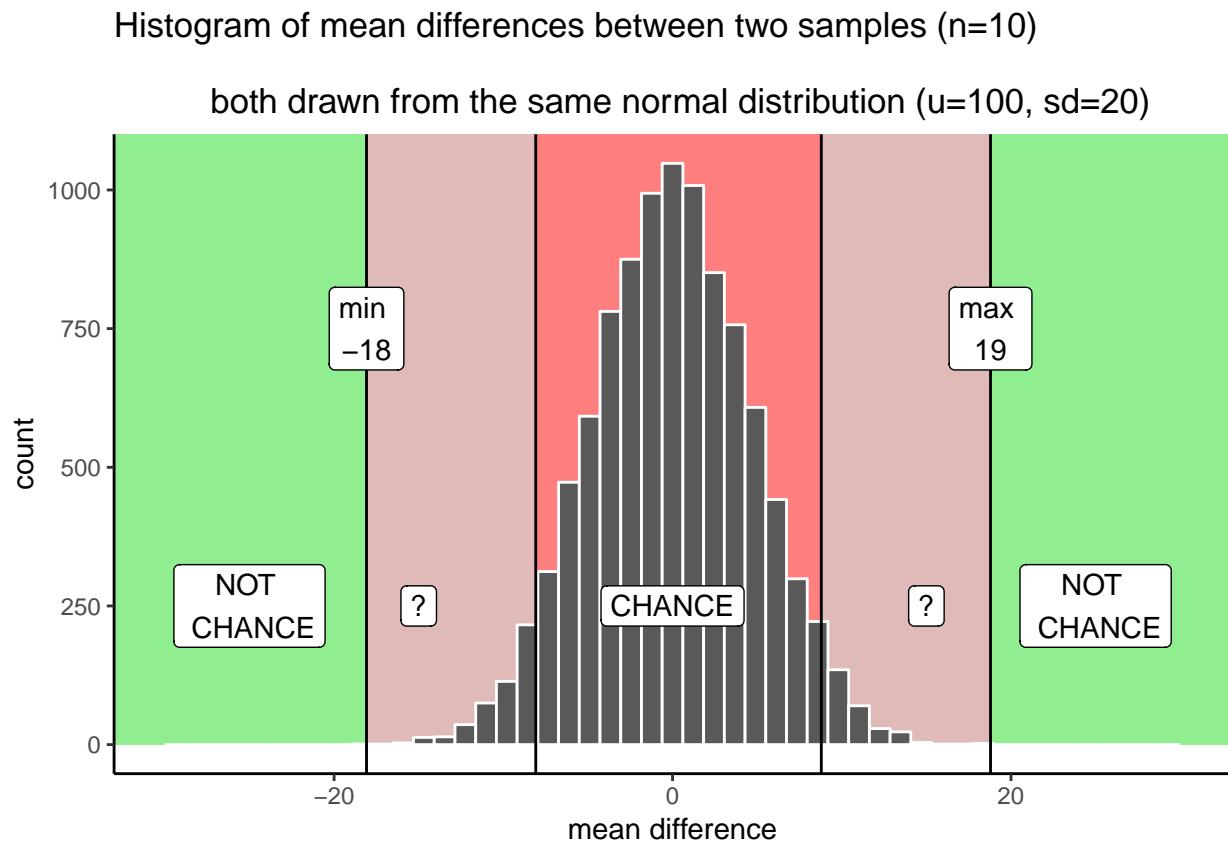


Figure 5.18: The question marks refer to an area where you have some uncertainty. Differences inside the question mark region do not happen very often by chance. When you find differences of these sizes, should you reject the idea that chance caused your difference? You will always have some uncertainty associated with this decision because it is clear that chance could have caused the difference. But, chance usually does not produce differences of these sizes

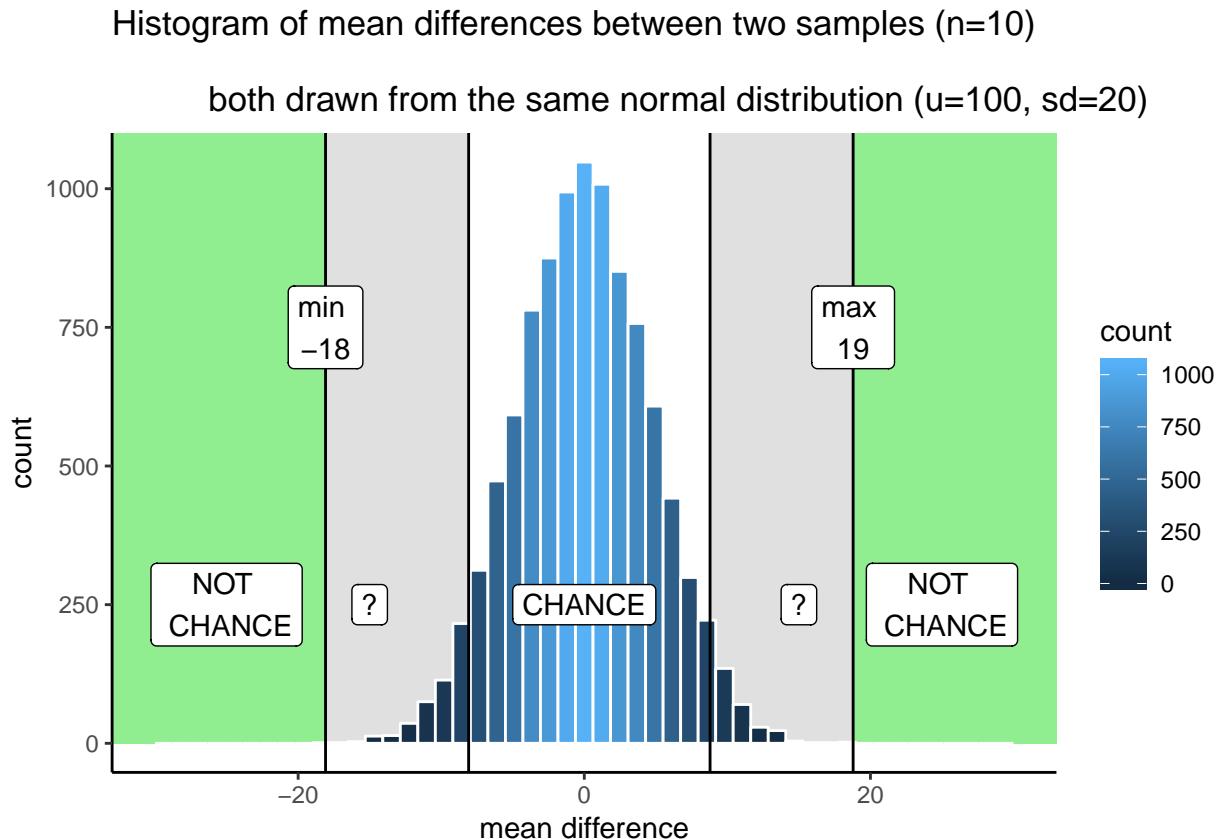


Figure 5.19: The shading of the blue bars indicates levels of confidence in whether a difference could have been produced by chance. Darker bars represent increased confidence that the difference was not produced by chance. Bars get darker as the mean difference increases in absolute value

used as a guide for your confidence. That is, your confidence in the belief that your manipulation caused the difference rather than chance. If you found a difference near a really dark bar, those don't happen often by chance, so you might be really confident that chance didn't do it. If you find a difference near a slightly lighter blue bar, you might be slightly less confident. That is all. You run your experiment, you get your data, then you have some amount of confidence that it wasn't produced by chance. This way of thinking is elaborated to very interesting degrees in the Bayesian world of statistics. We don't wade too much into that, but mention it a little bit here and there. It's worth knowing it's out there.

5.5.4.2 Making Bad Decisions

No matter how you plan to make decisions about your data, you will always be prone to making some mistakes. You might call one finding real, when in fact it was caused by chance. This is called a **type I** error, or a false positive. You might ignore one finding, calling it chance, when in fact it wasn't chance (even though it was in the window). This is called a ** type II**, or a false negative.

How you make decisions can influence how often you make errors over time. If you are a researcher, you will run lots of experiments, and you will make some amount of mistakes over time. If you do something like the very strict method of only accepting results as real when they are in the "no chance" zone, then you won't make many type I errors. Pretty much all of your result will be real. But, you'll also make type II errors, because you will miss things real things that your decision criteria says are due to chance. The opposite also holds. If you are willing to be more liberal, and accept results in the grey as real, then you will make more type I errors, but you won't make as many type II errors. Under the decision strategy of using these cutoff regions for decision-making there is a necessary trade-off. The Bayesian view gets around this a little bit. Bayesians talk about updating their beliefs and confidence over time. In that view, all you ever have is some level of confidence about whether something is real, and by running more experiments you can increase or decrease your level of confidence. This, in some fashion, avoids some trade-off between type I and type II errors.

Regardless, there is another way to avoid type I and type II errors, and to increase your confidence in your results, even before you do the experiment. It's called "knowing how to design a good experiment".

5.5.5 Part 4: Experiment Design

We've seen what chance can do. Now we run an experiment. We manipulate something between groups A and B, get the data, calculate the group means, then look at the difference. Then we cross all of our fingers and toes, and hope beyond hope that the difference is big enough to not be caused by chance. That's a lot of hope.

Here's the thing, we don't often know how strong our manipulation is in the first place. So, even if it can cause a change, we don't necessarily know how much change it can cause. That's why we're running the experiment. Many manipulations in Psychology are not strong enough to cause big changes. This is a problem for detecting these smallish causal forces. In our fake example, you could easily manipulate something that has a tiny influence, and will never push the mean difference past say 5 or 10. In our simulation, we need something more like a 15 or 17 or a 21, or hey, a 30 would be great, chance never does that. Let's say your manipulation is listening to music or not listening to music. Music listening might change something about X, but if it only changes X by +5, you'll never be able to confidently say it wasn't chance. And, it's not that easy to completely change music and make music super strong in the music condition so it really causes a change in X compared to the no music condition.

EXPERIMENT DESIGN TO THE RESCUE! Newsflash, it is often possible to change how you run your experiment so that it is **more sensitive** to smaller effects. How do you think we can do this? Here is a hint. It's the stuff you learned about the sampling distribution of the sample mean, and the role of sample-size. What happens to the sampling distribution of the sample mean when N (sample size)? The distribution gets narrower and narrower, and starts to look like a single number (the hypothetical mean of the hypothetical

population). That's great. If you switch to thinking about mean difference scores, like the distribution we created in this test, what do you think will happen to that distribution as we increase N ? It will shrink. As we increase N to infinity, it will shrink to 0. Which means that, when N is infinity, chance never produces any differences at all. We can use this.

For example, we could run our experiment with 20 subjects in each group. Or, we could decide to invest more time and run 40 subjects in each group, or 80, or 160. When you are the experimenter, you get to decide the design. These decisions matter big time. Basically, the more subjects you have, the more sensitive your experiment. With bigger N , you will be able to reliably detect smaller mean differences, and be able to confidently conclude that chance did not produce those small effects.

Check out this next set of histograms. All we are doing is the very same simulation as before, but this time we do it for different sample-sizes: 20, 40, 80, 160. We are doubling our sample-size across each simulation just to see what happens to the width of the chance window.

Sampling distribution of mean differences

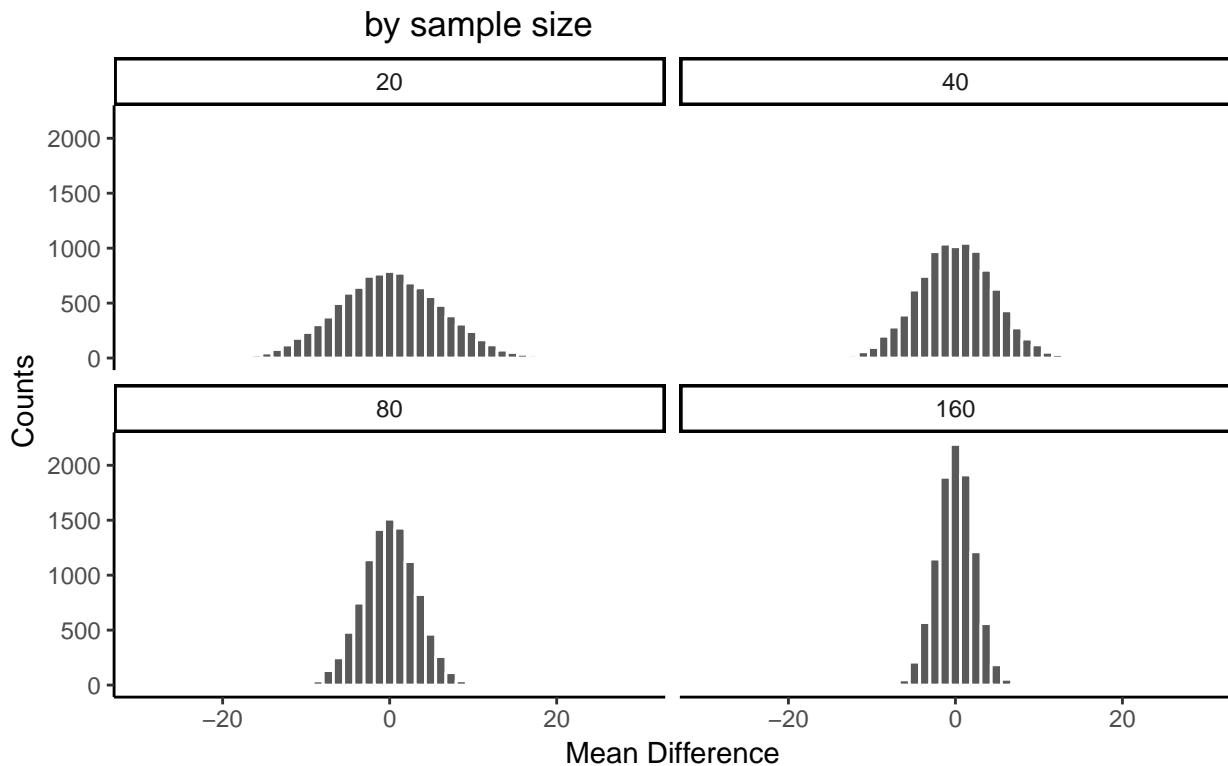


Figure 5.20: The range or width of the differences produced by chance shrinks as sample-size increases.

There you have it. The **sampling distribution of the mean differences** shrinks toward 0 as sample-size increases. This means if you run an experiment with a larger sample-size, you will be able to detect smaller mean differences, and be confident they aren't due to chance. Let's look at a table of the minimum and maximum values that chance produced across these four sample-sizes:

sample_size	smallest	biggest
20	-24.137128	26.80421
40	-17.462085	16.24287
80	-11.058030	12.14587
160	-7.782496	8.52066

The table is telling... The range of chance's behavior is very wide for sample-size = 20, but about half as

wide for sample-size = 160.

If it turns out your manipulation will cause a difference of +11, then what should you do? Run an experiment with 20 people? I hope not. If you did that, you could get +11s fairly often by chance. If you ran the experiment with 160 people, then you would definitely be able to say that +11 was not due to chance, it would be outside the range of what chance can do. You could even consider running the experiment with 80 subjects. A +11 there wouldn't happen often by chance, and you'd be cost-effective, spending less time on the experiment.

The point is: **the design of the experiment determines the sizes of the effects it can detect.** If you want to detect a small effect. Make your sample size bigger. It's really important to say this is not the only thing you can do. You can also make your cell-sizes bigger. For example, often times we take several measurements from a single subject. The more measurements you take (cell-size), the more stable your estimate of the subject's mean. We discuss these issues more later. You can also make a stronger manipulation, when possible.

5.5.6 Part 5: I have the power

By the power of greyskull, I HAVE THE POWER - He-man

The last thing we'll talk about here is something called power. In fact, we are going to talk about the concept of power, not actual power. It's confusing now, but later we will define power in terms of some particular ideas about statistical inference. Here, we will just talk about the idea. And, we'll show how to make sure your design has 100% power. Because, why not. Why run a design that doesn't have the power?

The big idea behind power is the concept of sensitivity. The concept of sensitivity assumes that there is something to be sensitive to. That is, there is some real difference that can be measured. So, the question is, how sensitive is your experiment? We've already seen that the number of subjects (sample-size), changes the sensitivity of the design. More subjects = more sensitivity to smaller effects.

Let's take a look at one more plot. What we will do is simulate a measure of sensitivity across a whole bunch of sample sizes, from 10 to 300. We'll do this in steps of 10. For each simulation, we'll compute the mean differences as we have done. But, rather than showing the histogram, we'll just compute the smallest value and the largest value. This is a pretty good measure of the outer reach of chance. Then we'll plot those values as a function of sample size and see what we've got.

What we have here is a reasonably precise window of sensitivity as a function of sample size. For each sample size, we can see the maximum difference that chance produced and the minimum difference. In those simulations, chance never produced bigger or smaller differences. So, each design is sensitive to any difference that is underneath the bottom line, or above the top line. It's really that simple.

Here's another way of putting it. Which of the sample sizes will be sensitive to a difference of +10 or -10. That is, if a difference of +10 or -10 was observed, then we could very confidently say that the difference was not due to chance, because according to these simulations, chance never produced differences that big. To help us see which ones are sensitive, let's draw some horizontal lines at -10 and +10.

I would say all of the designs with sample size = 100 or greater are all perfectly sensitive to real differences of 10 (if they exist). We can see that all of the dots after sample size 100 are underneath the red line. So effects that are as big as the red line, or bigger will almost never occur due to chance. But, if they do occur in nature, those experiments will detect them straight away. That is sensitivity. And, designing your experiment so that you know it is sensitive to the thing you are looking for is the big idea behind power. It's worth knowing this kind of thing before you run your experiment. Why waste your own time and run an experiment that doesn't have a chance of detecting the thing you are looking for.

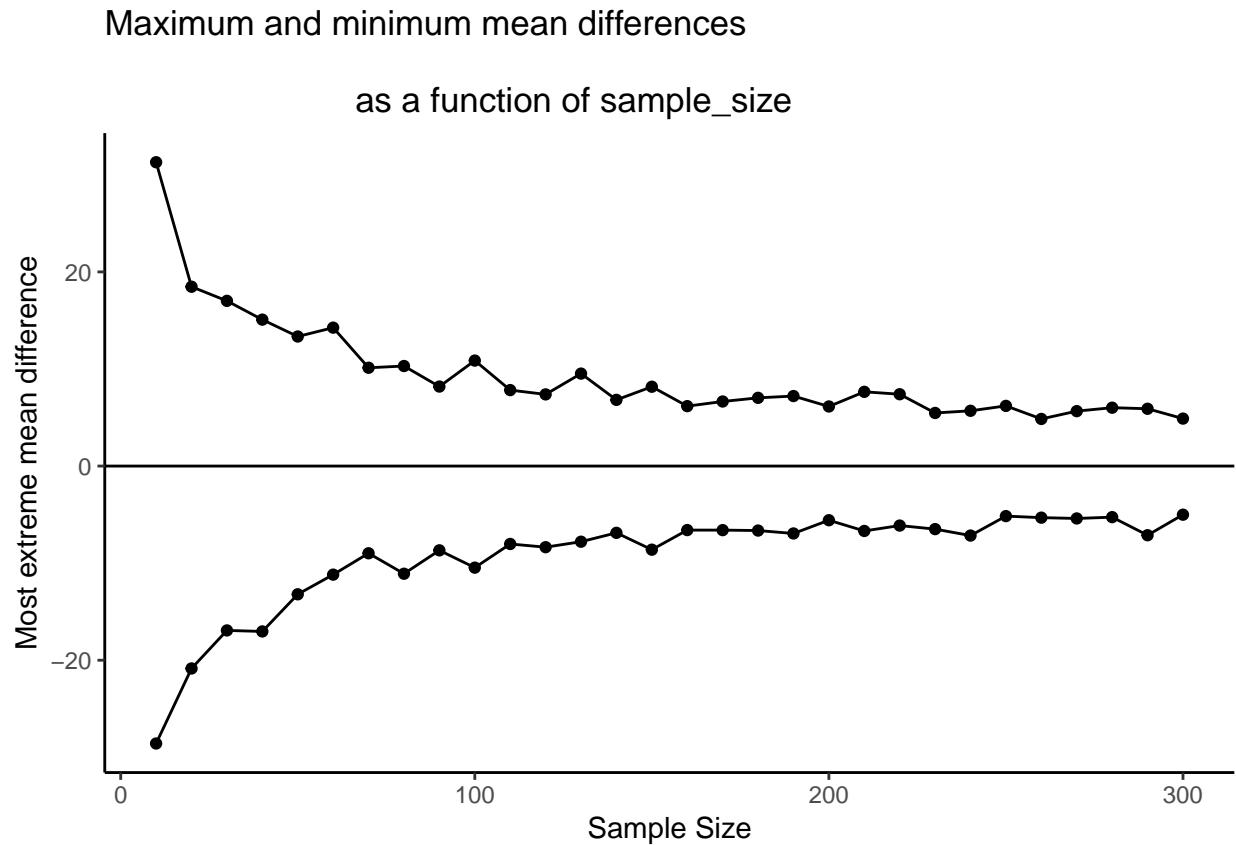


Figure 5.21: A graph of the maximum and minimum mean differences produced by chance as a function of sample-size. The range narrows as sample-size increases showing that chance alone produces a smaller range of mean differences as sample-size increases

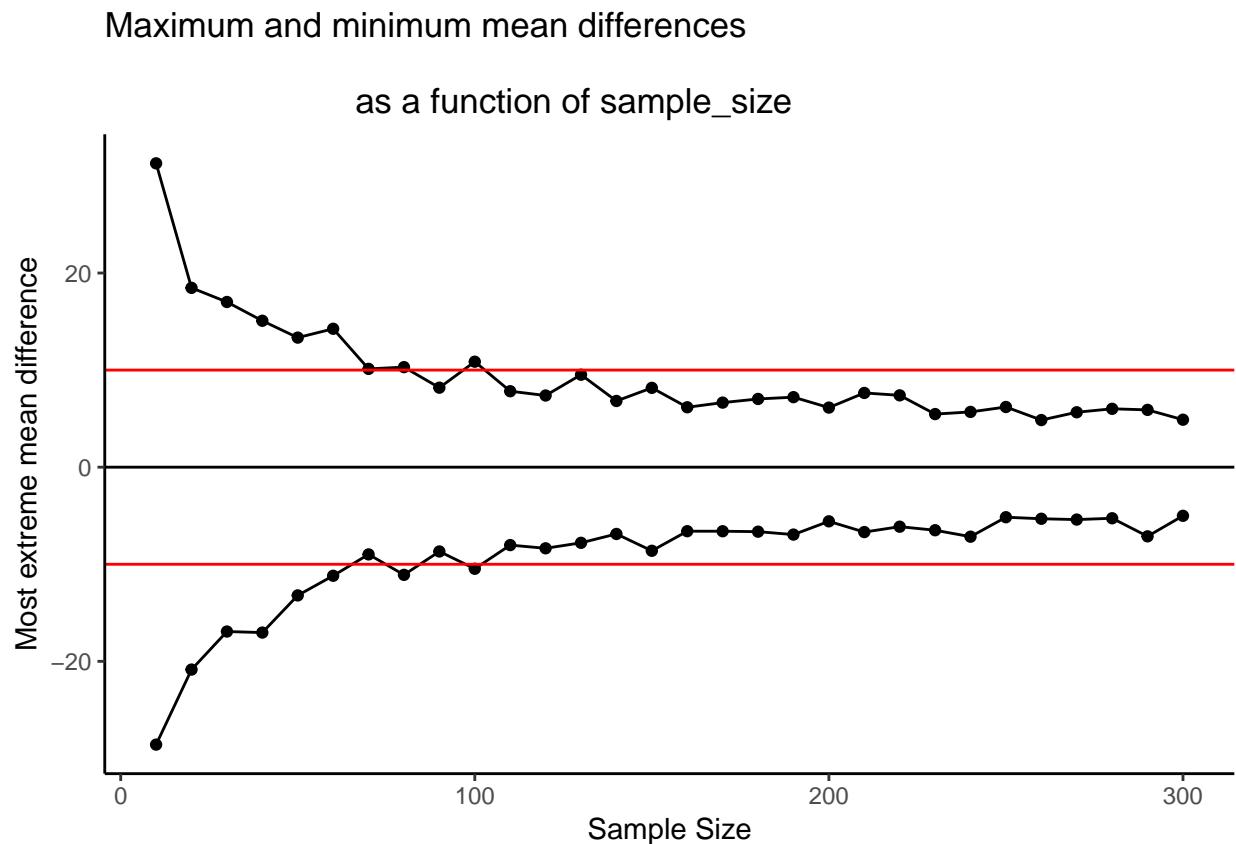


Figure 5.22: The red line represents the size of a mean difference that a researcher may be interested in detecting. All of the dots outside (above or below) the red line represent designs with small sample-sizes. When a difference of 10 occurs for these designs, we can rule out chance with confidence. The dots between the red lines represent designs with larger sample-sizes. These designs never produce differences as large as 10, so when those differences occur, we can be confident chance did not produce them.

5.5.7 Summary of Crump Test

What did we learn from this so-called fake Crump test that nobody uses? Well, we learned the basics of what we'll be doing moving forward. And, we did it all without any hard math or formulas. We sampled numbers, we computed means, we subtracted means, then we did that a lot and counted up the means and put them in a histogram. This showed us what chance do in an experiment. Then, we discussed how to make decisions around these facts. And, we showed how we can manipulate the role of chance just by changing things like sample size.

5.6 The randomization test (permutation test)

Welcome to the first official inferential statistic in this textbook. Up till now we have been building some intuitions for you. Next, we will get slightly more formal and show you how we can use random chance to tell us whether our experimental finding was likely due to chance or not. We do this with something called a randomization test. The ideas behind the randomization test are the very same ideas behind the rest of the inferential statistics that we will talk about in later chapters. And, surprise, we have already talked about all of the major ideas already. Now, we will just put the ideas together, and give them the name **randomization test**.

Here's the big idea. When you run an experiment and collect some data you get to find out what happened that one time. But, because you ran the experiment only once, you don't get to find out what **could have happened**. The randomization test is a way of finding out what **could have happened**. And, once you know that, you can compare **what did happen** in your experiment, with **what could have happened**.

5.6.1 Pretend example does chewing gum improve your grades?

Let's say you run an experiment to find out if chewing gum causes students to get better grades on statistics exams. You randomly assign 20 students to the chewing gum condition, and 20 different students to the no-chewing gum condition. Then, you give everybody statistics tests and measure their grades. If chewing gum causes better grades, then the chewing gum group should have higher grades on average than the group who did not chew gum.

Let's say the data looked like this:

student	gum	no_gum
1	96	69
2	96	79
3	72	82
4	80	83
5	71	54
6	71	68
7	94	43
8	86	61
9	73	87
10	96	51
11	75	45
12	87	47
13	94	82
14	71	61
15	82	62
16	70	60
17	99	45
18	83	65
19	96	54
20	78	80
Sums	1670	1278
Means	83.5	63.9

So, did the students chewing gum do better than the students who didn't chew gum? Look at the mean test performance at the bottom of the table. The mean for students chewing gum was 83.5, and the mean for students who did not chew gum was 63.9. Just looking at the means, it looks like chewing gum worked!

"STOP THE PRESSES, this is silly". We already know this is silly because we are making pretend data. But, even if this was real data, you might think, "Chewing gum won't do anything, this difference could have been caused by chance, I mean, maybe the better students just happened to be put into the chewing group, so because of that their grades were higher, chewing gum didn't do anything...". We agree. But, let's take a closer look. We already know how the data come out. What we want to know is how they could have come out, what are all the possibilities?

For example, the data would have come out a bit different if we happened to have put some of the students from the gum group into the no gum group, and vice versa. Think of all the ways you could have assigned the 40 students into two groups, there are lots of ways. And, the means for each group would turn out differently depending on how the students are assigned to each group.

Practically speaking, it's not possible to run the experiment every possible way, that would take too long. But, we can nevertheless estimate how all of those experiments might have turned out using simulation.

Here's the idea. We will take the 40 measurements (exam scores) that we found for all the students. Then we will randomly take 20 of them and pretend they were in the gum group, and we'll take the remaining 20 and pretend they were in the no gum group. Then we can compute the means again to find out what would have happened. We can keep doing this over and over again. Every time computing what happened in that version of the experiment.

5.6.1.1 Doing the randomization

Before we do that, let's show how the randomization part works. We'll use fewer numbers to make the process easier to look at. Here are the first 5 exam scores for students in both groups.

student	gum	no_gum
1	96	69
2	96	79
3	72	82
4	80	83
5	71	54
Sums	415	367
Means	83	73.4

Things could have turned out differently if some of the subjects in the gum group were switched with the subjects in the no gum group. Here's how we can do some random switching. We will do this using R.

```
all_scores      <- c(gum[1:5],no_gum[1:5])
randomize_scores <- sample(all_scores)
new_gum          <- randomize_scores[1:5]
new_no_gum       <- randomize_scores[6:10]
print(new_gum)

## [1] 79 72 80 96 82
print(new_no_gum)

## [1] 71 69 83 54 96
```

We have taken the first 5 numbers from the original data, and put them all into a variable called `all_scores`. Then we use the `sample` function in R to shuffle the scores. Finally, we take the first 5 scores from the shuffled numbers and put them into a new variable called `new_gum`. Then, we put the last five scores into the variable `new_no_gum`. Then we printed them, so we can see them.

If we do this a couple of times and put them in a table, we can indeed see that the means for gum and no gum would be different if the subjects were shuffled around. Check it out:

student	gum	no_gum	gum2	no_gum2	gum3	no_gum3
1	96	69	80	54	96	96
2	96	79	71	79	54	82
3	72	82	83	82	79	72
4	80	83	69	96	69	83
5	71	54	72	96	71	80
Sums	415	367	375	407	369	413
Means	83	73.4	75	81.4	73.8	82.6

5.6.1.2 Simulating the mean differences across the different randomizations

In our pretend experiment we found that the mean for students chewing gum was 83.5, and the mean for students who did not chew gum was 63.9. The mean difference (gum - no gum) was 19.6. This is a pretty big difference. This is **what did happen**. But, **what could have happened?** If we tried out all of the experiments where different subjects were switched around, what does the distribution of the possible mean differences look like? Let's find out. This is what the randomization test is all about.

When we do our randomization test we will measure the mean difference in exam scores between the gum group and the no gum group. Every time we randomize we will save the mean difference.

Let's look at a short animation of what is happening in the randomization test. Note, what you are about to see is data from a different fake experiment, but the principles are the same. We'll return to the gum no gum experiment after the animation. The animation is showing you three important things. First, the purple dots show you the mean scores in two groups (didn't study vs study). It looks like there is a difference, as 1 dot is lower than the other. We want to know if chance could produce a difference this big. At the

Animation not available in .pdf version

Figure 5.23: Animation of a randomization test. The purple dots represent the location of the original sample means in each condition. The yellow dots represent the means of each randomized sample. The blue and red dots show how the original scores are shuffled across each randomization.

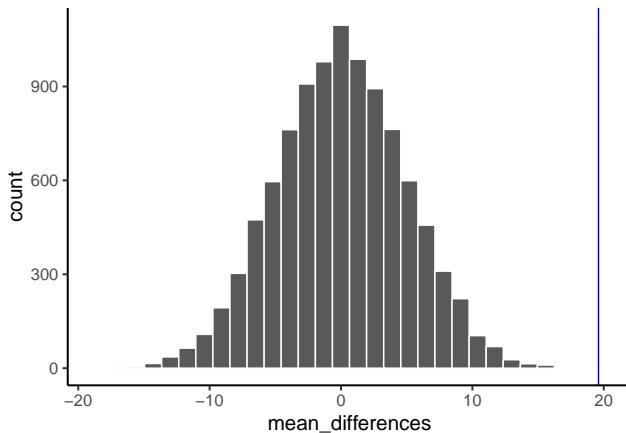


Figure 5.24: A histogram of simulated mean differences for a randomization test

beginning of the animation, the light green and red dots show the individual scores from each of 10 subjects in the design (the purple dots are the means of these original scores). Now, during the randomizations, we randomly shuffle the original scores between the groups. You can see this happening throughout the animation, as the green and red dots appear in different random combinations. The moving yellow dots show you the new means for each group after the randomization. The differences between the yellow dots show you the range of differences that chance could produce.

We are engaging in some visual statistical inference. By looking at the range of motion of the yellow dots, we are watching what kind of differences chance can produce. In this animation, the purple dots, representing the original difference, are generally outside of the range of chance. The yellow dots don't move past the purple dots, as a result chance is an unlikely explanation of the difference.

If the purple dots were inside the range of the yellow dots, then when would know that chance is capable of producing the difference we observed, and that it does so fairly often. As a result, we should not conclude the manipulation caused the difference, because it could have easily occurred by chance.

Let's return to the gum example. After we randomize our scores many times, and computed the new means, and the mean differences, we will have loads of mean differences to look at, which we can plot in a histogram. The histogram gives a picture of **what could have happened**. Then, we can compare **what did happen** with **what could have happened**.

Here's the histogram of the mean differences from the randomization test. For this simulation, we randomized the results from the original experiment 1000 times. This is what could have happened. The blue line in the figure shows us where our observed difference lies on the x-axis.

What do you think? Could the difference represented by the blue line have been caused by chance? My answer is probably not. The histogram shows us the window of chance. The blue line is not inside the window. This means we can be pretty confident that the difference we observed was not due to chance.

We are looking at another window of chance. We are seeing a histogram of the kinds of mean differences

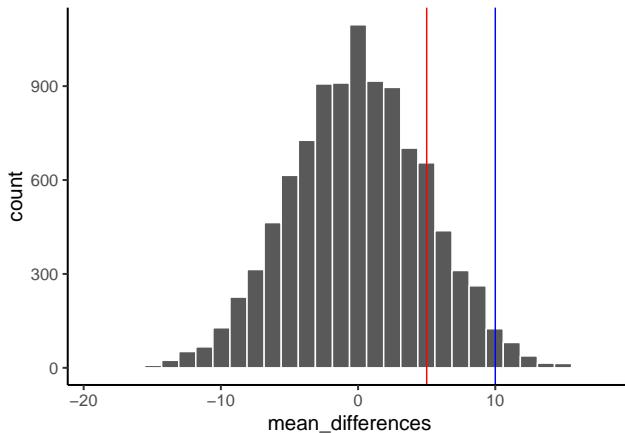


Figure 5.25: Would you expect a mean difference represented by the blue line to occur more or less often by chance compared to the mean difference represented by the red line?

that could have occurred in our experiment, if we had assigned our subjects to the gum and no gum groups differently. As you can see, the mean differences range from negative to positive. The most frequent difference is 0. Also, the distribution appears to be symmetrical about zero, which shows we had roughly same the chances of getting a positive or negative difference. Also, notice that as the differences get larger (in the positive or negative direction, they become less frequent). The blue line shows us **the observed difference**, this is the one we found in our fake experiment. Where is it? It's way out to the right. It is well outside the histogram. In other words, when we look at **what could have happened**, we see that **what did happen** doesn't occur very often.

IMPORTANT: In this case, when we speak of **what could have happened**. We are talking about what could have happened **by chance**. When we compare what did happen to what chance could have done, we can get a better idea of whether our result was caused by chance.

OK, let's pretend we got a much smaller mean difference when we first ran the experiment. We can draw new lines (blue and red) to represent a smaller mean we might have found.

Look at the blue line. If you found a mean difference of 10, would you be convinced that your difference was not caused by chance? As you can see, the blue line is inside the chance window. Notably, differences of +10 don't very often. You might infer that your difference was not likely to be due to chance (but you might be a little bit skeptical, because it could have been). How about the red line? The red line represents a difference of +5. If you found a difference of +5 here, would you be confident that your difference was not caused by chance? I wouldn't be. The red line is totally inside the chance window, this kind of difference happens fairly often. I'd need some more evidence to consider the claim the some independent variable actually caused the difference. I'd be much more comfortable assuming that sampling error probably caused the difference.

5.6.2 Take homes so far

Have you noticed that we haven't used any formulas yet, but we have been able to accomplish inferential statistics. We will see some formulas as we progress, but these aren't as the idea behind the formulas.

Inferential statistics is an attempt to solve the problem: **where did my data from?**. In the randomization test example, our question was: **where did the differences between the means in my data come from?**. We know that the differences could be produced by chance alone. We simulated what chance can do using randomization. Then we plotted what chance can do using a histogram. Then, we used to picture to help us make an inference. Did our observed difference come from the distribution, or not? When the

observed difference is clearly inside the chance distribution, then we can infer that our difference **could have been produced by chance**. When the observed difference is not clearly inside the chance distribution, then we can infer that our difference was **probably not produced by chance**.

In my opinion, these pictures are very, very helpful. If one of our goals is to help ourselves summarize a bunch of complicated numbers to arrive at an inference, then the pictures do a great job. We don't even need a summary number, we just need to look at the picture and see if the observed difference is inside or outside of the window. This is what it is all about. Creating intuitive and meaningful ways to make inferences from our data. As we move forward, the main thing that we will do is formalize our process, and talk more about "standard" inferential statistics. For example, rather than looking at a picture (which is a good thing to do), we will create some helpful numbers. For example, what if you wanted to the probability that your difference could have been produced by chance? That could be a single number, like 95%. If there was a 95% probability that chance can produce the difference you observed, you might not be very confident that something like your experimental manipulation was causing the difference. If there was only 1% probability that chance could produce your difference, then you might be more confident that **chance did not** produce the difference; and, you might instead be comfortable with the possibility that your experimental manipulation actually caused the difference. So, how can we arrive at those numbers? In order to get there we will introduce you to some more foundational tools for statistical inference.

Chapter 6

t-Tests

One day, many moons ago, William Sealy Gosset got a job working for Guinness Breweries. They make the famous Irish stout called Guinness. What happens next went something like this (total fabrication, but mostly on point).

Guinness wanted all of their beers to be the best beers. No mistakes, no bad beers. They wanted to improve their quality control so that when Guinness was poured anywhere in the world, it would always come out fantastic: 5 stars out of 5 every time, the best.

Guinness had some beer tasters, who were super-experts. Every time they tasted a Guinness from the factory that wasn't 5 out of 5, they knew right away.

But, Guinness had a big problem. They would make a keg of beer, and they would want to know if every single pint that would come out would be a 5 out of 5. So, the beer tasters drank pint after pint out of the keg, until it was gone. Some kegs were all 5 out of 5s. Some weren't, Guinness needed to fix that. But, the biggest problem was that, after the testing, there **was no beer left to sell**, the testers drank it all (remember I'm making this part up to illustrate a point, they probably still had beer left to sell).

Guinness had a sampling and population problem. They wanted to know that the entire population of the beers they made were all 5 out of 5 stars. But, if they sampled the entire population, they would drink all of their beer, and wouldn't have any left to sell.

Enter William Sealy Gosset. Gosset figured out the solution to the problem. He asked questions like this:

1. How many samples do I need to take to know the whole population is 5 out of 5?
2. What's the fewest amount of samples I need to take to know the above, that would mean Guinness could test fewer beers for quality, sell more beers for profit, and make the product testing time shorter.

Gosset solved those questions, and he invented something called the *Student's t-test*. Gosset was working for Guinness, and could be fired for releasing trade-secrets that he invented (the t-test). But, Gosset published the work anyways, under a pseudonym (Student, 1908). He called himself Student, hence Student's t-test. Now you know the rest of the story.

It turns out this was a very nice thing for Gosset to have done. t-tests are used all the time, and they are useful, that's why they are used. In this chapter we learn how they work.

You'll be surprised to learn that what we've already talked about, (the Crump Test, and the Randomization Test), are both very very similar to the t-test. So, in general, you have already been thinking about the things you need to think about to understand t-tests. You're probably wondering what is this *t*, what does *t* mean? We will tell you. Before we tell what it means, we first tell you about one more idea.

6.1 Check your confidence in your mean

We've talked about getting a sample of data. We know we can find the mean, we know we can find the standard deviation. We know we can look at the data in a histogram. These are all useful things to do for us to learn something about the properties of our data.

You might be thinking of the mean and standard deviation as very different things that we would not put together. The mean is about central tendency (where most of the data is), and the standard deviation is about variance (where most of the data isn't). Yes, they are different things, but we can use them together to create useful new things.

What if I told you my sample mean was 50, and I told you nothing else about my sample. Would you be confident that most of the numbers were near 50? Would you wonder if there was a lot of variability in the sample, and many of the numbers were very different from 50. You should wonder all of those things. The mean alone, just by itself, doesn't tell you anything about how well the mean represents all of the numbers in the sample.

It could be a representative number, when the standard deviation is very small, and all the numbers are close to 50. It could be a non-representative number, when the standard deviation is large, and many of the numbers are not near 50. You need to know the standard deviation in order to be confident in how well the mean represents the data.

How can we put the mean and the standard deviation together, to give us a new number that tells us about confidence in the mean?

We can do this using a ratio:

$$\frac{\text{mean}}{\text{standard deviation}}$$

Think about what happens here. We are dividing a number by a number. Look at what happens:

$$\frac{\text{number}}{\text{same number}} = 1$$

$$\frac{\text{number}}{\text{smaller number}} = \text{big number}$$

compared to:

$$\frac{\text{number}}{\text{bigger number}} = \text{smaller number}$$

Imagine we have a mean of 50, and a truly small standard deviation of 1. What do we get with our formula?

$$\frac{50}{1} = 50$$

Imagine we have a mean of 50, and a big standard deviation of 100. What do we get with our formula?

$$\frac{50}{100} = 0.5$$

Notice, when we have a mean paired with a small standard deviation, our formula gives us a big number, like 50. When we have a mean paired with a large standard deviation, our formula gives us a small number, like 0.5. These numbers can tell us something about confidence in our mean, in a general way. We can be 50 confident in our mean in the first case, and only 0.5 (not at a lot) confident in the second case.

What did we do here? We created a descriptive statistic by dividing the mean by the standard deviation. And, we have a sense of how to interpret this number, when it's big we're more confident that the mean represents all of the numbers, when it's small we are less confident. This is a useful kind of number, a ratio between what we think about our sample (the mean), and the variability in our sample (the standard deviation). Get used to this idea. Almost everything that follows in this textbook is based on this kind of ratio. We will see that our ratio turns into different kinds of "statistics", and the ratios will look like this in general:

$$\text{name of statistic} = \frac{\text{measure of what we know}}{\text{measure of what we don't know}}$$

or, to say it using different words:

$$\text{name of statistic} = \frac{\text{measure of effect}}{\text{measure of error}}$$

In fact, this is the general formula for the t-test. Big surprise!

6.2 One-sample t-test: A new t-test

Now we are ready to talk about t-test. We will talk about three of them. We start with the one-sample t-test.

Commonly, the one-sample t-test is used to estimate the chances that your sample came from a particular population. Specifically, you might want to know whether the mean that you found from your sample, could have come from a particular population having a particular mean.

Straight away, the one-sample t-test becomes a little confusing (and I haven't even described it yet). Officially, it uses known parameters from the population, like the mean of the population and the standard deviation of the population. However, most times you don't know those parameters of the population! So, you have to estimate them from your sample. Remember from the chapters on descriptive statistics and sampling, our sample mean is an unbiased estimate of the population mean. And, our sample standard deviation (the one where we divide by $n-1$) is an unbiased estimate of the population standard deviation. When Gosset developed the t-test, he recognized that he could use these estimates from his samples, to make the t-test. Here is the formula for the one sample t-test, we first use words, and then become more specific:

6.2.1 Formulas for one-sample t-test

$$\text{name of statistic} = \frac{\text{measure of effect}}{\text{measure of error}}$$

$$t = \frac{\text{measure of effect}}{\text{measure of error}}$$

$$t = \frac{\text{Mean difference}}{\text{standard error}}$$

$$t = \frac{\bar{X} - u}{S_{\bar{X}}}$$

$$t = \frac{\text{Sample Mean} - \text{Population Mean}}{\text{Sample Standard Error}}$$

$$\text{Estimated Standard Error} = \text{Standard Error of Sample} = \frac{s}{\sqrt{N}}$$

Where, s is the sample standard deviation.

Some of you may have gone cross-eyed looking at all of this. Remember, we've seen it before when we divided our mean by the standard deviation in the first bit. The t-test is just a measure of a sample mean, divided by the standard error of the sample mean. That is it.

6.2.2 What does t represent?

t gives us a measure of confidence, just like our previous ratio for dividing the mean by a standard deviations. The only difference with t , is that we divide by the standard error of mean (remember, this is also a standard deviation, it is the standard deviation of the sampling distribution of the mean)

What does the t in t-test stand for? Apparently nothing. Gosset originally labelled it z . And, Fisher later called it t , perhaps because t comes after s , which is often used for the sample standard deviation.

t is a property of the data that you collect. You compute it with a sample mean, and a sample standard error (there's one more thing in the one-sample formula, the population mean, which we get to in a moment). This is why we call t , a sample-statistic. It's a statistic we compute from the sample.

What kinds of numbers should we expect to find for these ts ? How could we figure that out?

Let's start small and work through some examples. Imagine your sample mean is 5. You want to know if it came from a population that also has a mean of 5. In this case, what would t be? It would be zero: we first subtract the sample mean from the population mean, $5 - 5 = 0$. Because the numerator is 0, t will be zero. So, $t = 0$, occurs, when there is no difference.

Let's say you take another sample, do you think the mean will be 5 every time, probably not. Let's say the mean is 6. So, what can t be here? It will be a positive number, because $6-5=+1$. But, will t be +1? That depends on the standard error of the sample. If the standard error of the sample is 1, then t could be 1, because $1/1 = 1$.

If the sample standard error is smaller than 1, what happens to t ? It gets bigger right? For example, 1 divided by 0.5 = 2. If the sample standard error was 0.5, t would be 2. And, what could we do with this information? Well, it be like a measure of confidence. As t gets bigger we could be more confident in the mean difference we are measuring.

Can t be smaller than 1? Sure, it can. If the sample standard error is big, say like 2, then t will be smaller than one (in our case), e.g., $1/2 = .5$. The direction of the difference between the sample mean and population mean, can also make the t become negative. What if our sample mean was 4. Well, then t will be negative, because the mean difference in the numerator will be negative, and the number in the bottom (denominator) will always be positive (remember why, it's the standard error, computed from the sample standard deviation, which is always positive because of the squaring that we did.).

So, that is some intuitions about what the kinds of values t can take. t can be positive or negative, and big or small.

Let's do one more thing to build our intuitions about what t can look like. How about we sample some numbers and then measure the sample mean **and** the standard error of the mean, and then plot those two things against each other. This will show us how a sample mean typically varies with respect to the standard error of the mean.

In the following figure, I pulled 1,000 samples of $N=10$ from a normal distribution (mean = 0, sd = 1). Each time I measured the mean and standard error of the sample. That gave two descriptive statistics for each sample, letting us plot each sample as dot in a scatterplot

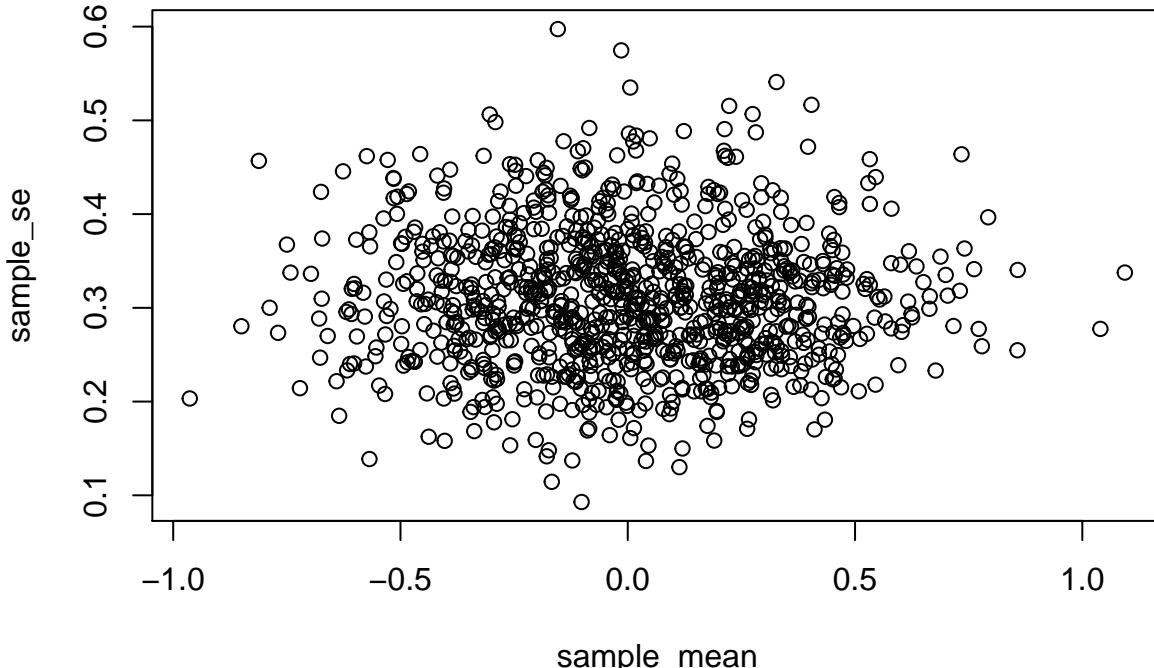


Figure 6.1: A scatterplot with sample mean on the x-axis, and standard error of the mean on the y-axis

What we get is a cloud of dots. You might notice the cloud has a circular quality. There's more dots in the middle, and fewer dots as they radiate out from the middle. The dot cloud shows us the general range of the sample mean, for example most of the dots are in between -1 and 1. Similarly, the range for the sample standard error is roughly between .2 and .5. Remember, each dot represents one sample.

We can look at the same data a different way. For example, rather than using a scatterplot, we can divide the mean for each dot, by the standard error for each dot. Below is a histogram showing what this looks like:

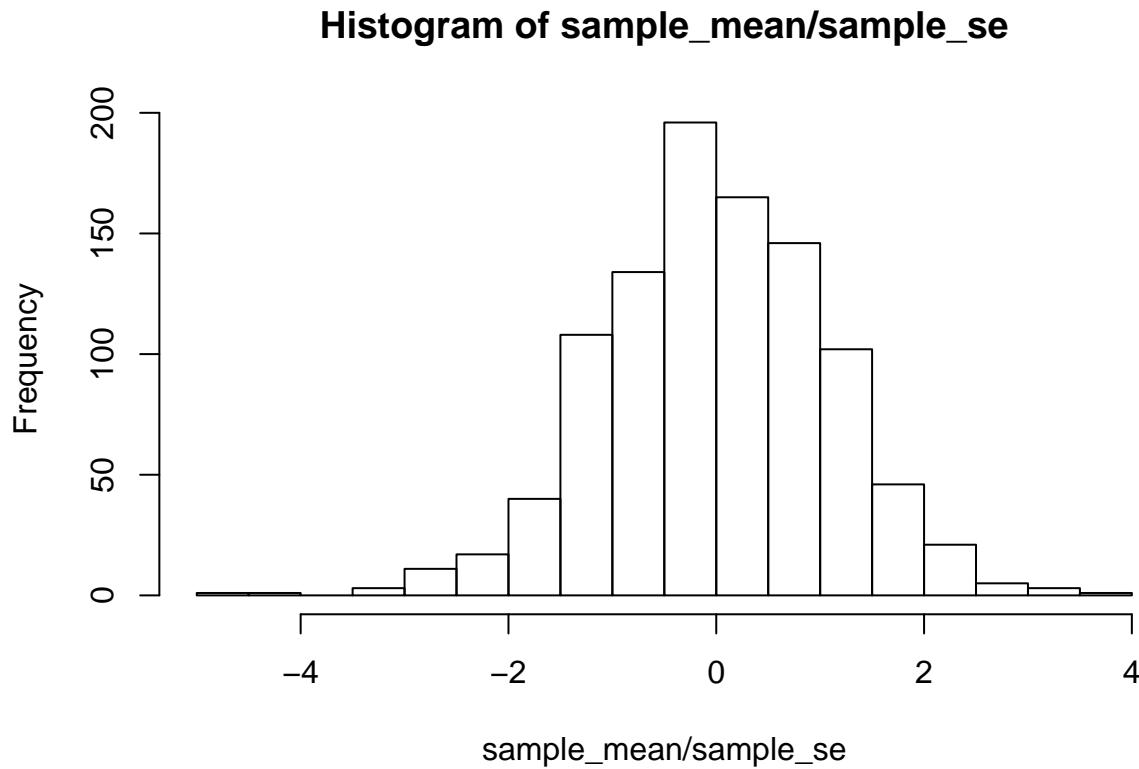


Figure 6.2: A histogram of the sample means divided by the sample standard errors, this is a t-distribution

Interesting, we can see the histogram is shaped like a normal curve. It is centered on 0, which is the most common value. As values become more extreme, they become less common. If you remember, our formula for t , was the mean divided by the standard error of the mean. That's what we did here. This histogram is showing you a t -distribution.

6.2.3 Calculating t from data

Let's briefly calculate a t-value from a small sample. Let's say we had 10 students do a true/false quiz with 5 questions on it. There's a 50% chance of getting each answer correct.

Every student completes the 5 questions, we grade them, and then we find their performance (mean percent correct). What we want to know is whether the students were guessing. If they were all guessing, then the sample mean should be about 50%, it shouldn't be different from chance, which is 50%. Let's look at the table:

students	scores	mean	Difference_from_Mean	Squared_Deviations
1	50	61	-11	121
2	70	61	9	81
3	60	61	-1	1
4	40	61	-21	441
5	80	61	19	361
6	30	61	-31	961
7	90	61	29	841
8	60	61	-1	1
9	70	61	9	81
10	60	61	-1	1
Sums	610	610	0	2890
Means	61	61	0	289
			sd	17.92
			SEM	5.67
			t	1.94003527336861

You can see the **scores** column has all of the test scores for each of the 10 students. We did the things we need to do to compute the standard deviation.

Remember the sample standard deviation is the square root of the sample variance, or:

$$\text{sample standard deviation} = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{N-1}}$$

$$sd = \sqrt{\frac{2890}{10-1}} = 17.92$$

The standard error of the mean, is the standard deviation divided by the square root of N

$$SEM = \frac{s}{\sqrt{N}} = \frac{17.92}{\sqrt{10}} = 5.67$$

t is the difference between our sample mean (61), and our population mean (50, assuming chance), divided by the standard error of the mean.

$$t = \frac{\bar{X} - u}{SEM} = \frac{61 - 50}{5.67} = 1.94$$

And, that is you how calculate *t*, by hand. It's a pain. I was annoyed doing it this way. In the lab, you learn how to calculate *t* using software, so it will just spit out *t*. For example in R, all you have to do is this:

```
t.test(scores, mu=50)
```

```
##
##  One Sample t-test
##
## data:  scores
## t = 1.9412, df = 9, p-value = 0.08415
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  48.18111 73.81889
## sample estimates:
## mean of x
##          61
```

6.2.4 How does *t* behave?

If *t* is just a number that we can compute from our sample (it is), what can we do with it? How can we use *t* for statistical inference?

Remember back to the chapter on sampling and distributions, that's where we discussed the sampling distribution of the sample mean. Remember, we made a lot of samples, then computed the mean for each sample, then we plotted a histogram of the sample means. Later, in that same section, we mentioned that we could generate sampling distributions for any statistic. For each sample, we could compute the mean, the standard deviation, the standard error, and now even t , if we wanted to. We could generate 10,000 samples, and draw four histograms, one for each sampling distribution for each statistic.

This is exactly what I did, and the results are shown in the four figures below. I used a sample size of 20, and drew random observations for each sample from a normal distribution, with mean = 0, and standard deviation = 1. Let's look at the sampling distributions for each of the statistics. t was computed assuming with the population mean assumed to be 0.

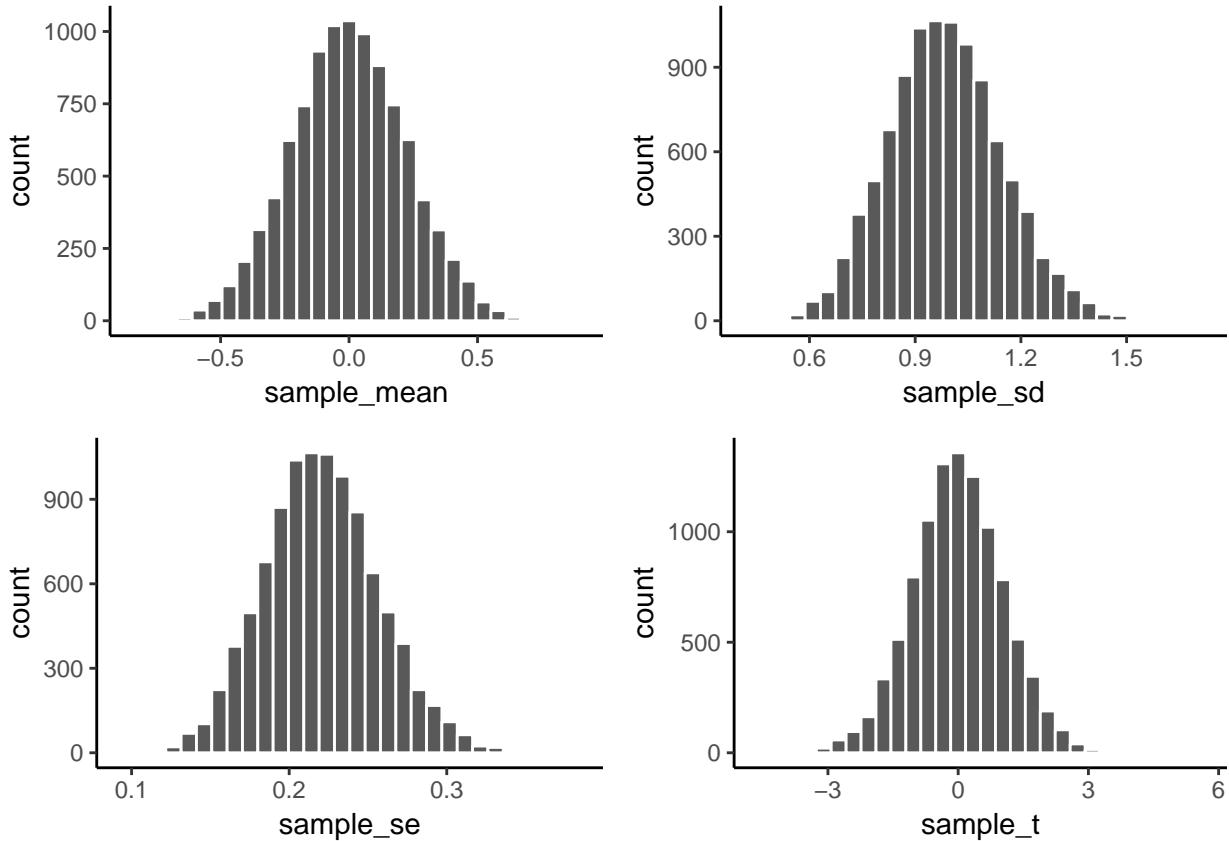


Figure 6.3: Sampling distributions for the mean, standard deviation, standard error of the mean, and t

We see four sampling distributions. This is how statistical summaries of these summaries behave. We have used the word chance windows before. These are four chance windows, measuring different aspects of the sample. In this case, all of the samples came from the same normal distribution. Because of sampling error, each sample is not identical. The means are not identical, the standard deviations are not identical, sample standard error of the means are not identical, and the ts of the samples are not identical. They all have some variation, as shown by the histograms. This is how samples of size 20 behave.

We can see straight away, that in this case, we are unlikely to get a sample mean of 2. That's way outside the window. The range for the sampling distribution of the mean is around -.5 to +.5, and is centered on 0 (the population mean, would you believe!).

We are unlikely to get sample standard deviations of between .6 and 1.5, that is a different range, specific to the sample standard deviation.

Same thing with the sample standard error of the mean, the range here is even smaller, mostly between .1,

and .3. You would rarely find a sample with a standard error of the mean greater than .3. Virtually never would you find one of say 1 (for this situation).

Now, look at t . Its range is basically between -3 and +3 here. 3s barely happen at all. You pretty much never see a 5 or -5 in this situation.

All of these sampling windows are chance windows, and they can all be used in the same way as we have used similar sampling distributions before (e.g., Crump Test, and Randomization Test) for statistical inference. For all of them we would follow the same process:

1. Generate these distributions
2. Look at your sample statistics for the data you have (mean, SD, SEM, and t)
3. Find the likelihood of obtaining that value or greater
4. Obtain that probability
5. See if you think your sample statistics were probable or improbable.

We'll formalize this in a second. I just want you to know that what you will be doing is something that you have already done before. For example, in the Crump test and the Randomization test we focused on the distribution of mean differences. We could do that again here, but instead, we will focus on the distribution of t values. We then apply the same kinds of decision rules to the t distribution, as we did for the other distributions. Below you will see a graph you have already seen, except this time it is a distribution of ts , not mean differences:

Remember, if we obtained a single t from one sample we collected, we could consult this chance window below to find out the t we obtained from the sample was likely or unlikely to occur by chance.

Histogram of mean sample_ts between two samples (n=20)

both drawn from the same normal distribution ($\mu=0$, $sd=1$)

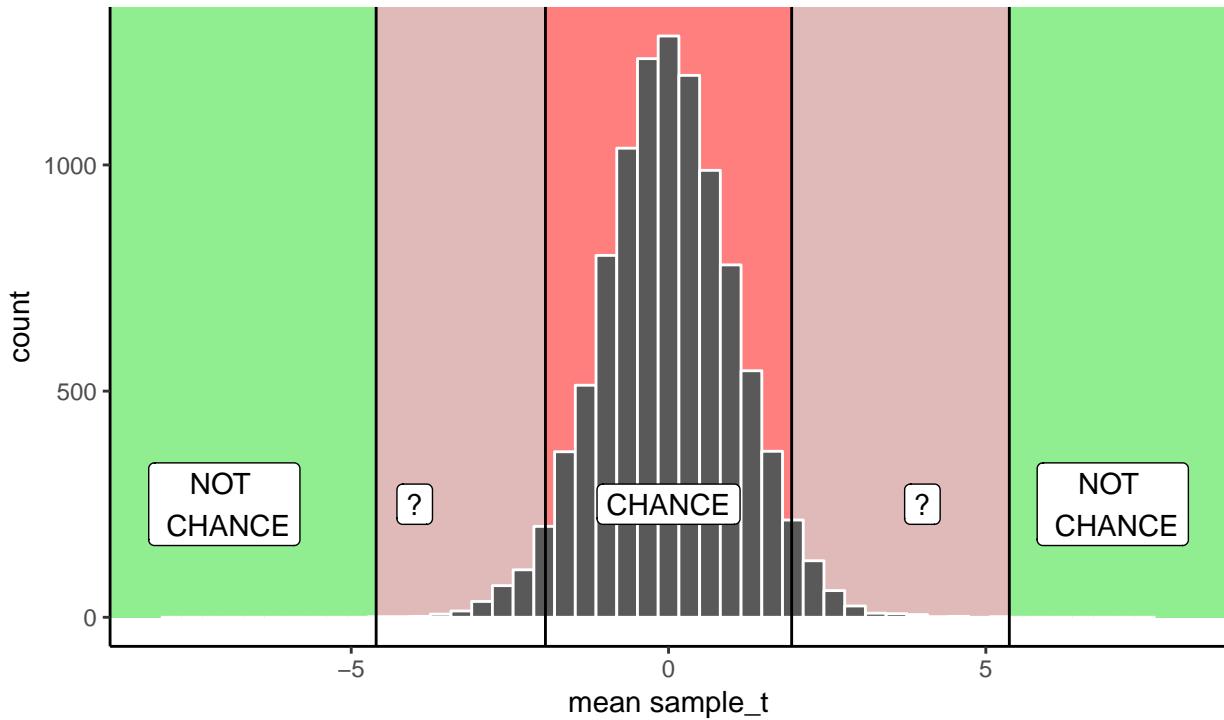


Figure 6.4: Applying decision criteria to the t-distribution.

6.2.5 Making a decision

From our early example involving the TRUE/FALSE quizzes, we are now ready to make some kind of decision about what happened there. We found a mean difference of 11. We found a $t = 1.9411765$. The probability of this t or larger occurring is $p = 0.0841503$. We were testing the idea that our sample mean of 61 could have come from a normal distribution with mean = 50. The t test tells us that the t for our sample, or a larger one, would happen with $p = 0.0841503$. In other words, chance can do it a kind of small amount of time, but not often. In English, this means that all of the students could have been guessing, but it wasn't that likely that were just guessing.

We're guessing that you are still a little bit confused about t values, and what we are doing here. We are going to skip ahead to the next t -test, called a **paired samples t-test**. We will also fill in some more things about t -tests that are more obvious when discussing paired samples t-test. In fact, spoiler alert, we will find out that a paired samples t-test is actually a one-sample t-test in disguise (WHAT!), yes it is. If the one-sample t -test didn't make sense to you, read the next section.

6.3 Paired-samples t -test

For me (Crump), many analyses often boil down to a paired samples t-test. It just happens that many things I do reduce down to a test like this.

I am a cognitive psychologist, I conduct research about how people do things like remember, pay attention, and learn skills. There are lots of Psychologists like me, who do very similar things.

We all often conduct the same kinds of experiments. They go like this, and they are called **repeated measures** designs. They are called **repeated measures** designs, because we measure how one person does something more than once, we **repeat** the measure.

So, I might measure somebody doing something in condition A, and measure the same person doing something in Condition B, and then I see that same person does different things in the two conditions. I **repeatedly measure** the same person in both conditions. I am interested in whether the experimental manipulation changes something about how people perform the task in question.

6.3.1 Mehr, Song, and Spelke (2016)

We will introduce the paired-samples t-test with an example using real data, from a real study. Mehr et al. (2016) were interested in whether singing songs to infants helps infants become more sensitive to social cues. For example, infants might need to learn to direct their attention toward people as a part of learning how to interact socially with people. Perhaps singing songs to infants aids this process of directing attention. When an infant hears a familiar song, they might start to pay more attention to the person singing that song, even after they are done singing the song. The person who sang the song might become more socially important to the infant. You will learn more about this study in the lab for this week. This example, prepares you for the lab activities. Here is a brief summary of what they did.

First, parents were trained to sing a song to their infants. After many days of singing this song to the infants, a parent came into the lab with their infant. In the first session, parents sat with their infants on their knees, so the infant could watch two video presentations. There were two videos. Each video involved two unfamiliar new people the infant had never seen before. Each new person in the video (the singers) sang one song to the infant. One singer sang the “familiar” song the infant had learned from their parents. The other singer sang an “unfamiliar” song the infant had not heard before.

There were two really important measurement phases: the baseline phase, and the test phase.

The baseline phase occurred before the infants saw and heard each singer sing a song. During the baseline phase, the infants watched a video of both singers at the same time. The researchers recorded the proportion

of time that the infant looked at each singer. The baseline phase was conducted to determine whether infants had a preference to look at either person (who would later sing them a song).

The test phase occurred **after** infants saw and heard each song, sung by each singer. During the test phase, each infant had an opportunity to watch silent videos of both singers. The researchers measured the proportion of time the infants spent looking at each person. The question of interest, was whether the infants would spend a greater proportion of time looking at the singer who sang the familiar song, compared to the singer who sang the unfamiliar song.

There is more than one way to describe the design of this study. We will describe it like this. It was a repeated measures design, with one independent (manipulation) variable called Viewing phase: Baseline versus Test. There was one dependent variable (the measurement), which was proportion looking time (to singer who sung familiar song). This was a repeated measures design because the researchers measured proportion looking time twice (they repeated the measure), once during baseline (before infants heard each singer sing a song), and again during test (after infants heard each singer sing a song).

The important question was whether infants would change their looking time, and look more at the singer who sang the familiar song during the test phase, than they did during the baseline phase. This is a question about a change within individual infants. In general, the possible outcomes for the study are:

1. No change: The difference between looking time toward the singer of the familiar song during baseline and test is zero, no difference.
2. Positive change: Infants will look longer toward the singer of the familiar song during the test phase (after they saw and heard the singers), compared to the baseline phase (before they saw and heard the singers). This is a positive difference if we use the formula: Test Phase Looking time - Baseline phase looking time (to familiar song singer).
3. Negative change: Infants will look longer toward the singer of the unfamiliar song during the test phase (after they saw and heard the singers), compared to the baseline phase (before they saw and heard the singers). This is a negative difference if we use the same formula: Test Phase Looking time - Baseline phase looking time (to familiar song singer).

6.3.2 The data

Let's take a look at the data for the first 5 infants in the study. This will help us better understand some properties of the data before we analyze it. We will see that the data is structured in a particular way that we can take advantage of with a paired samples t-test. Note, we look at the first 5 infants to show how the computations work. The results of the paired-samples t-test change when we use all of the data from the study.

Here is a table of the data:

infant	Baseline	Test
1	0.44	0.60
2	0.41	0.68
3	0.75	0.72
4	0.44	0.28
5	0.47	0.50

The table shows proportion looking times toward the singer of the familiar song during the Baseline and Test phases. Notice there are five different infants, (1 to 5). Each infant is measured twice, once during the Baseline phase, and once during the Test phase. To repeat from before, this is a repeated-measures design, because the infants are measured repeatedly (twice in this case). Or, this kind of design is also called a **paired-samples** design. Why? because each participant comes with a pair of samples (two samples), one for each level of the design.

Great, so what are we really interested in here? We want to know if the mean looking time toward the singer of the familiar song for the Test phase is higher than the Baseline phase. We are comparing the two sample means against each other and looking for a difference. We already know that differences could be obtained by chance alone, simply because we took two sets of samples, and we know that samples can be different. So, we are interested in knowing whether chance was likely or unlikely to have produced any difference we might observe.

In other words, we are interested in looking at the difference scores between the baseline and test phase for each infant. The question here is, for each infant, did their proportion looking time to the singer of the familiar song, increase during the test phase as compared to the baseline phase.

6.3.3 The difference scores

Let's add the difference scores to the table of data so it is easier to see what we are talking about. The first step in creating difference scores is to decide how you will take the difference, there are two options:

1. Test phase score - Baseline Phase Score
2. Baseline phase score - Test Phase score

Let's use the first formula. Why? Because it will give us positive differences when the test phase score is higher than the baseline phase score. This makes a positive score meaningful with respect to the study design, we know (because we defined it to be this way), that positive scores will refer to longer proportion looking times (to singer of familiar song) during the test phase compared to the baseline phase.

infant	Baseline	Test	differences
1	0.44	0.60	0.16
2	0.41	0.68	0.27
3	0.75	0.72	-0.03
4	0.44	0.28	-0.16
5	0.47	0.50	0.03

There we have it, the difference scores. The first thing we can do here is look at the difference scores, and ask how many infants showed the effect of interest. Specifically, how many infants showed a positive difference score. We can see that three of five infants showed a positive difference (they looked more at the singer of the familiar song during the test than baseline phase), and two the infants showed the opposite effect (negative difference, they looked more at the singer of the familiar song during baseline than test).

6.3.4 The mean difference

As we have been discussing, the effect of interest in this study is the mean difference between the baseline and test phase proportion looking times. We can calculate the **mean difference**, by finding the **mean of the difference scores**. Let's do that, in fact, for fun let's calculate the mean of the baseline scores, the test scores, and the difference scores.

infant	Baseline	Test	differences
1	0.44	0.6	0.16
2	0.41	0.68	0.27
3	0.75	0.72	-0.03
4	0.44	0.28	-0.16
5	0.47	0.5	0.03
Sums	2.51	2.78	0.27
Means	0.502	0.556	0.054

We can see there was a positive mean difference of 0.054, between the test and baseline phases.

Can we rush to judgment and conclude that infants are more socially attracted to individuals who have sung them a familiar song? I would hope not based on this very small sample. First, the difference in proportion looking isn't very large, and of course we recognize that this difference could have been produced by chance.

We will more formally evaluate whether this difference could have been caused by chance with the paired-samples t-test. But, before we do that, let's again calculate t and discuss what t tells us over and above what our measure of the mean of the difference scores tells us.

6.3.5 Calculate t

OK, so how do we calculate t for a paired-samples t -test? Surprise, we use the one-sample t-test formula that you already learned about! Specifically, we use the one-sample t -test formula on the difference scores. We have one sample of difference scores (you can see they are in one column), so we can use the one-sample t -test on the difference scores. Specifically, we are interested in comparing whether the mean of our difference scores came from a distribution with mean difference = 0. This is a special distribution we refer to as the **null distribution**. It is the distribution no differences. Of course, this **null distribution** can produce differences due to sampling error, but those differences are not caused by any experimental manipulation, they caused by the random sampling process.

We calculate t in a moment. Let's now consider again why we want to calculate t ? Why don't we just stick with the mean difference we already have?

Remember, the whole concept behind t , is that it gives an indication of how confident we should be in our mean. Remember, t involves a measure of the mean in the numerator, divided by a measure of variation (standard error of the sample mean) in the denominator. The resulting t value is small when the mean difference is small, or when the variation is large. So small t -values tell us that we shouldn't be that confident in the estimate of our mean difference. Large t -values occur when the mean difference is large and/or when the measure of variation is small. So, large t -values tell us that we can be more confident in the estimate of our mean difference. Let's find t for the mean difference scores. We use the same formulas as we did last time:

infant	Baseline	Test	differences	diff_from_mean	Squared_differences
1	0.44	0.6	0.16	0.106	0.011236
2	0.41	0.68	0.27	0.216	0.046656
3	0.75	0.72	-0.03	-0.084	0.0070560000000001
4	0.44	0.28	-0.16	-0.214	0.045796
5	0.47	0.5	0.03	-0.024	0.000575999999999999
Sums	2.51	2.78	0.27	0	0.11132
Means	0.502	0.556	0.054	0	0.022264
				sd	0.167
				SEM	0.075
				t	0.72

If we did this test using R, we would obtain almost the same numbers (there is a little bit of rounding in the table).

```
## 
## One Sample t-test
## 
## data: differences
## t = 0.72381, df = 4, p-value = 0.5092
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -0.1531384 0.2611384
## sample estimates:
## mean of x
```

```
##      0.054
```

Here is a quick write up of our t-test results, $t(4) = .72$, $p = .509$.

What does all of that tell us? There's a few things we haven't gotten into much yet. For example, the 4 represents degrees of freedom, which we discuss later. The important part, the t value should start to be a little bit more meaningful. We got a kind of small t -value didn't we. It's .72. What can we tell from this value? First, it is positive, so we know the mean difference is positive. The sign of the t -value is always the same as the sign of the mean difference (ours was +0.054). We can also see that the p-value was .509. We've seen p-values before. This tells us that our t value or larger, occurs about 50.9% of the time... Actually it means more than this. And, to understand it, we need to talk about the concept of two-tailed and one-tailed tests.

6.3.6 Interpreting ts

Remember what it is we are doing here. We are evaluating whether our sample data could have come from a particular kind of distribution. The null distribution of no differences. This is the distribution of t -values that would occur for samples of size 5, with a mean difference of 0, and a standard error of the sample mean of .075 (this is the SEM that we calculated from our sample). We can see what this particular null-distribution looks like by plotting it like this:

Null-distribution of t-values for our data

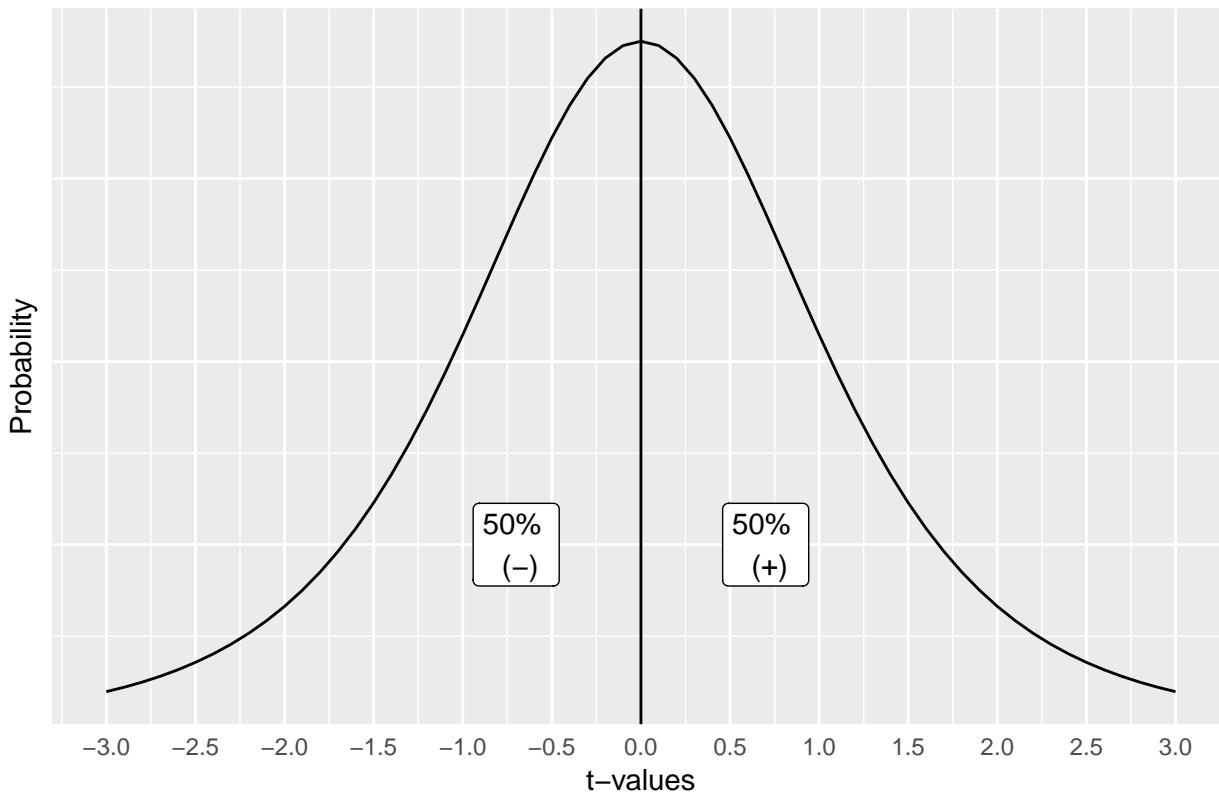


Figure 6.5: A distribution of t -values that can occur by chance alone, when there is no difference between the sample and a population

The t -distribution above shows us the kinds of values t will take by chance alone, when we measure the mean differences for pairs of 5 samples (like our current). t is most likely to be zero, which is good, because we are looking at the distribution of no-differences, which should most often be 0! But, sometimes, due to

sampling error, we can get t s that are bigger than 0, either in the positive or negative direction. Notice the distribution is symmetrical, a t from the null-distribution will be positive half of the time, and negative half of the time, that is what we would expect by chance.

So, what kind of information do we want know when we find a particular t value from our sample? We want to know how likely the t value like the one we found occurs just by chance. This is actually a subtly nuanced kind of question. For example, any particular t value doesn't have a specific probability of occurring. When we talk about probabilities, we are talking about ranges of probabilities. Let's consider some probabilities. We will use the letter p , to talk about the probabilities of particular t values.

1. What is the probability that t is zero or positive or negative? The answer is $p=1$, or 100%. We will always have a t value that is zero or non-zero...Actually, if we can't compute the t -value, for example when the standard deviation is undefined, I guess then we would have a non-number. But, assuming we can calculate t , then it will always be 0 or positive or negative.
2. What is the probability of $t = 0$ or greater than 0? The answer is $p=.5$, or 50%. 50% of t -values are 0 or greater.
3. What is the of $t = 0$ or smaller than 0? The answer is $p=.5$, or 50%. 50% of t -values are 0 or smaller.

We can answer all of those questions just by looking at our t -distribution, and dividing it into two equal regions, the left side (containing 50% of the t values), and the right side containing 50% of the t -values).

What if we wanted to take a more fine-grained approach, let's say we were interested in regions of 10%. What kinds of t s occur 10% of the time. We would apply lines like the following. Notice, the likelihood of bigger numbers (positive or negative) gets smaller, so we have to increase the width of the bars for each of the intervals between the bars to contain 10% of the t -values, it looks like this:

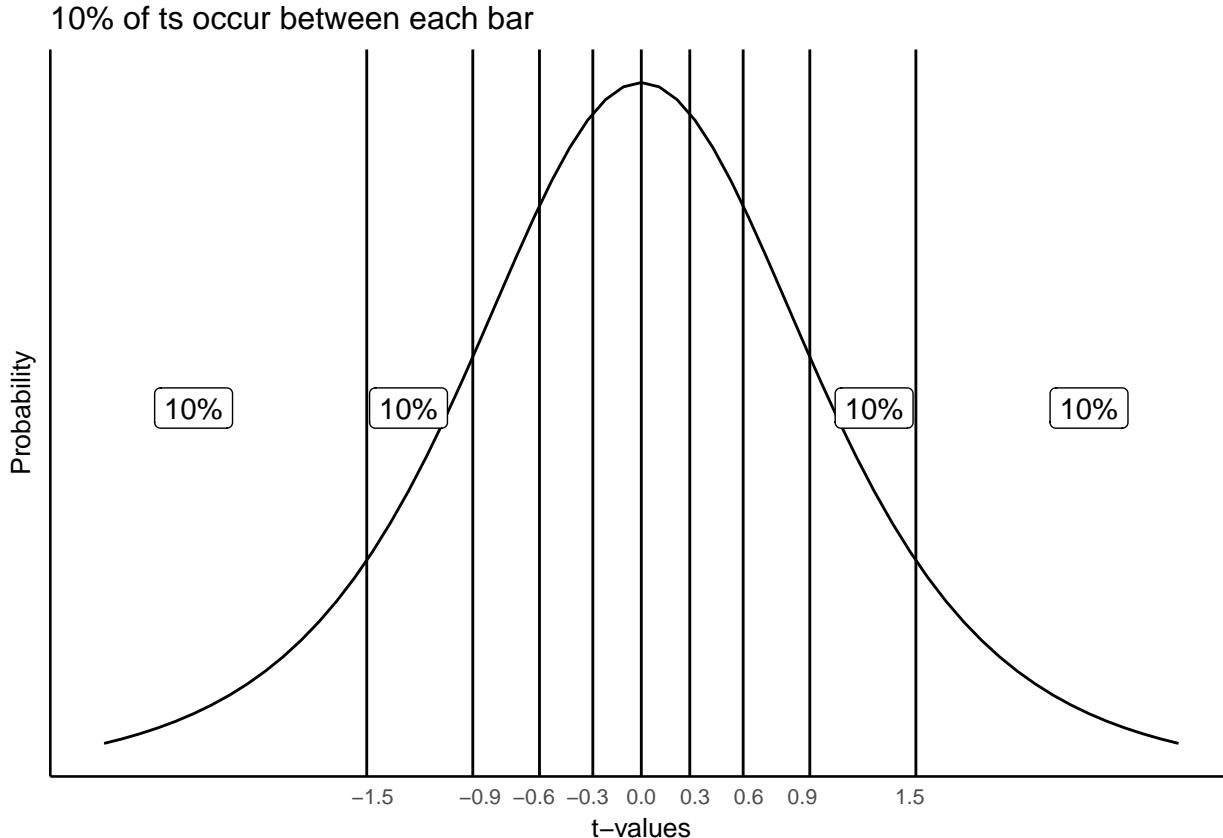


Figure 6.6: Splitting the t distribution up into regions each containing 5% of the t -values. The width between the bars narrows as they approach the center of the distribution, where there are more t -values

Consider the probabilities (p) of t for the different ranges.

1. $t \leq -1.5$ (t is less than or equal to -1.5), $p = 10\%$
2. $-1.5 \geq t \geq -0.9$ (t is equal to or between -1.5 and -0.9), $p = 10\%$
3. $-0.9 \geq t \leq -0.6$ (t is equal to or between -0.9 and -0.6), $p = 10\%$
4. $t \geq 1.5$ (t is greater than or equal to 1.5), $p = 10\%$

Notice, that the ps are always 10%. ts occur in these ranges with 10% probability.

6.3.7 Getting the p-values for t-values

You might be wondering where I am getting some of these values from. For example, how do I know that 10% of t values (for this null distribution) have a value of approximately 1.5 or greater than 1.5? The answer is I used R to tell me.

In most statistics textbooks the answer would be: there is a table at the back of the book where you can look these things up...This textbook has no such table. We could make one for you. And, we might do that. But, we didn't do that yet...

So, where do these values come from, how can you figure out what they are? The complicated answer is that we are not going to explain the math behind finding these values because, 1) the authors (some of us) admittedly don't know the math well enough to explain it, and 2) it would sidetrack us to much, 3) you will learn how to get these numbers in the lab with software, 4) you will learn how to get these numbers in lab without the math, just by doing a simulation, and 5) you can do it in R, or excel, or you can use an online calculator.

This is all to say that you can find the ts and their associated ps using software. But, the software won't tell you what these values mean. That's we are doing here. You will also see that software wants to know a few more things from you, such as the degrees of freedom for the test, and whether the test is one-tailed or two tailed. We haven't explained any of these things yet. That's what we are going to do now. Note, we explain degrees of freedom last. First, we start with a one-tailed test.

6.3.8 One-tailed tests

A **one-tailed test** is sometimes also called a directional test. It is called a directional test, because a researcher might have a hypothesis in mind suggesting that the difference they observe in their means is going to have a particular direction, either a positive difference, or a negative difference.

Typically, a researcher would set an **alpha criterion**. The alpha criterion describes a line in the sand for the researcher. Often, the alpha criterion is set at $p=.05$. What does this mean? Let's look at again at the graph of the t -distribution, and show the alpha criterion.

The figure shows that t values of $+2.13$ or greater occur 5% of the time. Because the t -distribution is symmetrical, we also know that t values of -2.13 or smaller also occur 5% of the time. Both of these properties are true under the null distribution of no differences. This means, that when there really are no differences, a researcher can expect to find t values of 2.13 or larger 5% of the time.

Let's review and connect some of the terms:

1. **alpha criterion:** the criterion set by the researcher to make decisions about whether they believe chance did or did not cause the difference. The alpha criterion here is set to $p=.05$
2. **Critical t .** The critical t is the t -value associated with the alpha-criterion. In this case for a one-tailed test, it is the t value where 5% of all ts are this number or greater. In our example, the critical t is 2.13 . 5% of all t values (with degrees of freedom = 4) are $+2.13$, or greater than $+2.13$.
3. **Observed t .** The observed t is the one that you calculated from your sample. In our example about the infants, the observed t was $t(4) = 0.72$.

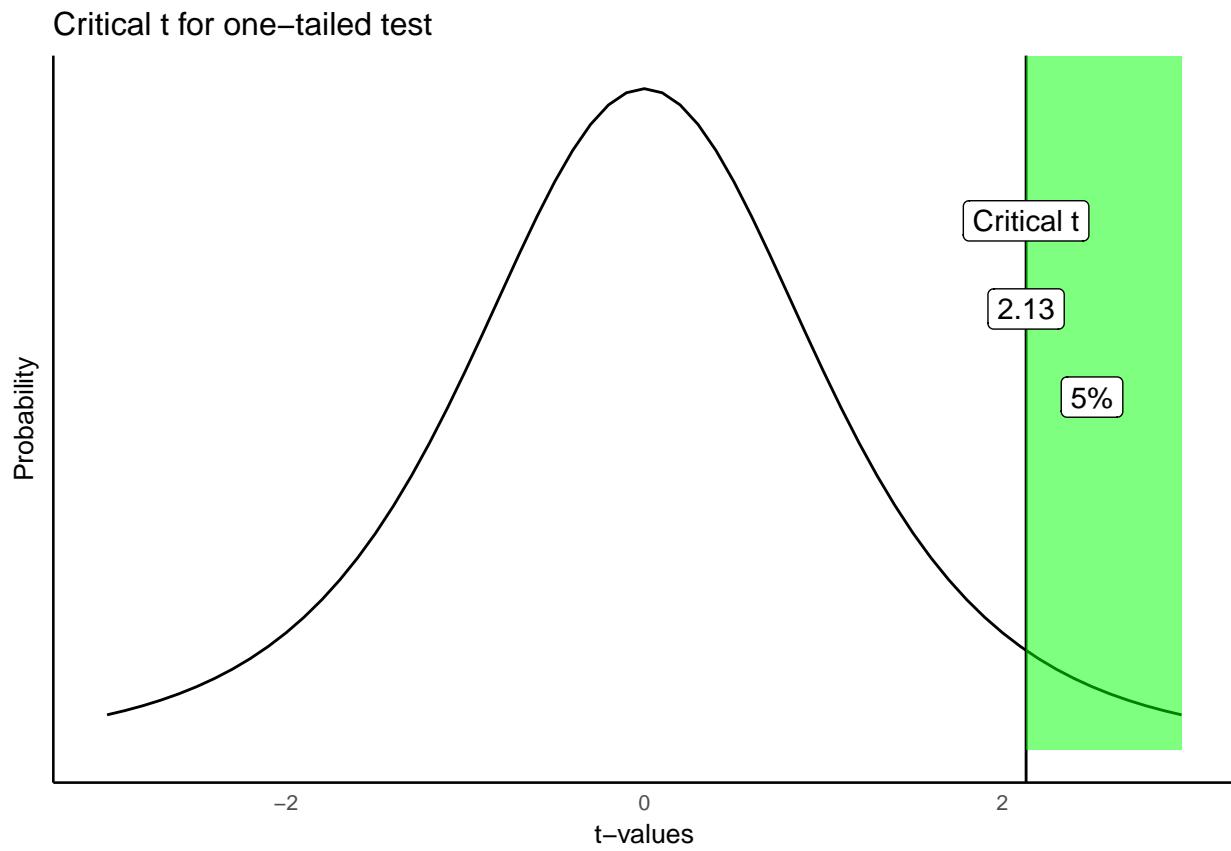


Figure 6.7: The critical value of t for an alpha criterion of 0.05. 5% of all ts are at this value or larger

4. **p-value.** The *p*-value is the probability of obtaining the observed *t* value or larger. Now, you could look back at our previous example, and find that the *p*-value for $t(4) = .72$, was $p=.509$. HOWEVER, this *p*-value was not calculated for a one-directional test... (we talk about what $.509$ means in the next section).

Let's see what the *p*-value for $t(4) = .72$ using a one-directional test would be, and what it would look like:

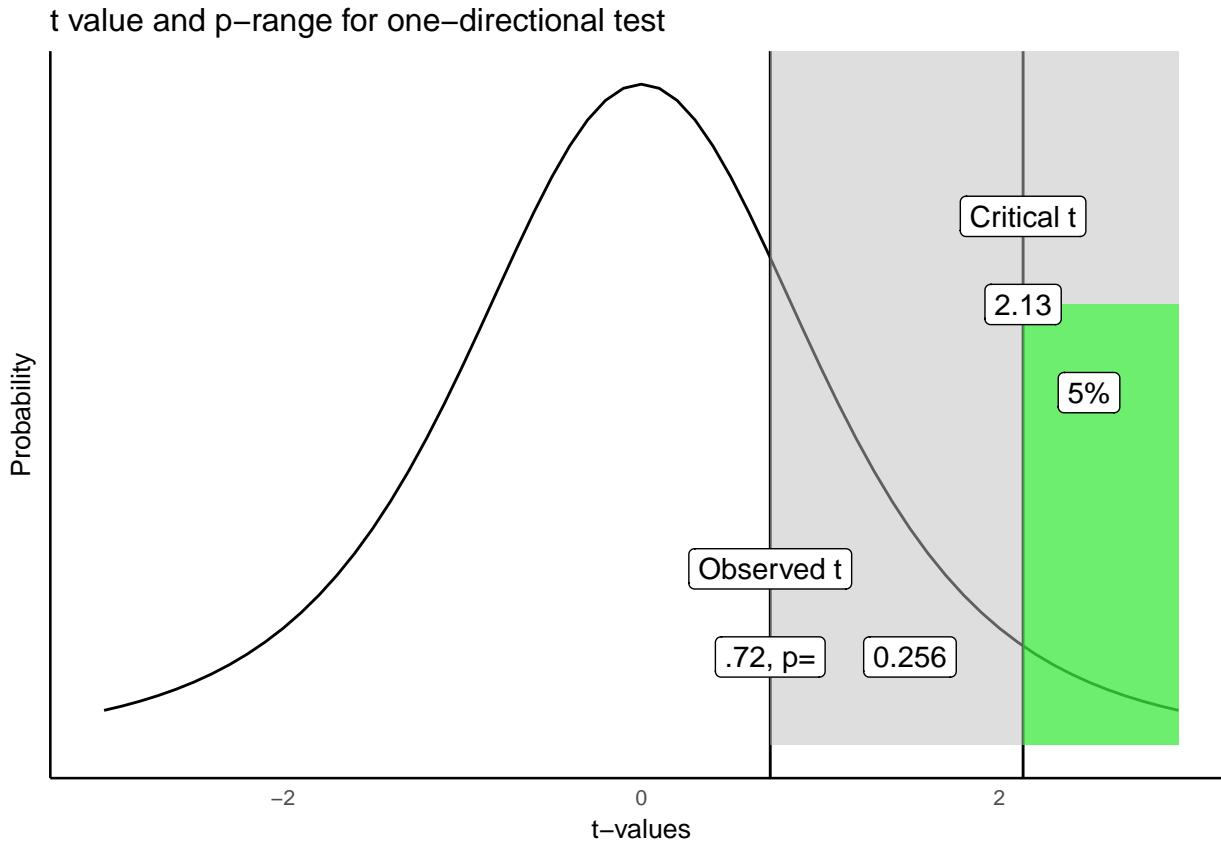


Figure 6.8: Critical value for a one-directional t-test

Let's take this one step at a time. We have located the observed *t* of $.72$ on the graph. We shaded the right region all grey. What we see is that the grey region represents $.256$ or 25.6% of all *t* values. In other words, 25.6% of *t* values are $.72$ or larger than $.72$. You could expect, by chance alone, to find a *t* value of $.72$ or larger, 25.6% of the time. That's fairly often. We did find a *t* value of $.72$. Now that you know this kind of *t* value or larger occurs 25.6% of the time, would you be confident that the mean difference was not due to chance? Probably not, given that chance can produce this difference fairly often.

Following the “standard” decision making procedure, we would claim that our *t* value was **not statistically significant**, because it was not large enough. If our observed value was larger than the critical *t* (larger than 2.13), defined by our alpha criterion, then we would claim that our *t* value was **statistically significant**. This would be equivalent to saying that we believe it is unlikely that the difference we observed was due to chance. In general, for any observed *t* value, the associated *p*-value tells you how likely a *t* of the observed size or larger would be observed. The *p*-value **always** refers to a **range** of *t*-values, never to a single *t*-value. Researchers use the alpha criterion of $.05$, as a matter of convenience and convention. There are other ways to interpret these values that do not rely on a strict (significant versus not) dichotomy.

6.3.9 Two-tailed tests

OK, so that was one-tailed tests... What are two tailed tests, what is that? The p -value that we originally calculated from our paired-samples t -test was for a 2-tailed test. Often, the default is that the p -value is for a two-tailed test.

The two-tailed test, is asking a more general question about whether a difference is likely to have been produced by chance. The question is: what is probability of any difference. It is also called a **non-directional** test, because here we don't care about the direction or sign of the difference (positive or negative), we just care if there is any kind of difference.

The same basic things as before are involved. We define an alpha criterion ($\alpha = 0.05$). And, we say that any observed t value that has a probability of $p < .05$ (p is less than .05) will be called **statistically significant**, and ones that are more likely ($p > .05$, p is greater than .05) will be called null-results, or not statistically significant. The only difference is how we draw the alpha range. Before it was on the right side of the t distribution (we were conducting a one-sided test remember, so we were only interested in one side).

Let's just take a look at what the most extreme 5% of the t -values are, when we ignore if they are positive or negative:

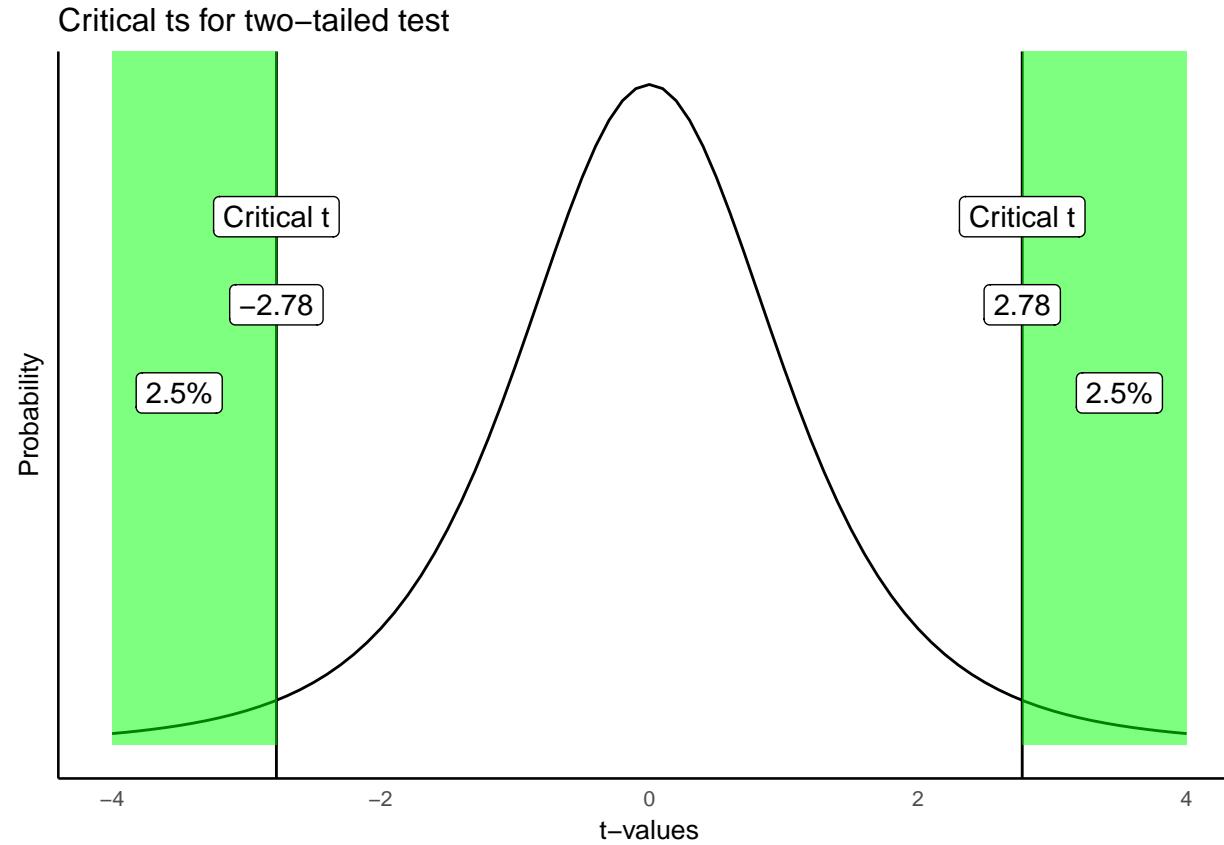


Figure 6.9: Critical values for a two-tailed test. Each line represents the location where 2.5% of all t s are larger or smaller than critical value. The total for both tails is 5%

Here is what we are seeing. A distribution of no differences (the null, which is what we are looking at), will produce t s that are 2.78 or greater 2.5% of the time, and t s that are -2.78 or smaller 2.5% of the time. 2.5% + 2.5% is a total of 5% of the time. We could also say that t s larger than ± 2.78 occur 5% of the time.

As a result, the critical t value is $(\pm) 2.78$ for a two-tailed test. As you can see, the two-tailed test is blind to the direction or sign of the difference. Because of this, the critical t value is also higher for a two-tailed

test, than for the one-tailed test that we did earlier. Hopefully, now you can see why it is called a two-tailed test. There are two tails of the distribution, one on the left and right, both shaded in green.

6.3.10 One or two tailed, which one?

Now that you know there are two kinds of tests, one-tailed, and two-tailed, which one should you use? There is some conventional wisdom on this, but also some debate. In the end, it is up to you to be able to justify your choice and why it is appropriate for your data. That is the real answer.

The conventional answer is that you use a one-tailed test when you have a theory or hypothesis that is making a directional prediction (the theory predicts that the difference will be positive, or negative). Similarly, use a two-tailed test when you are looking for any difference, and you don't have a theory that makes a directional prediction (it just makes the prediction that there will be a difference, either positive or negative).

Also, people appear to choose one or two-tailed tests based on how risky they are as researchers. If you always ran one-tailed tests, your critical t values for your set alpha criterion would always be smaller than the critical ts for a two-tailed test. Over the long run, you would make more type I errors, because the criterion to detect an effect is a lower bar for one than two tailed tests.

Remember type 1 errors occur when you reject the idea that chance could have caused your difference. You often never know when you make this error. It happens anytime that sampling error was the actual cause of the difference, but a researcher dismisses that possibility and concludes that their manipulation caused the difference.

Similarly, if you always ran two-tailed tests, even when you had a directional prediction, you would make fewer type I errors over the long run, because the t for a two-tailed test is higher than the t for a one-tailed test. It seems quite common for researchers to use a more conservative two-tailed test, even when they are making a directional prediction based on theory. In practice, researchers tend to adopt a standard for reporting that is common in their field. Whether or not the practice is justifiable can sometimes be an open question. The important task for any researcher, or student learning statistics, is to be able to justify their choice of test.

6.3.11 Degrees of freedom

Before we finish up with paired-samples t -tests, we should talk about degrees of freedom. Our sense is that students don't really understand degrees of freedom very well. If you are reading this textbook, you are probably still wondering what is degrees of freedom, seeing as we haven't really talked about it all.

For the t -test, there is a formula for degrees of freedom. For the one-sample and paired sample t -tests, the formula is:

Degrees of Freedom = $df = n - 1$. Where n is the number of samples in the test.

In our paired t -test example, there were 5 infants. Therefore, degrees of freedom = $5-1 = 4$.

OK, that's a formula. Who cares about degrees of freedom, what does the number mean? And why do we report it when we report a t -test... you've probably noticed the number in parentheses e.g., $t(4)=.72$, the 4 is the df , or degrees of freedom.

Degrees of freedom is both a concept, and a correction. The concept is that if you estimate a property of the numbers, and you use this estimate, you will be forcing some constraints on your numbers.

Consider the numbers: 1, 2, 3. The mean of these numbers is 2. Now, let's say I told you that the mean of three numbers is 2. Then, how many of these three numbers have freedom? Funny question right. What we mean is, how many of the three numbers could be any number, or have the freedom to be any number.

The first two numbers could be any number. But, once those two numbers are set, the final number (the third number), MUST be a particular number that makes the mean 2. The first two numbers have freedom. The third number has no freedom.

To illustrate. Let's freely pick two numbers: 51 and -3. I used my personal freedom to pick those two numbers. Now, if our three numbers are 51, -3, and x, and the mean of these three numbers is 2. There is only one solution, x has to be -42, otherwise the mean won't be 2. This is one way to think about degrees of freedom. The degrees of freedom for these three numbers is $n-1 = 3-1 = 2$, because 2 of the numbers can be free, but the last number has no freedom, it becomes fixed after the first two are decided.

Now, statisticians often apply degrees of freedom to their calculations, especially when a second calculation relies on an estimated value. For example, when we calculate the standard deviation of a sample, we first calculate the mean of the sample right! By estimating the mean, we are fixing an aspect of our sample, and so, our sample now has $n-1$ degrees of freedom when we calculate the standard deviation (remember for the sample standard deviation, we divide by $n-1$...there's that $n-1$ again.)

6.3.11.1 Simulating how degrees of freedom affects the t distribution

There are at least two ways to think the degrees of freedom for a *t*-test. For example, if you want to use math to compute aspects of the *t* distribution, then you need the degrees of freedom to plug in to the formula... If you want to see the formulas I'm talking about, scroll down on the *t*-test wikipedia page and look for the probability density or cumulative distribution functions...We think that is quite scary for most people, and one reason why degrees of freedom are not well-understood.

If we wanted to simulate the *t* distribution we could more easily see what influence degrees of freedom has on the shape of the distribution. Remember, *t* is a sample statistic, it is something we measure from the sample. So, we could simulate the process of measuring *t* from many different samples, then plot the histogram of *t* to show us the simulated *t* distribution.

Notice that the red distribution for $df = 4$, is a little bit shorter, and a little bit wider than the bluey-green distribution for $df = 100$. As degrees of freedom increase, the *t*-distribution gets taller (in the middle), and narrower in the range. It gets more peaky. Can you guess the reason for this? Remember, we are estimating a sample statistic, and degrees of freedom is really just a number that refers to the number of subjects (well minus one). And, we already know that as we increase *n*, our sample statistics become better estimates (less variance) of the distributional parameters they are estimating. So, *t* becomes a better estimate of its "true" value as sample size increase, resulting in a more narrow distribution of *ts*.

There is a slightly different *t* distribution for every degrees of freedom, and the critical regions associated with 5% of the extreme values are thus slightly different every time. This is why we report the degrees of freedom for each *t*-test, they define the distribution of *t* values for the sample-size in question. Why do we use $n-1$ and not *n*? Well, we calculate *t* using the sample standard deviation to estimate the standard error or the mean, that estimate uses $n-1$ in the denominator, so our *t* distribution is built assuming $n-1$. That's enough for degrees of freedom...

6.4 The paired samples t-test strikes back

You must be wondering if we will ever be finished talking about paired samples *t*-tests... why are we doing round 2, oh no! Don't worry, we're just going to 1) remind you about what we were doing with the infant study, and 2) do a paired samples *t*-test on the entire data set and discuss.

Remember, we were wondering if the infants would look longer toward the singer who sang the familiar song during the test phase compared to the baseline phase. We showed you data from 5 infants, and walked through the computations for the *t*-test. As a reminder, it looked like this:

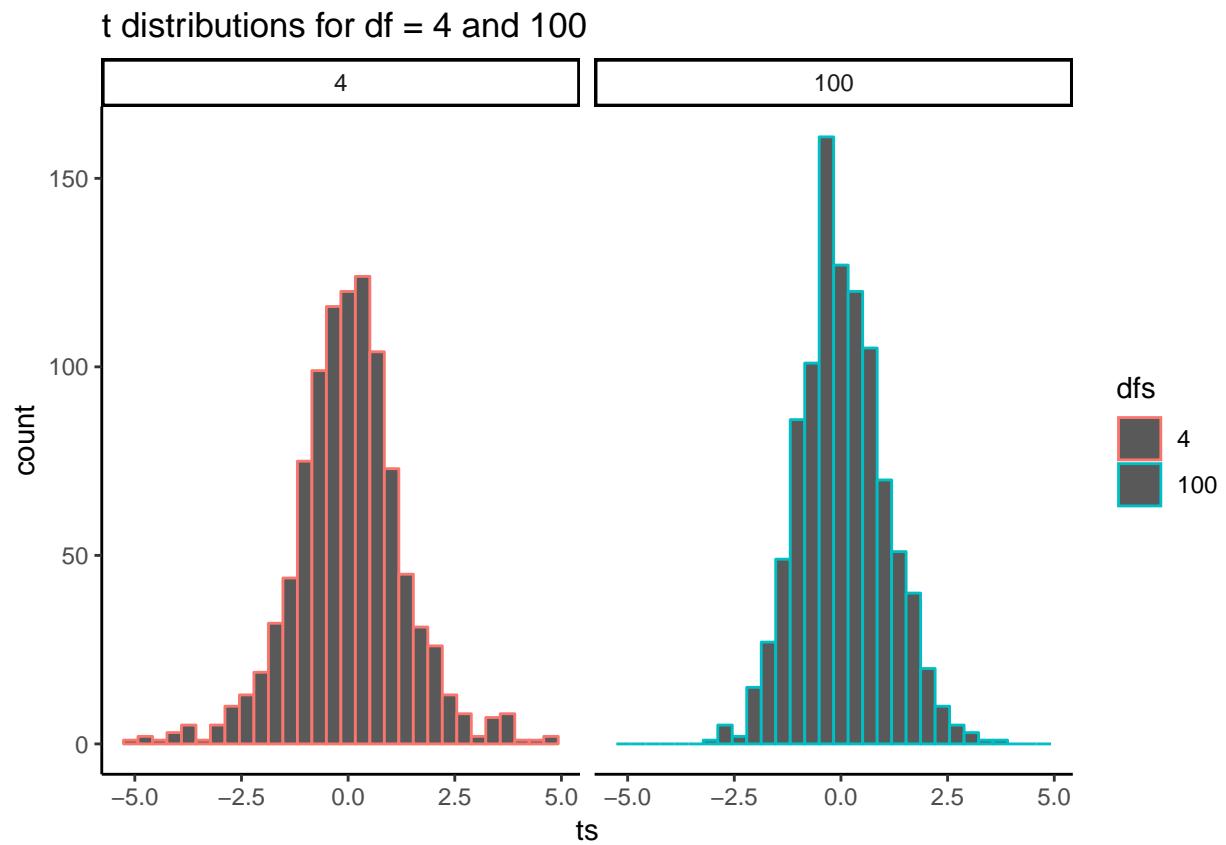


Figure 6.10: The width of the t distribution shrinks as sample size increases

infant	Baseline	Test	differences	diff_from_mean	Squared_differences
1	0.44	0.6	0.16	0.106	0.011236
2	0.41	0.68	0.27	0.216	0.046656
3	0.75	0.72	-0.03	-0.084	0.00705600000000001
4	0.44	0.28	-0.16	-0.214	0.045796
5	0.47	0.5	0.03	-0.024	0.000575999999999999
Sums	2.51	2.78	0.27	0	0.11132
Means	0.502	0.556	0.054	0	0.022264
				sd	0.167
				SEM	0.075
				t	0.72

```
##
## One Sample t-test
##
## data: round(differences, digits = 2)
## t = 0.72381, df = 4, p-value = 0.5092
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -0.1531384 0.2611384
## sample estimates:
## mean of x
## 0.054
```

Let's write down the finding one more time: The mean difference was 0.054, $t(4) = .72$, $p = .509$. We can also now confirm, that the p -value was from a two-tailed test. So, what does this all really mean.

We can say that a t value with an absolute of .72 or larger occurs 50.9% of the time. More precisely, the distribution of no differences (the null), will produce a t value this large or larger 50.9% of the time. In other words, chance alone good have easily produced the t value from our sample, and the mean difference we observed or .054, could easily have been a result of chance.

Let's quickly put all of the data in the t -test, and re-run the test using all of the infant subjects.

```
##
## One Sample t-test
##
## data: differences
## t = 2.4388, df = 31, p-value = 0.02066
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 0.01192088 0.13370412
## sample estimates:
## mean of x
## 0.0728125
```

Now we get a very different answer. We would summarize the results saying the mean difference was .073, $t(31) = 2.44$, $p = 0.020$. How many total infants were their? Well the degrees of freedom was 31, so there must have been 32 infants in the study. Now we see a much smaller p -value. This was also a two-tailed test, so we that observing a t value of 2.4 or greater (absolute value) only occurs 2% of the time. In other words, the distribution of no differences will produce the observed t-value very rarely. So, it is unlikely that the observed mean difference of .073 was due to chance (it could have been due to chance, but that is very unlikely). As a result, we can be somewhat confident in concluding that something about seeing and hearing a unfamiliar person sing a familiar song, causes an infant to draw their attention toward the singer, and this potentially benefits social learning on the part of the infant.

6.5 Independent samples t-test: The return of the t-test?

If you've been following the Star Wars references, we are on last movie (of the original trilogy)... the independent t-test. This is were basically the same story plays out as before, only slightly different.

Remember there are different *t*-tests for different kinds of research designs. When your design is a **between-subjects** design, you use an **independent samples t-test**. Between-subjects design involve different people or subjects in each experimental condition. If there are two conditions, and 10 people in each, then there are 20 total people. And, there are no paired scores, because every single person is measured once, not twice, no repeated measures. Because there are no repeated measures we can't look at the difference scores between conditions one and two. The scores are not paired in any meaningful way, so it doesn't make sense to subtract them. So what do we do?

The logic of the independent samples t-test is the very same as the other *t*-tests. We calculated the means for each group, then we find the difference. That goes into the numerator of the *t* formula. Then we get an estimate of the variation for the denominator. We divide the mean difference by the estimate of the variation, and we get *t*. It's the same as before.

The only wrinkle here is what goes into the denominator? How should we calculate the estimate of the variance? It would be nice if we could do something very straightforward like this, say for an experiment with two groups A and B:

$$t = \frac{\bar{A} - \bar{B}}{\left(\frac{SEM_A + SEM_B}{2} \right)}$$

In plain language, this is just:

1. Find the mean difference for the top part
2. Compute the SEM (standard error of the mean) for each group, and average them together to make a single estimate, pooling over both samples.

This would be nice, but unfortunately, it turns out that finding the average of two standard errors of the mean is not the best way to do it. This would create a biased estimator of the variation for the hypothesized distribution of no differences. We won't go into the math here, but instead of the above formula, we can use a different one that gives us an **unbiased estimate of the pooled standard error of the sample mean**. Our new and improved *t* formula would look like this:

$$t = \frac{\bar{X}_A - \bar{X}_B}{s_p * \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

and, s_p , which is the pooled sample standard deviation is defined as, note the s's in the formula are variances:

$$s_p = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}$$

Believe you me, that is so much more formula than I wanted to type out. Shall we do one independent *t*-test example by hand, just to see the computations? Let's do it...but in a slightly different way than you expect. I show the steps using R. I made some fake scores for groups A and B. Then, I followed all of the steps from the formula, but made R do each of the calculations. This shows you the needed steps by following the code. At the end, I print the *t*-test values I computed "by hand", and then the *t*-test value that the R software outputs using the *t*-test function. You should be able to get the same values for *t*, if you were brave enough to compute *t* by hand.

```
## By "hand" using R r code
a <- c(1,2,3,4,5)
b <- c(3,5,4,7,9)

mean_difference <- mean(a)-mean(b) # compute mean difference

variance_a <- var(a) # compute variance for A
variance_b <- var(b) # compute variance for B
```

```

# Compute top part and bottom part of sp formula

sp_numerator <- (4*variance_a + 4* variance_b)
sp_denominator <- 5+5-2
sp <- sqrt(sp_numerator/sp_denominator) # compute sp

# compute t following formulat

t <- mean_difference / ( sp * sqrt( (1/5) +(1/5) ) )

t # print results

## [1] -2.017991

# using the R function t.test
t.test(a,b, paired=FALSE, var.equal = TRUE)

##
## Two Sample t-test
##
## data: a and b
## t = -2.018, df = 8, p-value = 0.0783
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -5.5710785 0.3710785
## sample estimates:
## mean of x mean of y
## 3.0 5.6

```

6.6 Simulating data for t-tests

An “advanced” topic for *t*-tests is the idea of using R to simulations for *t*-tests.

If you recall, *t* is a property of a sample. We calculate *t* from our sample. The *t* distribution is the hypothetical behavior of our sample. That is, if we had taken thousands upon thousands of samples, and calculated *t* for each one, and then looked at the distribution of those *t*’s, we would have the sampling distribution of *t*!

It can be very useful to get in the habit of using R to simulate data under certain conditions, to see how your sample data, and things like *t* behave. Why is this useful? It mainly prepares you with some intuitions about how sampling error (random chance) can influence your results, given specific parameters of your design, such as sample-size, the size of the mean difference you expect to find in your data, and the amount of variation you might find. These methods can be used formally to conduct power-analyses. Or more informally for data sense.

6.6.1 Simulating a one-sample t-test

Here are the steps you might follow to simulate data for a one sample *t*-test.

1. Make some assumptions about what your sample (that you might be planning to collect) might look like. For example, you might be planning to collect 30 subjects worth of data. The scores of those data points might come from a normal distribution (mean = 50, SD = 10).

2. sample simulated numbers from the distribution, then conduct a t -test on the simulated numbers. Save the statistics you want (such as ts and ps), and then see how things behave.

Let's do this a couple different times. First, let's simulate samples with $N = 30$, taken from a normal ($\text{mean} = 50$, $\text{SD} = 25$). We'll do a simulation with 1000 simulations. For each simulation, we will compare the sample mean with a population mean of 50. There should be no difference on average here, this is the null distribution that we are simulating. The distribution of no differences

```
# steps to create fake data from a distribution
# and conduct t-tests on the simulated data
save_ps <- length(1000)
save_ts <- length(1000)
for ( i in 1:1000 ){
  my_sample <- rnorm(n=30, mean =50, sd =25)
  t_test <- t.test (my_sample, mu = 50)
  save_ps[i] <- t_test$p.value
  save_ts[i] <- t_test$statistic
}

#plot histograms of t and p values for 1000 simulations
hist(save_ts)
```

Histogram of save_ts

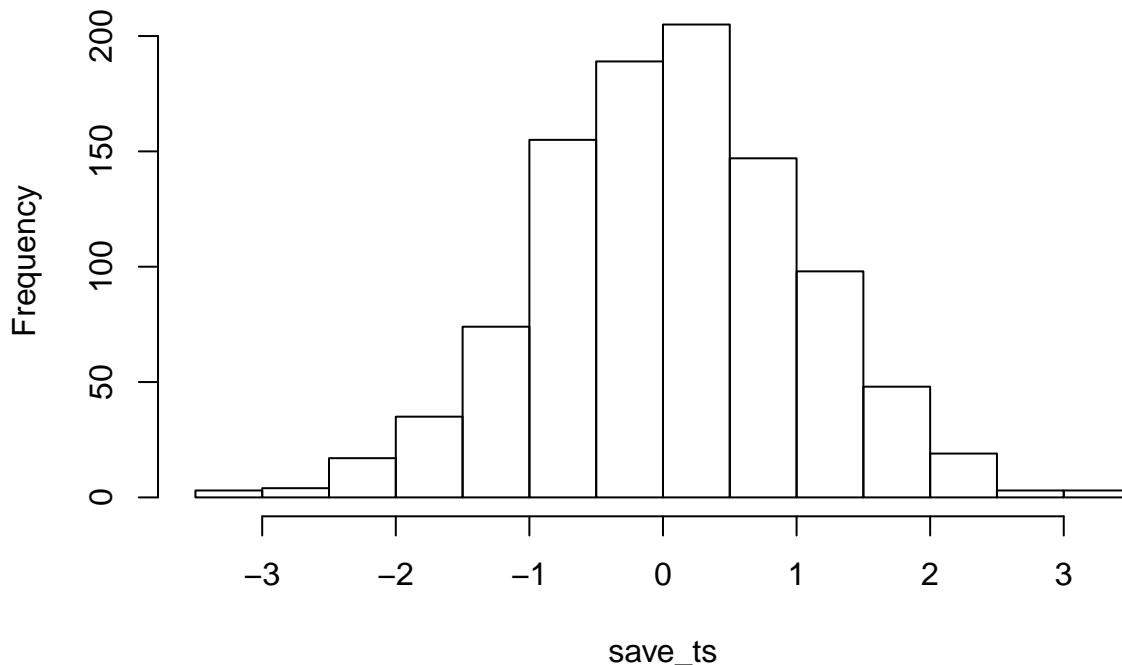


Figure 6.11: The distribution of p-values is flat under the null

```
hist(save_ps)
```

Neat. We see both a t distribution, that looks like t distribution as it should. And we see the p distribution. This shows us how often we get t values of particular sizes. You may find it interesting that the p -distribution is flat under the null, which we are simulating here. This means that you have the same chances of getting a t with a p -value between 0 and 0.05, as you would for getting a t with a p -value between .90 and .95. Those ranges are both ranges of 5%, so there are an equal amount of t values in them by definition.

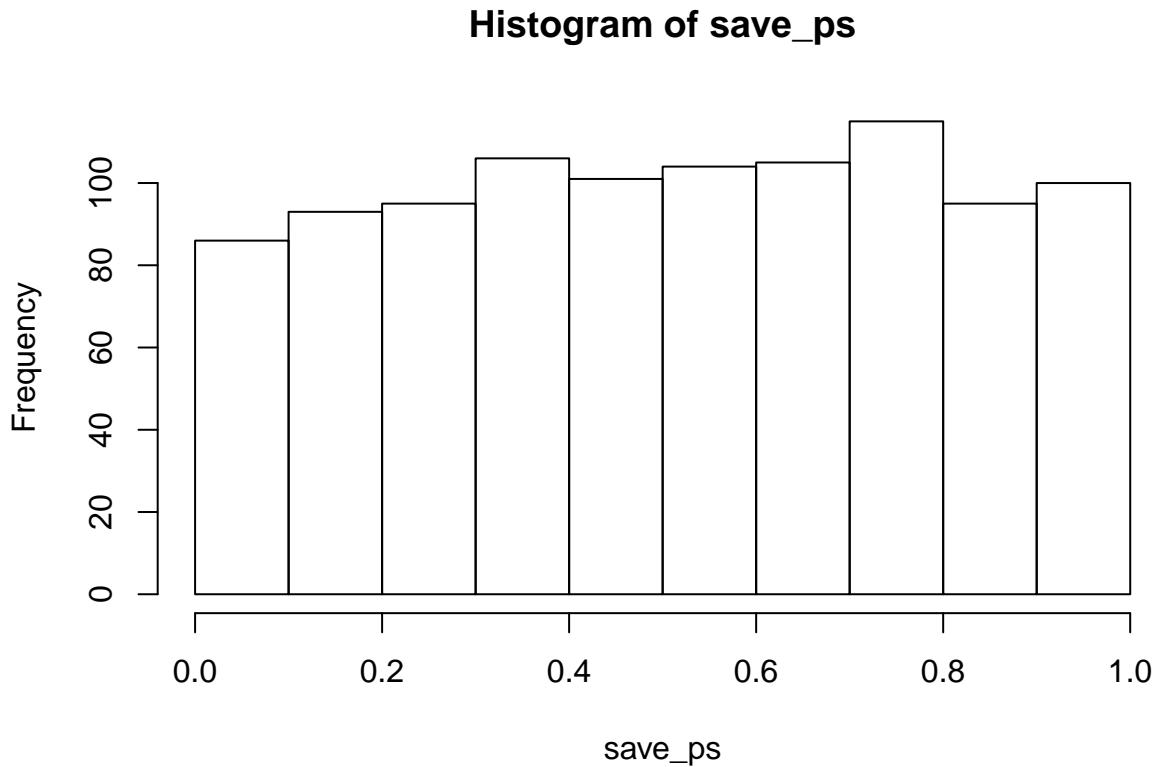


Figure 6.12: The distribution of p-values is flat under the null

Here's another way to do the same simulation in R, using the `replicate` function, instead a for loop:

```
simulated_ts <- replicate(1000,
                           t.test(rnorm(30,50,25))$statistic)
hist(simulated_ts)

simulated_ps <- replicate(1000,
                           t.test(rnorm(30,50,25))$p.value)
hist(simulated_ps)
```

6.6.2 Simulating a paired samples t-test

The code below is set up to sample 10 scores for condition A and B from the same normal distribution. The simulation is conducted 1000 times, and the *ts* and *ps* are saved and plotted for each.

```
save_ps <- length(1000)
save_ts <- length(1000)
for ( i in 1:1000 ){
  condition_A <- rnorm(10,10,5)
  condition_B <- rnorm(10,10,5)
  differences <- condition_A - condition_B
  t_test <- t.test(differences, mu=0)
  save_ps[i] <- t_test$p.value
  save_ts[i] <- t_test$statistic
}
```

According to the simulation. When there are no differences between the conditions, and the samples are

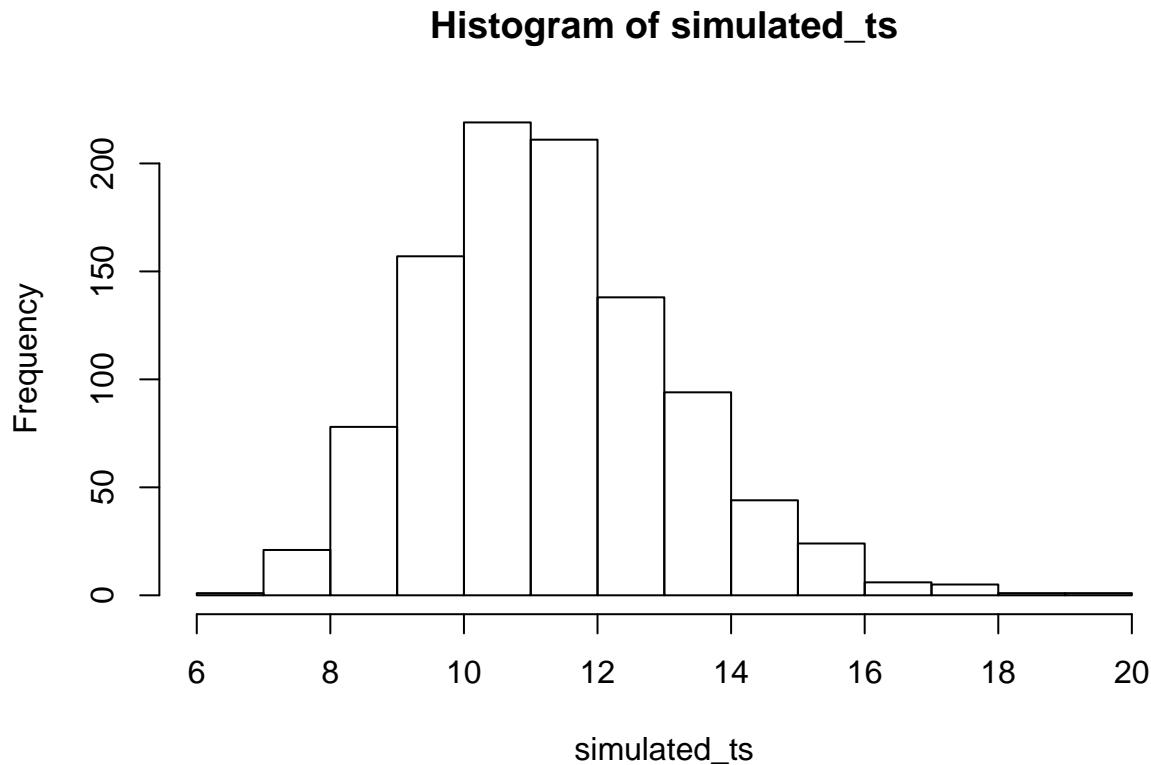


Figure 6.13: Simulating ts in R

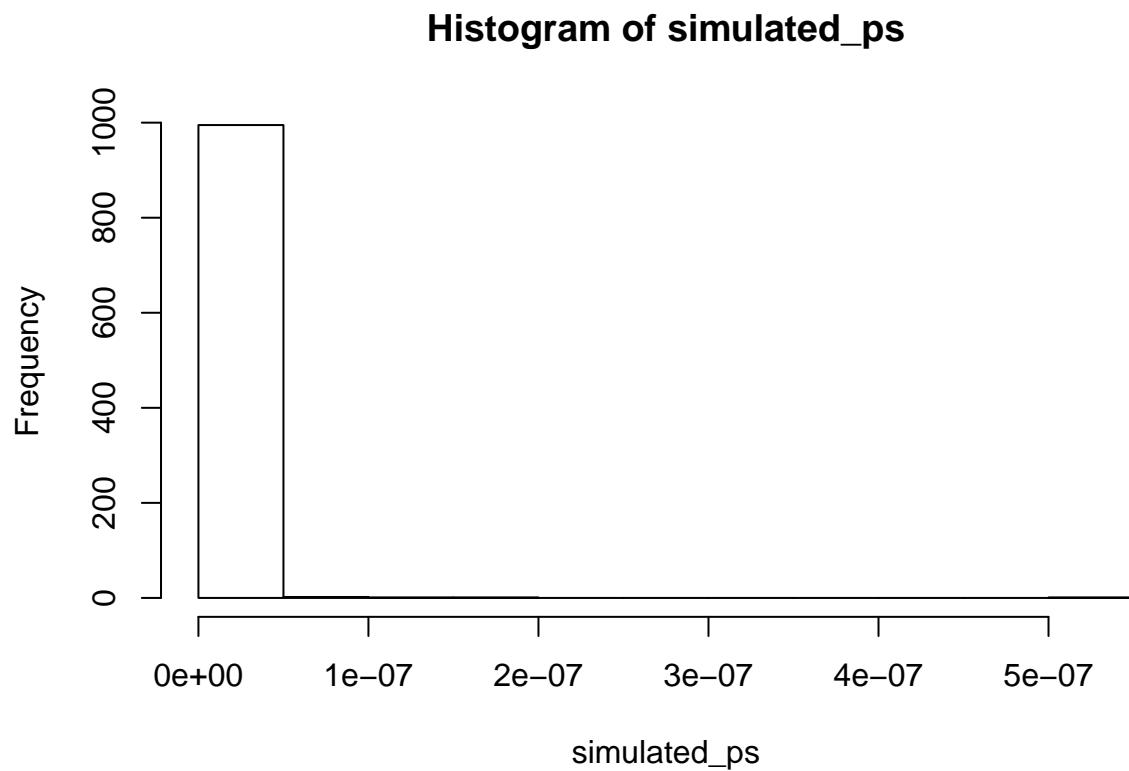


Figure 6.14: Simulating ps in R

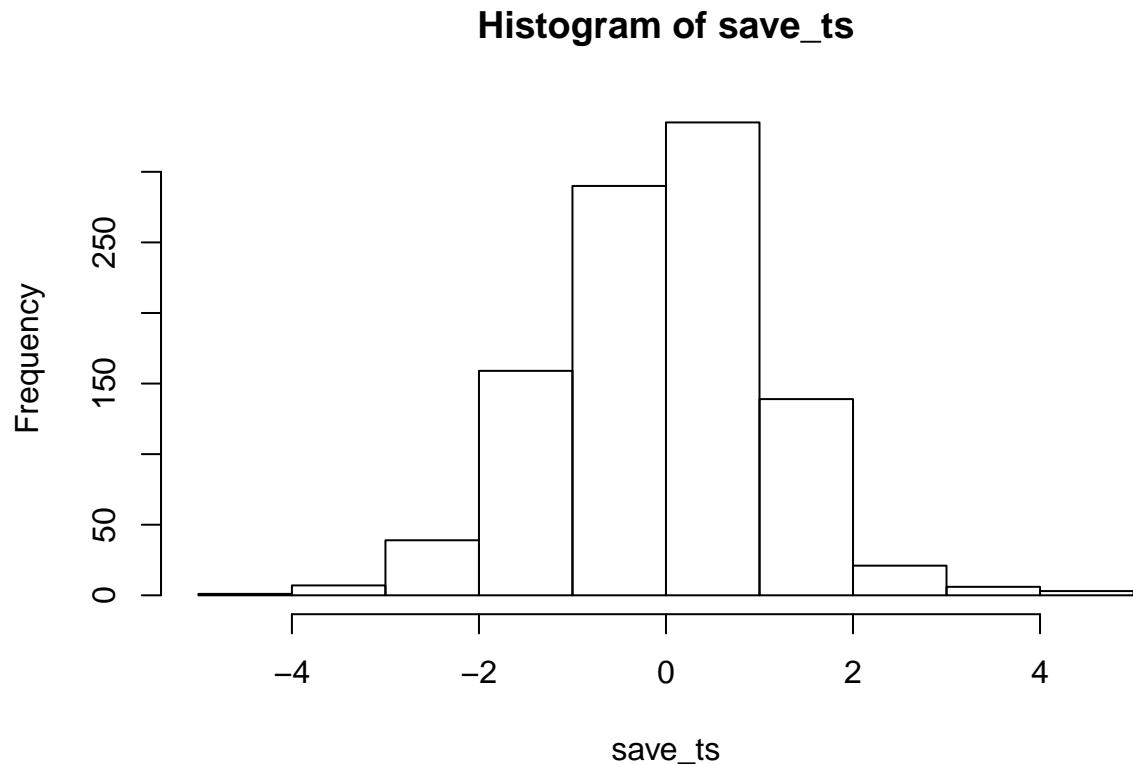


Figure 6.15: 1000 simulated ts from the null distribution

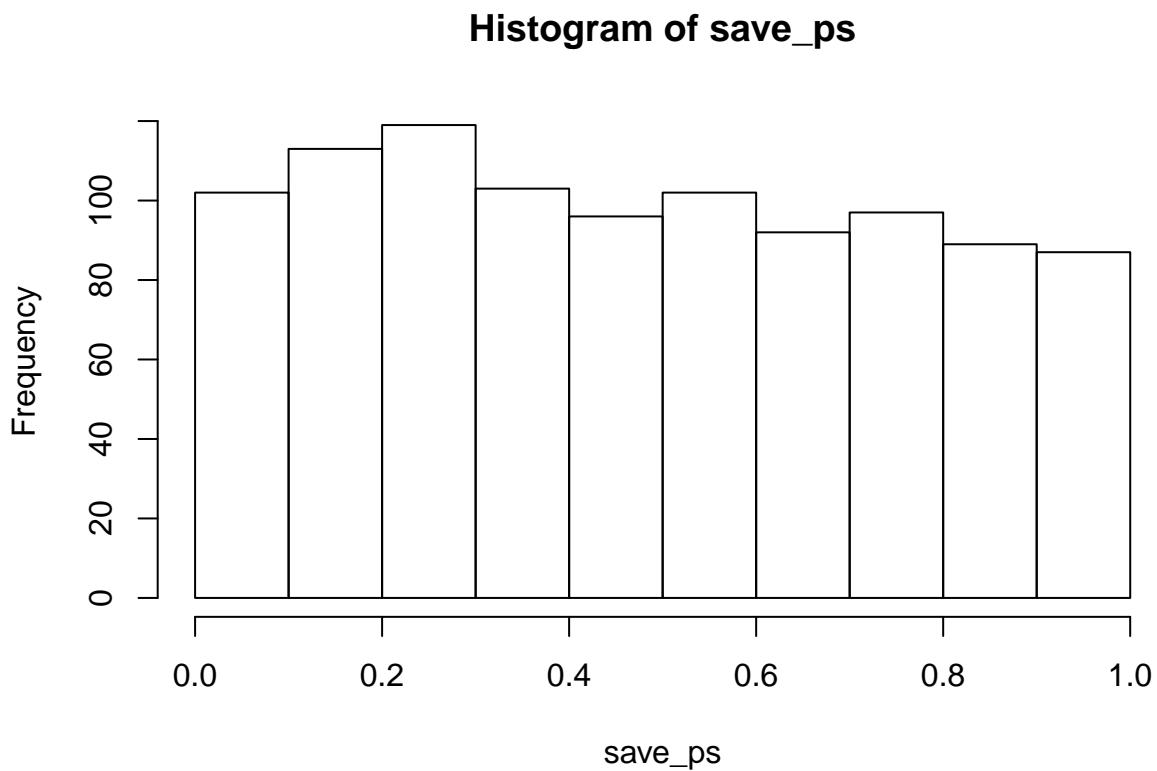


Figure 6.16: 1000 simulated ps from the null distribution

being pulled from the very same distribution, you get these two distributions for t and p . These again show how the null distribution of no differences behaves.

For any of these simulations, if you rejected the null-hypothesis (that your difference was only due to chance), you would be making a type I error. If you set your alpha criteria to $\alpha = .05$, we can ask how many type I errors were made in these 1000 simulations. The answer is:

```
length(save_ps[save_ps<.05])
## [1] 53
length(save_ps[save_ps<.05])/1000
## [1] 0.053
```

We happened to make 53. The expectation over the long run is 5% type I error rates (if your alpha is .05).

What happens if there actually is a difference in the simulated data, let's set one condition to have a larger mean than the other:

```
save_ps <- length(1000)
save_ts <- length(1000)
for ( i in 1:1000 ){
  condition_A <- rnorm(10,10,5)
  condition_B <- rnorm(10,13,5)
  differences <- condition_A - condition_B
  t_test <- t.test(differences, mu=0)
  save_ps[i] <- t_test$p.value
  save_ts[i] <- t_test$statistic
}
```

Histogram of save_ts

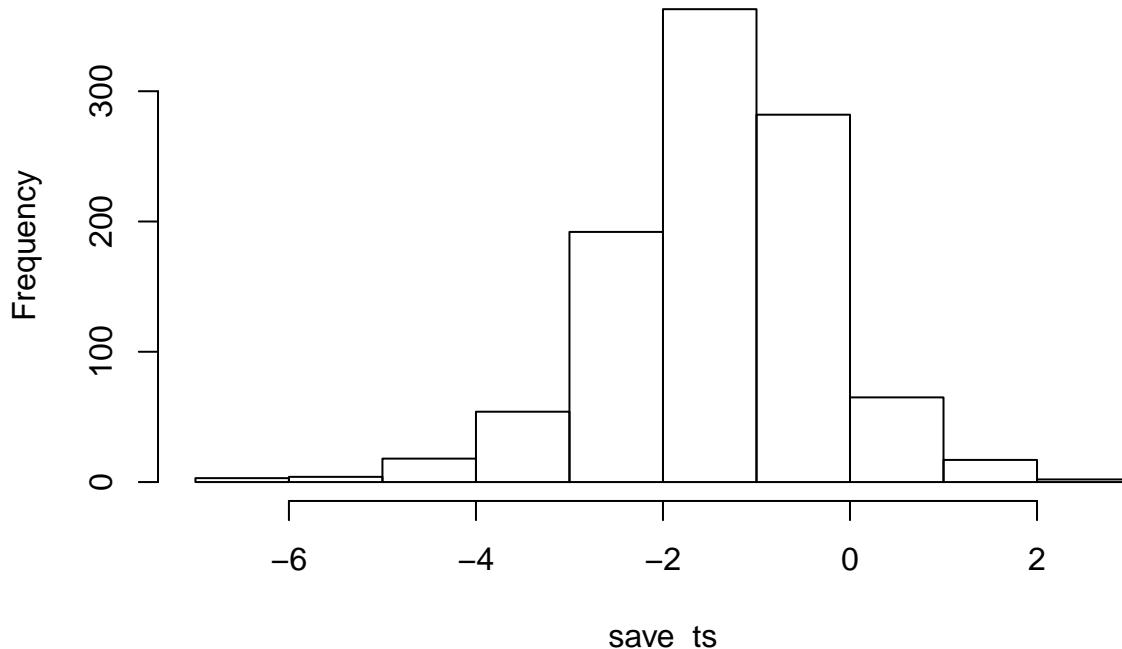


Figure 6.17: 1000 ts when there is a true difference

Histogram of save_ps

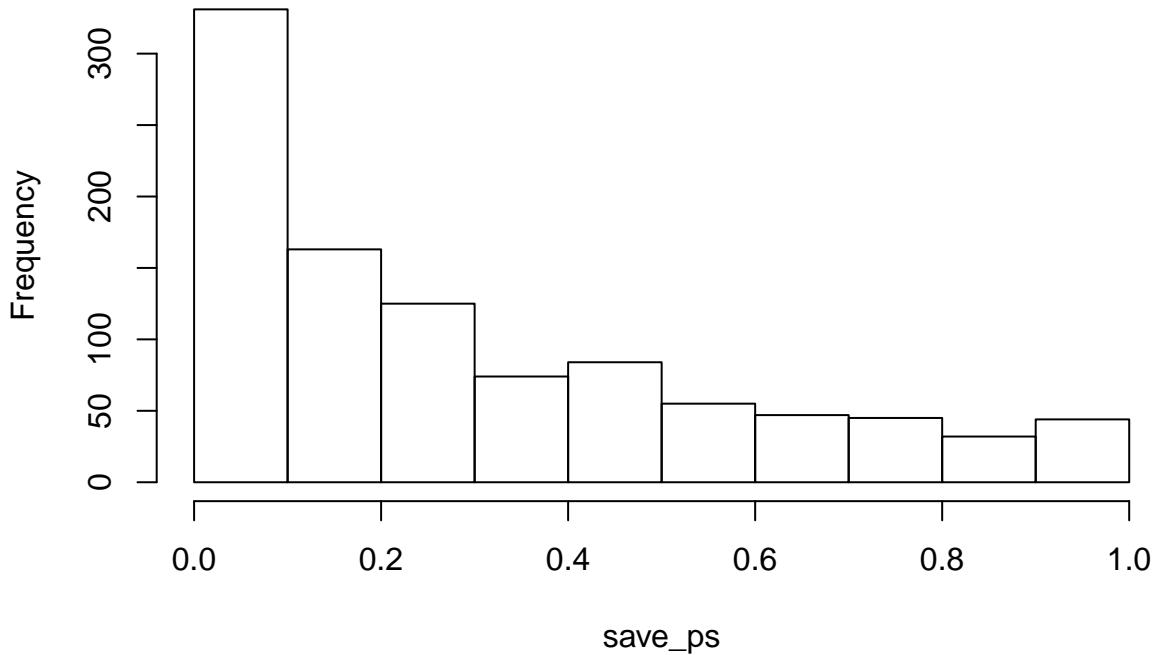


Figure 6.18: 1000 ps when there is a true difference

Now you can see that the p -value distribution is skewed to the left. This is because when there is a true effect, you will get p -values that are less than .05 more often. Or, rather, you get larger t values than you normally would if there were no differences.

In this case, we wouldn't be making a type I error if we rejected the null when p was smaller than .05. How many times would we do that out of our 1000 experiments?

```
length(save_ps[save_ps<.05])
## [1] 205
length(save_ps[save_ps<.05])/1000
## [1] 0.205
```

We happened to get 205 simulations where p was less than .05, that's only 0.205 experiments. If you were the researcher, would you want to run an experiment that would be successful only 0.205 of the time? I wouldn't. I would run a better experiment.

How would you run a better simulated experiment? Well, you could increase n , the number of subjects in the experiment. Let's increase n from 10 to 100, and see what happens to the number of "significant" simulated experiments.

```
save_ps <- length(1000)
save_ts <- length(1000)
for ( i in 1:1000 ){
  condition_A <- rnorm(100,10,5)
  condition_B <- rnorm(100,13,5)
  differences <- condition_A - condition_B
  t_test <- t.test(differences, mu=0)
```

```

  save_ps[i] <- t_test$p.value
  save_ts[i] <- t_test$statistic
}

```

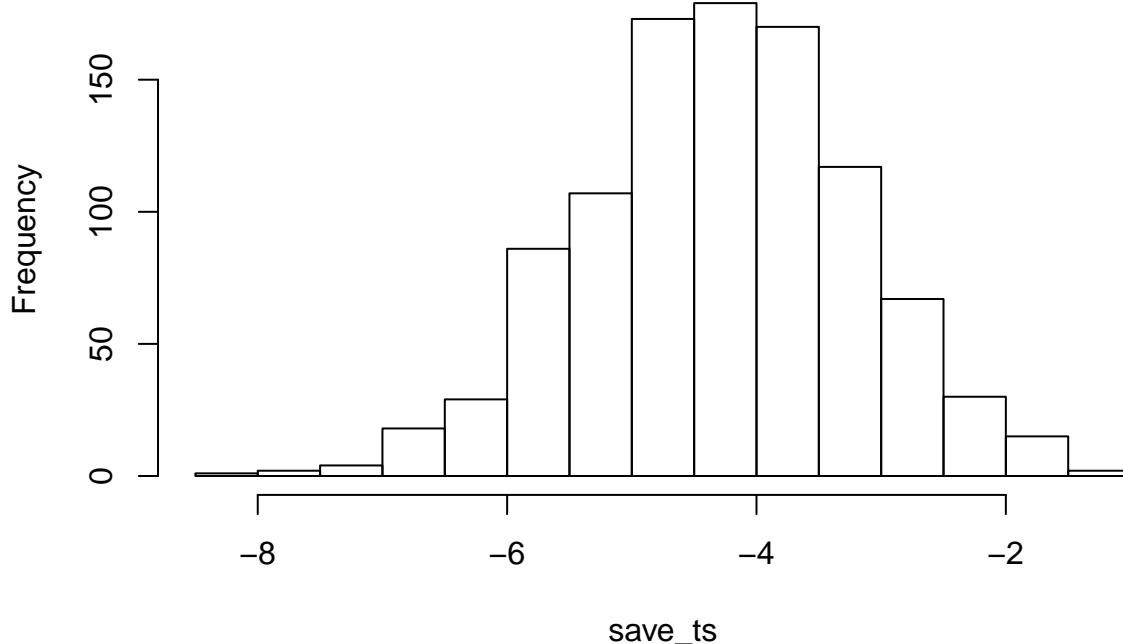
Histogram of save_ts

Figure 6.19: 1000 ts for n = 100, when there is a true effect

```

## [1] 984
## [1] 0.984

```

Cool, now almost all of the experiments show a *p*-value of less than .05 (using a two-tailed test, that's the default in R). See, you could use this simulation process to determine how many subjects you need to reliably find your effect.

6.6.3 Simulating an independent samples t.test

Just change the t.test function like so... this is for the null, assuming no difference between groups.

```

save_ps <- length(1000)
save_ts <- length(1000)
for ( i in 1:1000 ){
  group_A <- rnorm(10,10,5)
  group_B <- rnorm(10,10,5)
  t_test <- t.test(group_A, group_B, paired=FALSE, var.equal=TRUE)
  save_ps[i] <- t_test$p.value
  save_ts[i] <- t_test$statistic
}

```

```

## [1] 60

```

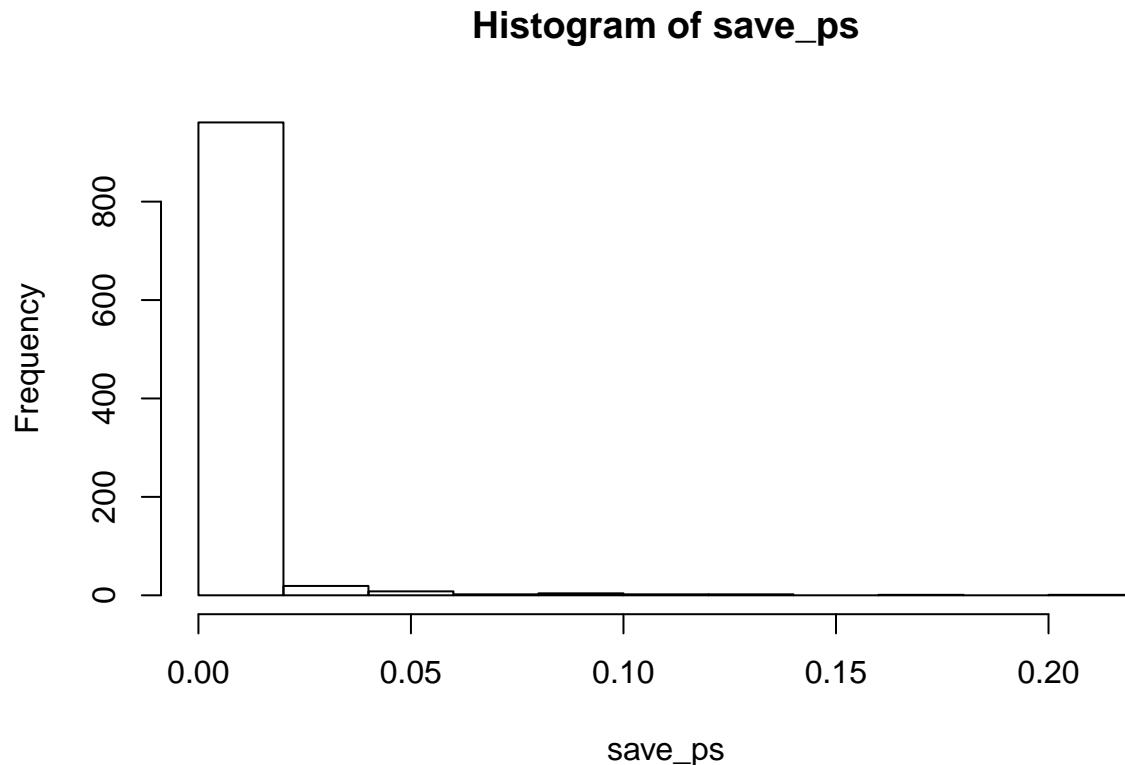


Figure 6.20: 1000 ps for $n = 100$, when there is a true effect

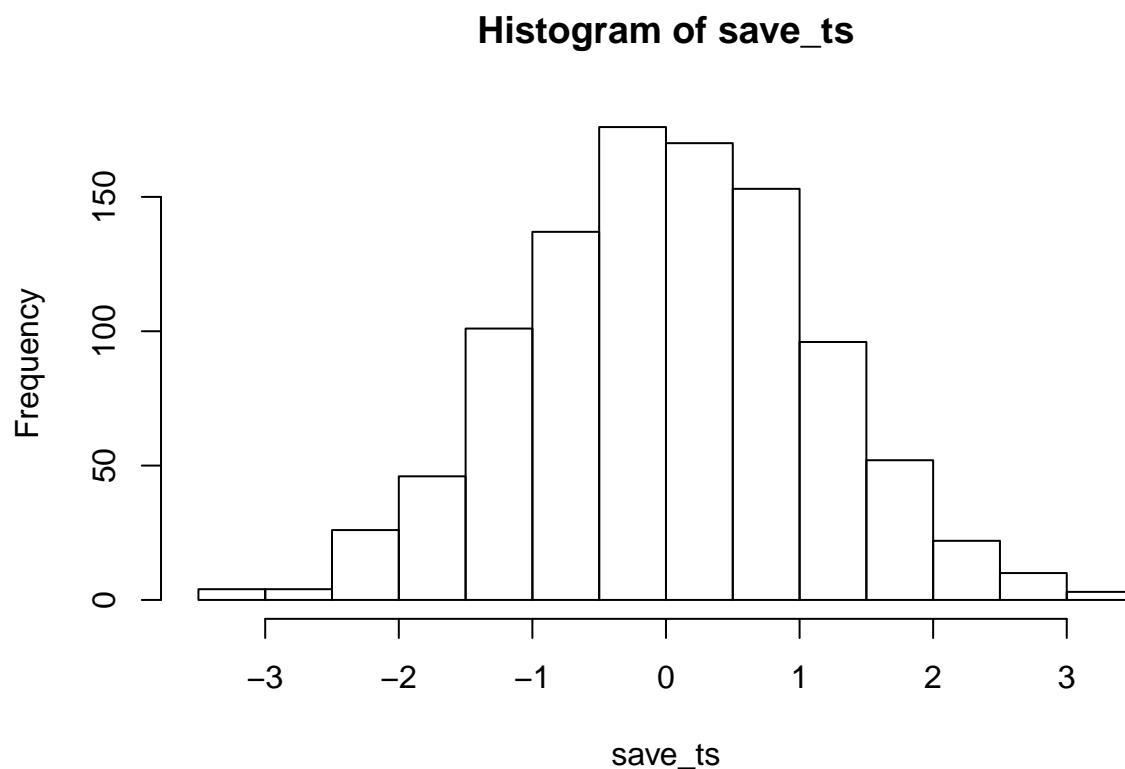


Figure 6.21: 1000 ts for $n = 100$, when there is a true effect

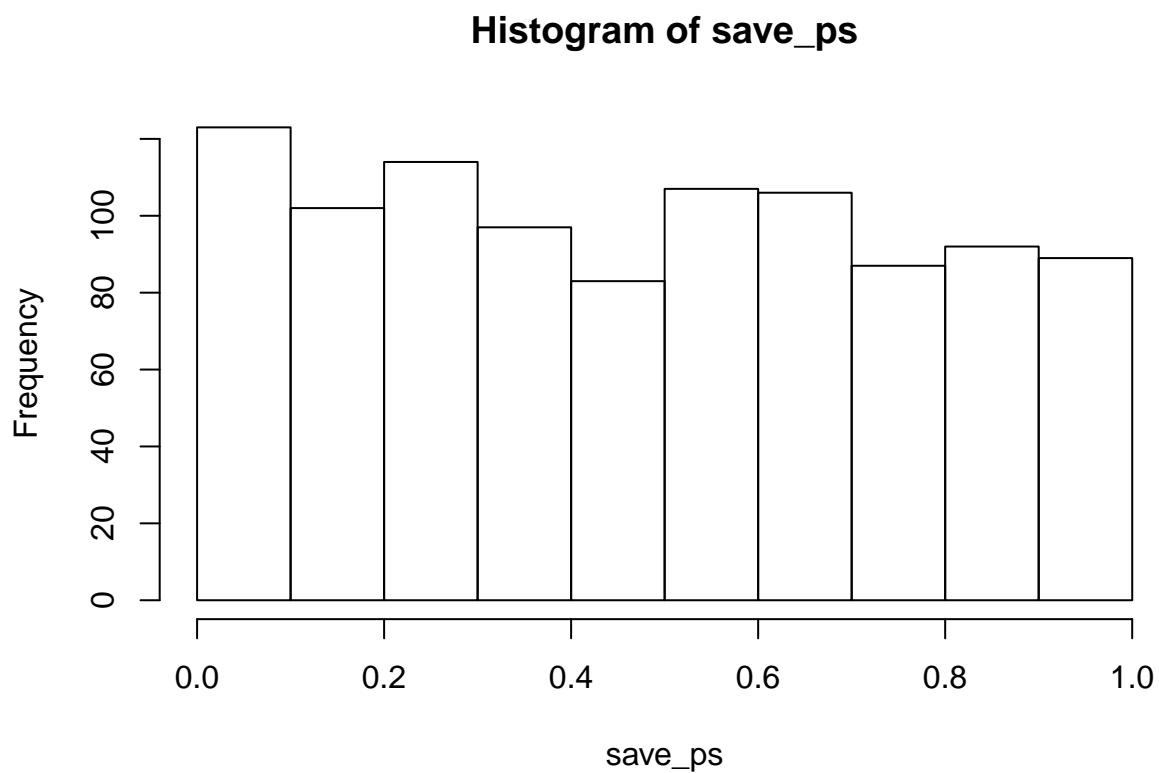


Figure 6.22: 1000 ps for n = 100, when there is a true effect

```
## [1] 0.06
```


Chapter 7

ANOVA

A fun bit of stats history (Salsburg, 2001). Sir Ronald Fisher invented the ANOVA, which we learn about in this section. He wanted to publish his new test in the journal Biometrika. The editor at the time was Karl Pearson (remember Pearson's r for correlation?). Pearson and Fisher were apparently not on good terms, they didn't like each other. Pearson refused to publish Fisher's new test. So, Fisher eventually published his work in the Journal of Agricultural Science. Funnily enough, the feud continued onto the next generation. Years after Fisher published his ANOVA, Karl Pearson's son Egon Pearson, and Jerzy Neyman revamped Fisher's ideas, and re-cast them into what is commonly known as null vs. alternative hypothesis testing. Fisher didn't like this very much.

We present the ANOVA in the Fisherian sense, and at the end describe the Neyman-Pearson approach that invokes the concept of null vs. alternative hypotheses.

7.1 ANOVA is Analysis of Variance

ANOVA stands for Analysis Of Variance. It is a widely used technique for assessing the likelihood that differences found between means in sample data could be produced by chance. You might be thinking, well don't we have t -tests for that? Why do we need the ANOVA, what do we get that's new that we didn't have before?

What's new with the ANOVA, is the ability to test a wider range of means beyond just two. In all of the t -test examples we were always comparing two things. For example, we might ask whether the difference between two sample means could have been produced by chance. What if our experiment had more than two conditions or groups? We would have more than 2 means. We would have one mean for each group or condition. That could be a lot depending on the experiment. How would we compare all of those means? What should we do, run a lot of t -tests, comparing every possible combination of means? Actually, you could do that. Or, you could do an ANOVA.

In practice, we will combine both the ANOVA test and t -tests when analyzing data with many sample means (from more than two groups or conditions). Just like the t -test, there are different kinds of ANOVAs for different research designs. There is one for between-subjects designs, and a slightly different one for repeated measures designs. We talk about both, beginning with the ANOVA for between-subjects designs.

7.2 One-factor ANOVA

The one-factor ANOVA is sometimes also called a between-subjects ANOVA, an independent factor ANOVA, or a one-way ANOVA (which is a bit of a misnomer as we discuss later). The critical ingredient for a one-

factor, between-subjects ANOVA, is that you have one independent variable, with at least two-levels. When you have one IV with two levels, you can run a *t*-test. You can also run an ANOVA. Interestingly, they give you almost the exact same results. You will get a *p*-value from both tests that is identical (they are really doing the same thing under the hood). The *t*-test gives a *t*-value as the important sample statistic. The ANOVA gives you the *F*-value (for Fisher, the inventor of the test) as the important sample statistic. It turns out that t^2 equals *F*, when there are only two groups in the design. They are the same test. Side-note, it turns out they are all related to Pearson's r too (but we haven't written about this relationship yet in this textbook).

Remember that *t* is computed directly from the data. It's like a mean and standard error that we measure from the sample. In fact it's the mean difference divided by the standard error of the sample. It's just another descriptive statistic isn't it.

The same thing is true about *F*. *F* is computed directly from the data. In fact, the idea behind *F* is the same basic idea that goes into making *t*. Here is the general idea behind the formula, it is again a ratio of the effect we are measuring (in the numerator), and the variation associated with the effect (in the denominator).

$$\text{name of statistic} = \frac{\text{measure of effect}}{\text{measure of error}}$$

$$F = \frac{\text{measure of effect}}{\text{measure of error}}$$

The difference with *F*, is that we use variances to describe both the measure of the effect and the measure of error. So, *F* is a ratio of two variances.

Remember what we said about how these ratios work. When the variance associated with the effect is the same size as the variance associated with sampling error, we will get two of the same numbers, this will result in an *F*-value of 1. When the variance due to the effect is larger than the variance associated with sampling error, then *F* will be greater than 1. When the variance associated with the effect is smaller than the variance associated with sampling error, *F* will be less than one.

Let's rewrite in plainer English. We are talking about two concepts that we would like to measure from our data. 1) A measure of what we can explain, and 2) a measure of error, or stuff about our data we can't explain. So, the *F* formula looks like this:

$$F = \frac{\text{Can Explain}}{\text{Can't Explain}}$$

When we can explain as much as we can't explain, *F* = 1. This isn't that great of a situation for us to be in. It means we have a lot of uncertainty. When we can explain much more than we can't we are doing a good job, *F* will be greater than 1. When we can explain less than what we can't, we really can't explain very much, *F* will be less than 1. That's the concept behind making *F*.

If you saw an *F* in the wild, and it was .6. Then you would automatically know the researchers couldn't explain much of their data. If you saw an *F* of 5, then you would know the researchers could explain 5 times more than the couldn't, that's pretty good. And the point of this is to give you an intuition about the meaning of an *F*-value, even before you know how to compute it.

7.2.1 Computing the *F*-value

Fisher's ANOVA is very elegant in my opinion. It starts us off with a big problem we always have with data. We have a lot of numbers, and there is a lot of variation in the numbers, what to do? Wouldn't it be nice to split up the variation into kinds, or sources. If we could know what parts of the variation were being caused by our experimental manipulation, and what parts were being caused by sampling error, we would be making really good progress. We would be able to know if our experimental manipulation was causing more change in the data than sampling error, or chance alone. If we could measure those two parts of the total variation, we could make a ratio, and then we would have an *F* value. This is what the ANOVA does. It splits the total variation in the data into two parts. The formula is:

$$\text{Total Variation} = \text{Variation due to Manipulation} + \text{Variation due to sampling error}$$

This is a nice idea, but it is also vague. We haven't specified our measure of variation. What should we use? Remember the sums of squares that we used to make the variance and the standard deviation? That's what we'll use. Let's take another look at the formula, using sums of squares for the measure of variation:

$$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

7.2.2 SS Total

The total sums of squares, or SST_{Total} is a way of thinking about all of the variation in a set of data. It's pretty straightforward to measure. No tricky business. All we do is find the difference between each score and the grand mean, then we square the differences and add them all up.

Let's imagine we had some data in three groups, A, B, and C. For example, we might have 3 scores in each group. The data could look like this:

groups	scores	diff	diff_squared
A	20	13	169
A	11	4	16
A	2	-5	25
B	6	-1	1
B	2	-5	25
B	7	0	0
C	2	-5	25
C	11	4	16
C	2	-5	25
Sums	63	0	302
Means	7	0	33.555555555556

The data is organized in long format, so that each row is a single score. There are three scores for the A, B, and C groups. The mean of all of the scores is called the **Grand Mean**. It's calculated in the table, the Grand Mean = 7.

We also calculated all of the difference scores **from the Grand Mean**. The difference scores are in the column titled **diff**. Next, we squared the difference scores, and those are in the next column called **diff_squared**.

Remember, the difference scores are a way of measuring variation. They represent how far each number is from the Grand Mean. If the Grand Mean represents our best guess at summarizing the data, the difference scores represent the error between the guess and each actual data point. The only problem with the difference scores is that they sum to zero (because the mean is the balancing point in the data). So, it is convenient to square the difference scores, this turns all of them into positive numbers. The size of the squared difference scores still represents error between the mean and each score. And, the squaring operation exacerbates the differences as the error grows larger (squaring a big number makes a really big number, squaring a small number still makes a smallish number).

OK fine! We have the squared deviations from the grand mean, we know that they represent the error between the grand mean and each score. What next? SUM THEM UP!

When you add up all of the individual squared deviations (difference scores) you get the sums of squares. That's why it's called the sums of squares (SS).

Now, we have the first part of our answer:

$$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

$$SS_{\text{total}} = 302 \text{ and}$$

$$302 = SS_{\text{Effect}} + SS_{\text{Error}}$$

What next? If you think back to what you learned about algebra, and solving for X, you might notice that we don't really need to find the answers to both missing parts of the equation. We only need one, and we can solve for the other. For example, if we found SS_{Effect} , then we could solve for SS_{Error} .

7.2.3 SS Effect

SS_{Total} gave us a number representing all of the change in our data, how all the scores are different from the grand mean.

What we want to do next is estimate how much of the total change in the data might be due to the experimental manipulation. For example, if we ran an experiment that causes changes in the measurement, then the means for each group will be different from other. As a result, the manipulation forces change onto the numbers, and this will naturally mean that some part of the total variation in the numbers is caused by the manipulation.

The way to isolate the variation due to the manipulation (also called effect) is to look at the means in each group, and calculate the difference scores between each group mean and the grand mean, and then sum the squared deviations to find SS_{Effect} .

Consider this table, showing the calculations for SS_{Effect} .

groups	scores	means	diff	diff_squared
A	20	11	4	16
A	11	11	4	16
A	2	11	4	16
B	6	5	-2	4
B	2	5	-2	4
B	7	5	-2	4
C	2	5	-2	4
C	11	5	-2	4
C	2	5	-2	4
Sums	63	63	0	72
Means	7	7	0	8

Notice we created a new column called `means`. For example, the mean for group A was 11. You can see there are three 11s, one for each observation in row A. The means for group B and C happen to both be 5. So, the rest of the numbers in the means column are 5s.

What we are doing here is thinking of each score in the data from the viewpoint of the group means. The group means are our best attempt to summarize the data in those groups. From the point of view of the mean, all of the numbers are treated as the same. The mean doesn't know how far off it is from each score, it just knows that all of the scores are centered on the mean.

Let's pretend you are the mean for group A. That means you are an 11. Someone asks you "hey, what's the score for the first data point in group A?". Because you are the mean, you say, I know that, it's 11. "What about the second score?"...it's 11... they're all 11, so far as I can tell..."Am I missing something...", asked the mean.

Now that we have converted each score to its mean value we can find the differences between each mean score and the grand mean, then square them, then sum them up. We did that, and found that the $SS_{\text{Effect}} = 72$.

SS_{Effect} represents the amount of variation that is caused by differences between the means. I also refer to this as the amount of variation that the researcher can explain (by the means, which represent differences between groups or conditions that were manipulated by the researcher).

Notice also that $SS_{\text{Effect}} = 72$, and that 72 is smaller than $SS_{\text{total}} = 302$. That is very important. SS_{Effect} by definition can never be larger than SS_{total} .

7.2.4 SS Error

Great, we made it to SS Error. We already found SS Total, and SS Effect, so now we can solve for SS Error just like this:

$$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

switching around:

$$SS_{\text{Error}} = SS_{\text{total}} - SS_{\text{Effect}}$$

$$SS_{\text{Error}} = 302 - 72 = 230$$

We could stop here and show you the rest of the ANOVA, we're almost there. But, the next step might not make sense unless we show you how to calculate SS_{Error} directly from the data, rather than just solving for it. We should do this just to double-check our work anyway.

groups	scores	means	diff	diff_squared
A	20	11	-9	81
A	11	11	0	0
A	2	11	9	81
B	6	5	-1	1
B	2	5	3	9
B	7	5	-2	4
C	2	5	3	9
C	11	5	-6	36
C	2	5	3	9
Sums	63	63	0	230
Means	7	7	0	25.55555555555556

Alright, we did almost the same thing as we did to find SS_{Effect} . Can you spot the difference? This time for each score we first found the group mean, then we found the error in the group mean estimate for each score. In other words, the values in the *diff* column are the differences between each score and its group mean. The values in the *diff_squared* column are the squared deviations. When we sum up the squared deviations, we get another Sums of Squares, this time it's the SS_{Error} . This is an appropriate name, because these deviations are the ones that the group means can't explain!

7.2.5 Degrees of freedom

Degrees of freedom come into play again with ANOVA. This time, their purpose is a little bit more clear. *Dfs* can be fairly simple when we are doing a relatively simple ANOVA like this one, but they can become complicated when designs get more complicated.

Let's talk about the degrees of freedom for the SS_{Effect} and SS_{Error} .

The formula for the degrees of freedom for SS_{Effect} is

$$df_{\text{Effect}} = \text{Groups} - 1, \text{ where Groups is the number of groups in the design.}$$

In our example, there are 3 groups, so the df is $3-1 = 2$. You can think of the df for the effect this way. When we estimate the grand mean (the overall mean), we are taking away a degree of freedom for the group means. Two of the group means can be anything they want (they have complete freedom), but in order for all three to be consistent with the Grand Mean, the last group mean has to be fixed.

The formula for the degrees of freedom for SS_{Error} is

$df_{\text{Error}} = \text{scores} - \text{groups}$, or the number of scores minus the number of groups. We have 9 scores and 3 groups, so our *df* for the error term is $9-3 = 6$. Remember, when we computed the difference score between each score and its group mean, we had to compute three means (one for each group) to do that. So, that

reduces the degrees of freedom by 3. 6 of the difference scores could be anything they want, but the last 3 have to be fixed to match the means from the groups.

7.2.6 Mean Squared Error

OK, so we have the degrees of freedom. What's next? There are two steps left. First we divide the $SSes$ by their respective degrees of freedom to create something new called Mean Squared Error. Let's talk about why we do this.

First of all, remember we are trying to accomplish this goal:

$$F = \frac{\text{measure of effect}}{\text{measure of error}}$$

We want to build a ratio that divides a measure of an effect by a measure of error. Perhaps you noticed that we already have a measure of an effect and error! How about the SS_{Effect} and SS_{Error} . They both represent the variation due to the effect, and the leftover variation that is unexplained. Why don't we just do this?

$$\frac{SS_{\text{Effect}}}{SS_{\text{Error}}}$$

Well, of course you could do that. What would happen is you can get some really big and small numbers for your inferential statistic. And, the kind of number you would get wouldn't be readily interpretable like a t value or a z score.

The solution is to **normalize** the SS terms. Don't worry, normalize is just a fancy word for taking the average, or finding the mean. Remember, the SS terms are all sums. And, each sum represents a different number of underlying properties.

For example, the SS_{Effect} represents the sum of variation for three means in our study. We might ask the question, well, what is the average amount of variation for each mean... You might think to divide SS_{Effect} by 3, because there are three means, but because we are estimating this property, we divide by the degrees of freedom instead ($\# \text{ groups} - 1 = 3 - 1 = 2$). Now we have created something new, it's called the MSE_{Effect} .

$$MSE_{\text{Effect}} = \frac{SS_{\text{Effect}}}{df_{\text{Effect}}}$$

$$MSE_{\text{Effect}} = \frac{72}{2} = 36$$

This might look alien and seem a bit complicated. But, it's just another mean. It's the mean of the sums of squares for the effect. If this reminds you of the formula for the variance, good memory. The MSE_{Effect} is a measure variance for the change in the data due to changes in the means (which are tied to the experimental conditions).

The SS_{Error} represents the sum of variation for nine scores in our study. That's a lot more scores, so the SS_{Error} is often way bigger than than SS_{Effect} . If we left our $SSes$ this way and divided them, we would almost always get numbers less than one, because the SS_{Error} is so big. What we need to do is bring it down to the average size. So, we might want to divide our SS_{Error} by 9, after all there were nine scores. However, because we are estimating this property, we divide by the degrees of freedom instead ($\text{scores-groups} = 9 - 3 = 6$). Now we have created something new, it's called the MSE_{Error} .

$$MSE_{\text{Error}} = \frac{SS_{\text{Error}}}{df_{\text{Error}}}$$

$$MSE_{\text{Error}} = \frac{230}{6} = 38.33$$

7.2.7 Calculate F

Now that we have done all of the hard work, calculating F is easy:

$$F = \frac{\text{measure of effect}}{\text{measure of error}}$$

$$F = \frac{MSE_{\text{Effect}}}{MSE_{\text{Error}}}$$

$$F = \frac{36}{38.33} = .939$$

Done!

7.2.8 The ANOVA TABLE

You might suspect we aren't totally done here. We've walked through the steps of computing F . Remember, F is a sample statistic, we computed F directly from the data. There were a whole bunch of pieces we needed, the dfs, the SSes, the MSEs, and then finally the F .

All of these little pieces are conveniently organized by ANOVA tables. ANOVA tables look like this:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
groups	2	72	36.00000	0.9391304	0.4417359
Residuals	6	230	38.33333	NA	NA

You are looking at the print-out of an ANOVA summary table from R. Notice, it had columns for Df , SS (Sum Sq), MSE (Mean Sq), F , and a p -value. There are two rows. The **groups** row is for the Effect (what our means can explain). The **Residuals** row is for the Error (what our means can't explain). Different programs give slightly different labels, but they are all attempting to present the same information in the ANOVA table. There isn't anything special about the ANOVA table, it's just a way of organizing all the pieces. Notice, the MSE for the effect (36) is placed above the MSE for the error (38.333), and this seems natural because we divide 36/38.33 in or to get the F -value!

7.3 What does F mean?

We've just noted that the ANOVA has a bunch of numbers that we calculated straight from the data. All except one, the p -value. We did not calculate the p -value from the data. Where did it come from, what does it mean? How do we use this for statistical inference. Just so you don't get too worried, the p -value for the ANOVA has the very same general meaning as the p -value for the t -test, or the p -value for any sample statistic. It tells us that the probability that we would observe our test statistic or larger, under the distribution of no differences (the null).

As we keep saying, F is a sample statistic. Can you guess what we do with sample statistics in this textbook? We did it for the Crump Test, the Randomization Test, and the t -test... We make fake data, we simulate it, we compute the sample statistic we are interested in, then we see how it behaves over many replications or simulations.

Let's do that for F . This will help you understand what F really is, and how it behaves. We are going to create the sampling distribution of F . Once we have that you will be able to see where the p -values come from. It's the same basic process that we followed for the t tests, except we are measuring F instead of t .

Here is the set-up, we are going to run an experiment with three levels. In our imaginary experiment we are going to test whether a new magic pill can make you smarter. The independent variable is the number of magic pills you take: 1, 2, or 3. We will measure your smartness using a smartness test. We will assume the smartness test has some known properties, the mean score on the test is 100, with a standard deviation of 10 (and the distribution is normal).

The only catch is that our magic pill does NOTHING AT ALL. The fake people in our fake experiment will all take sugar pills that do absolutely nothing to their smartness. Why would we want to simulate such a bunch of nonsense? The answer is that this kind of simulation is critical for making inferences about chance if you were to conduct a real experiment.

Here are some more details for the experiment. Each group will have 10 different subjects, so there will be a total of 30 subjects. We are going to run this experiment 10,000 times. Each time drawing numbers randomly from the very same normal distribution. We are going to calculate F from our sample data every

time, and then we are going to draw the histogram of F -values. This will show us the sampling distribution of F for our situation. Let's do that and see what it looks like:

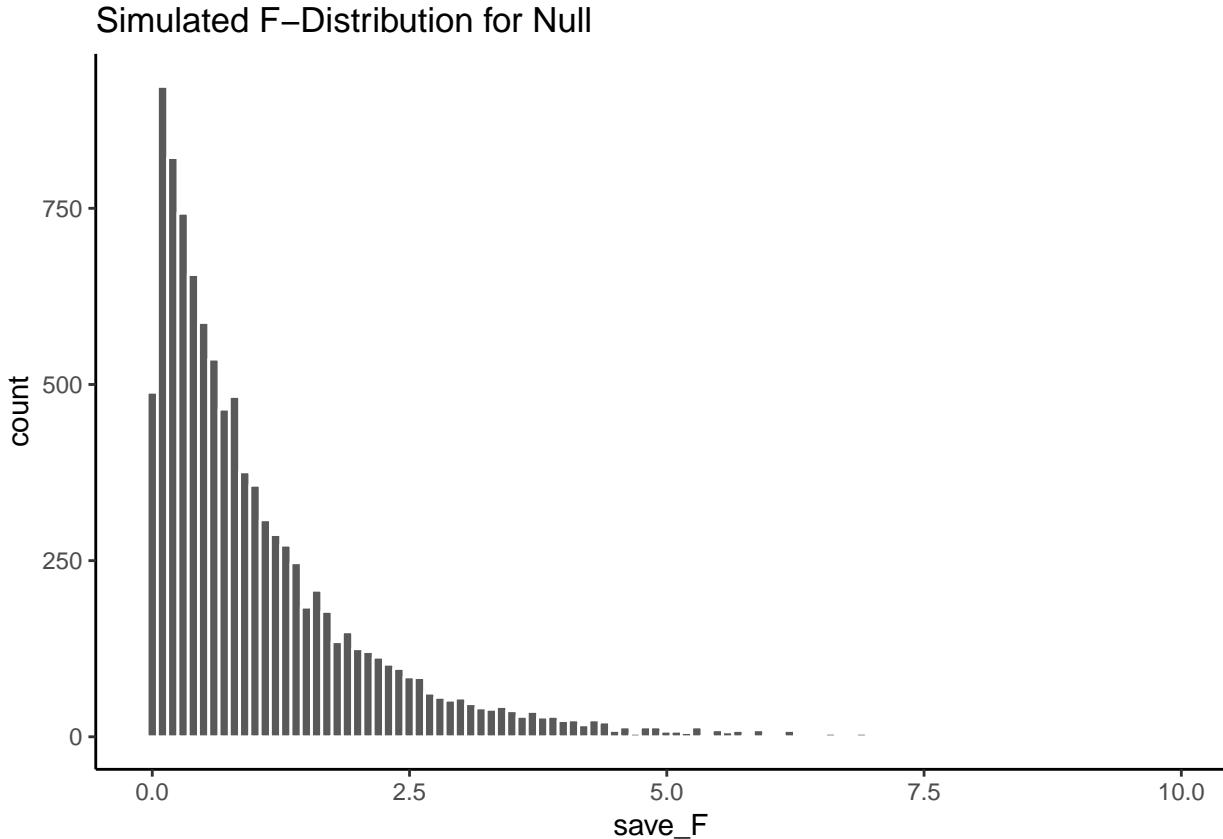


Figure 7.1: A simulation of 10,000 experiments from a null distribution where there is no differences. The histogram shows 10,000 F -values, one for each simulation. These are values that F can take in this situation. All of these F -values were produced by random sampling error

Let's note a couple things about the F distribution. 1) The smallest value is 0, and there are no negative values. Does this make sense? F can never be negative because it is the ratio of two variances, and variances are always positive because of the squaring operation. So, yes, it makes sense that the sampling distribution of F is always 0 or greater. 2) it does not look normal. No it does not. F can have many different looking shapes, depending on the degrees of freedom in the numerator and denominator. However, these aspects are too important for now.

Remember, before we talked about some intuitive ideas for understanding F , based on the idea that F is a ratio of what we can explain (variance due to mean differences), divided by what we can't explain (the error variance). When the error variance is higher than the effect variance, then we will always get an F -value less than one. You can see that we often got F -values less than one in the simulation. This is sensible, after all we were simulating samples coming from the very same distribution. On average there should be no differences between the means. So, on average the part of the total variance that is explained by the means should be less than one, or around one, because it should be roughly the same as the amount of error variance (remember, we are simulating no differences).

At the same time, we do see that some F -values are larger than 1. There are little bars that we can see going all the way up to about 5. If you were to get an F -value of 5, you might automatically think, that's a pretty big F -value. Indeed it kind of is, it means that you can explain 5 times more of variance than you can't explain. That seems like a lot. You can also see that larger F -values don't occur very often. As a final reminder, what you are looking at is how the F -statistic (measured from each of 10,000 simulated

experiments) behaves when the only thing that can cause differences in the means is random sampling error. Just by chance sometimes the means will be different. You are looking at another chance window. These are the F s that chance can produce.

7.3.1 Making Decisions

We can use the sampling distribution of F (for the null) to make decisions about the role of chance in a real experiment. For example, we could do the following.

1. Set an alpha criterion of $p = 0.05$
2. Find out the critical value for F , for our particular situation (with our dfs for the numerator and denominator).

Let's do that. I've drawn the line for the critical value onto the histogram:

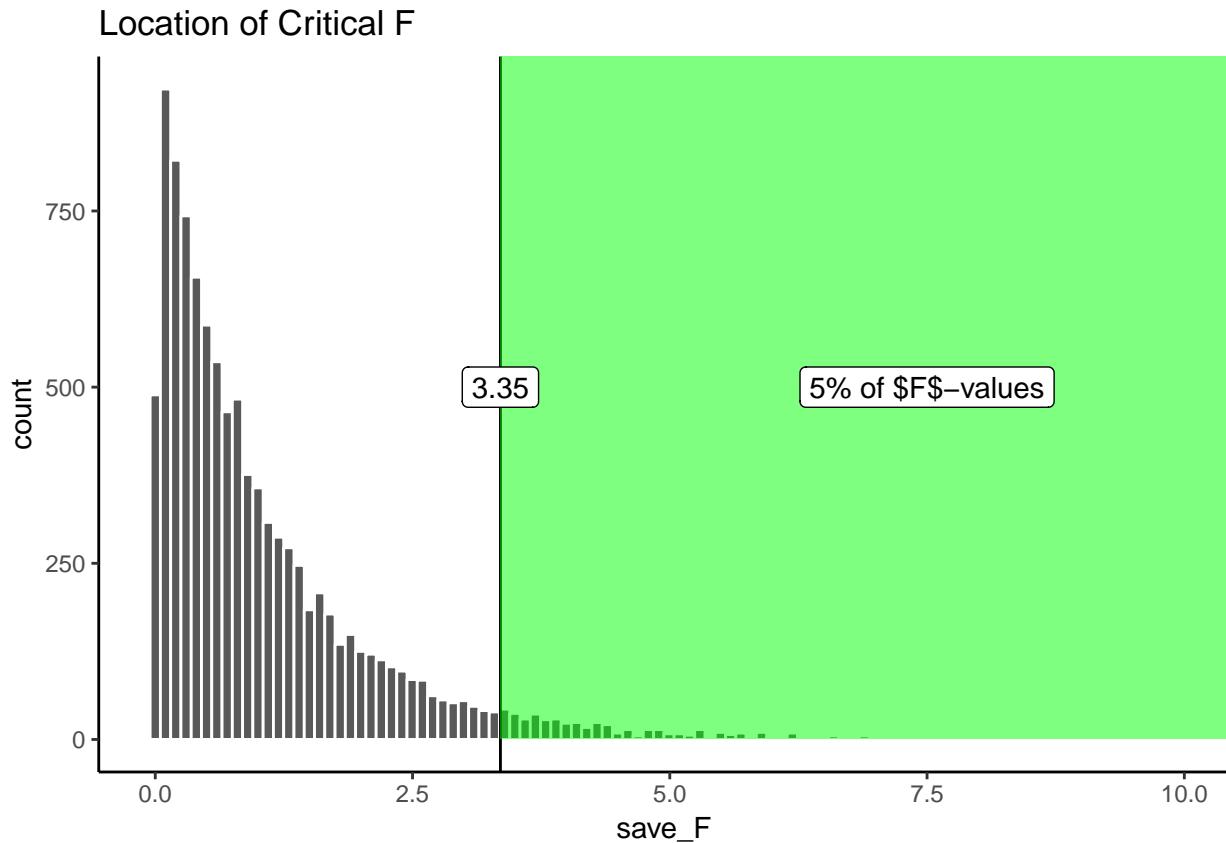


Figure 7.2: The critical value for F where 5% of all F -values lie beyond this point

Alright, now we can see that only 5% of all F -values from this sampling distribution will be 3.35 or larger. We can use this information.

How would we use it? Imagine we ran a real version of this experiment. And, we really used some pills that just might change smartness. If we ran the exact same design, with 30 people in total (10 in each group), we could set an F criterion of 3.35 for determining whether any of our results reflected a causal change in smartness due to the pills, and not due to random chance. For example, if we found an F -value of 3.34, which happens, just less than 5% of the time, we might conclude that random sampling error did not produce the differences between our means. Instead, we might be more confident that the pills actually did something, after all an F -value of 3.34 doesn't happen very often, it is unlikely (only 5 times out of 100) to occur by chance.

7.3.2 Fs and means

Up to here we have been building your intuition for understanding F . We went through the calculation of F from sample data. We went through the process of simulating thousands of F s to show you the null distribution. We have not talked so much about what researchers really care about...The MEANS! The actual results from the experiment. Were the means different? that's often what people want to know. So, now we will talk about the means, and F , together.

Notice, if I told you I ran an experiment with three groups, testing whether some manipulation changes the behavior of the groups, and I told you that I found a big $F!$, say an F of 6!. And, that the F of 6 had a p -value of .001. What would you know based on that information alone? You would only know that Fs of 6 don't happen very often by chance. In fact they only happen 0.1% of the time, that's hardly at all. If someone told me those values, I would believe that the results they found in their experiment were not likely due to chance. However, I still would not know what the results of the experiment were! Nobody told us what the means were in the different groups, we don't know what happened!

IMPORTANT: even though we don't know what the means were, we do know something about them, whenever we get F -values and p -values like that (big F s, and very small associated ps)... Can you guess what we know? I'll tell you. We automatically know that there **must have been some differences between the means**. If there was no differences between the means, then the variance explained by the means (the numerator for F) would not be very large. So, we know that there must be some differences, we just don't know what they are. Of course, if we had the data, all we would need to do is look at the means for the groups (the ANOVA table doesn't report this, we need to do it as a separate step).

7.3.2.1 ANOVA is an omnibus test

This property of the ANOVA is why the ANOVA is sometimes called the **omnibus test**. Omnibus is a fun word, it sounds like a bus I'd like to ride. The meaning of omnibus, according to the dictionary, is "comprising several items". The ANOVA is, in a way, one omnibus test, comprising several little tests.

For example, if you had three groups, A, B, and C. You get could differences between

1. A and B
2. B and C
3. A and C

That's three possible differences you could get. You could run separate t -tests, to test whether each of those differences you might have found could have been produced by chance. Or, you could run an ANOVA, like what we have been doing, to ask one more general question about the differences. Here is one way to think about what the omnibus test is testing:

Hypothesis of no differences anywhere: $A = B = C$

Any differences anywhere:

- a. $A \neq B = C$
- b. $A = B \neq C$
- c. $A \neq C = B$

The \neq symbol means "does not equal", it's an equal sign with a cross through it (no equals allowed!).

How do we put all of this together. Generally, when we get a small F -value, with a large p -value, we will not reject the hypothesis of no differences. We will say that we do not have evidence that the means of the three groups are in any way different, and the differences that are there could easily have been produced by chance. When we get a large F with a small p -value (one that is below our alpha criterion), we will generally reject the hypothesis of no differences. We would then assume that at least one group mean is not equal to one of the others. That is the omnibus test. Rejecting the null in this way is rejecting the idea there are no

differences. But, the F test still does not tell you which of the possible group differences are the ones that are different.

7.3.2.2 Looking at a bunch of group means

We ran 10,000 experiments just before, and we didn't even once look at the group means for any of the experiments. Let's quickly do that, so we get a better sense of what is going on.

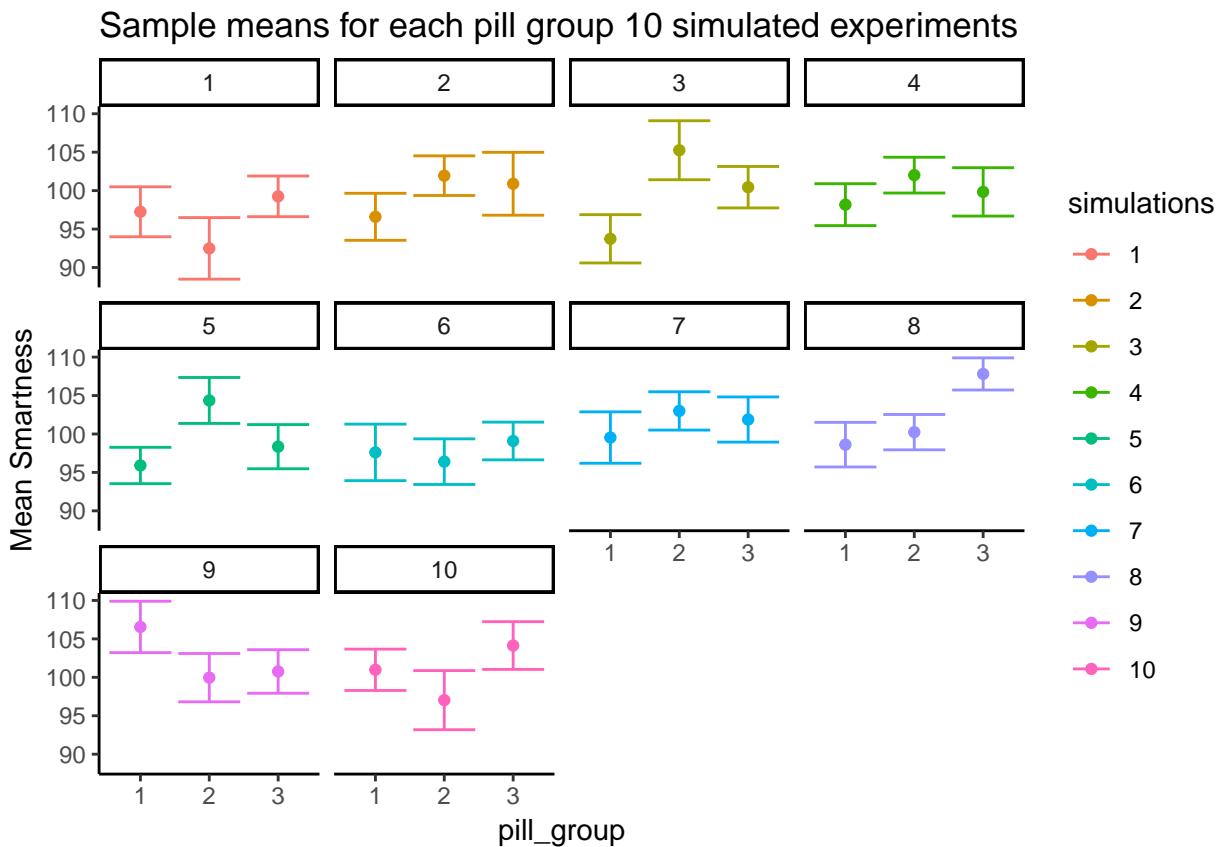


Figure 7.3: Different patterns of group means under the null (all scores for each group sampled from the same distribution)

Whoa, that's a lot to look at. What is going on here? Each little box represents the outcome of a simulated experiment. The dots are the means for each group (whether subjects took 1, 2, or 3 magic pills). The y-axis shows the mean smartness for each group. The error bars are standard errors of the mean.

You can see that each of the 10 experiments turn out different. Remember, we sampled 10 numbers for each group from the **same** normal distribution with $\text{mean} = 100$, and $\text{sd} = 10$. So, we know that the **correct** means for each sample should actually be 100 every single time. However, they are not 100 every single time because of?...**sampling error** (Our good friend that we talk about all the time).

For most of the simulations the error bars are all overlapping, this suggests visually that the means are not different. However, some of them look like they are not overlapping so much, and this would suggest that they are different. This is the siren song of chance (sirens lured sailors to their deaths at sea...beware of the siren call of chance). If we concluded that any of these sets of means had a true difference, we would be committing a type I error. Because we made the simulation, we know that none of these means are actually different. But, when you are running a real experiment, you don't get to know this for sure.

7.3.2.3 Looking at bar graphs

Let's look at the exact same graph as above, but this time use bars to visually illustrate the means, instead of dots. We'll re-do our simulation of 10 experiments, so the pattern will be a little bit different:

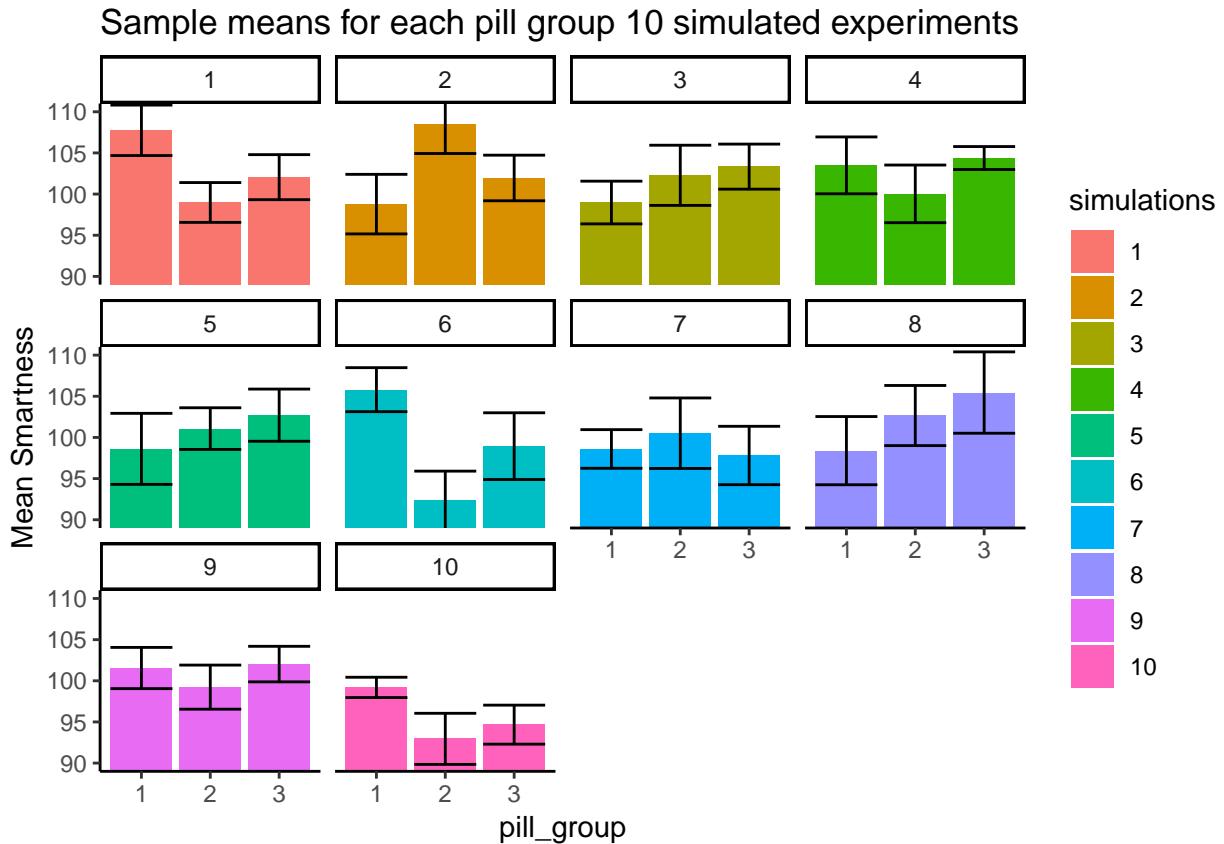


Figure 7.4: Different patterns of group means under the null (all scores for each group sampled from the same distribution)

Now the heights of the bars display the means for each pill group. In general we see the same thing. Some of the fake experiments look like there might be differences, and some of them don't.

7.3.2.4 What mean differences look like when F is < 1

We are now giving you some visual experience looking at what means look like from a particular experiment. This is for your stats intuition. We're trying to improve your data senses.

What we are going to do now is similar to what we did before. Except this time we are going to look at 10 simulated experiments, where all of the F -values were less than 1. All of these F -values would also be associated with fairly large p -values. When F is less than one, we would not reject the hypothesis of no differences. So, when we look at patterns of means when F is less than 1, we should see mostly the same means, and no big differences.

The numbers in the panels now tell us which simulations actually produced Fs of less than 1.

We see here that all the bars aren't perfectly flat, that's OK. What's more important is that for each panel, the error bars for each mean are totally overlapping with all the other error bars. We can see visually that our estimate of the mean for each sample is about the same for all of the bars. That's good, we wouldn't make any type I errors here.

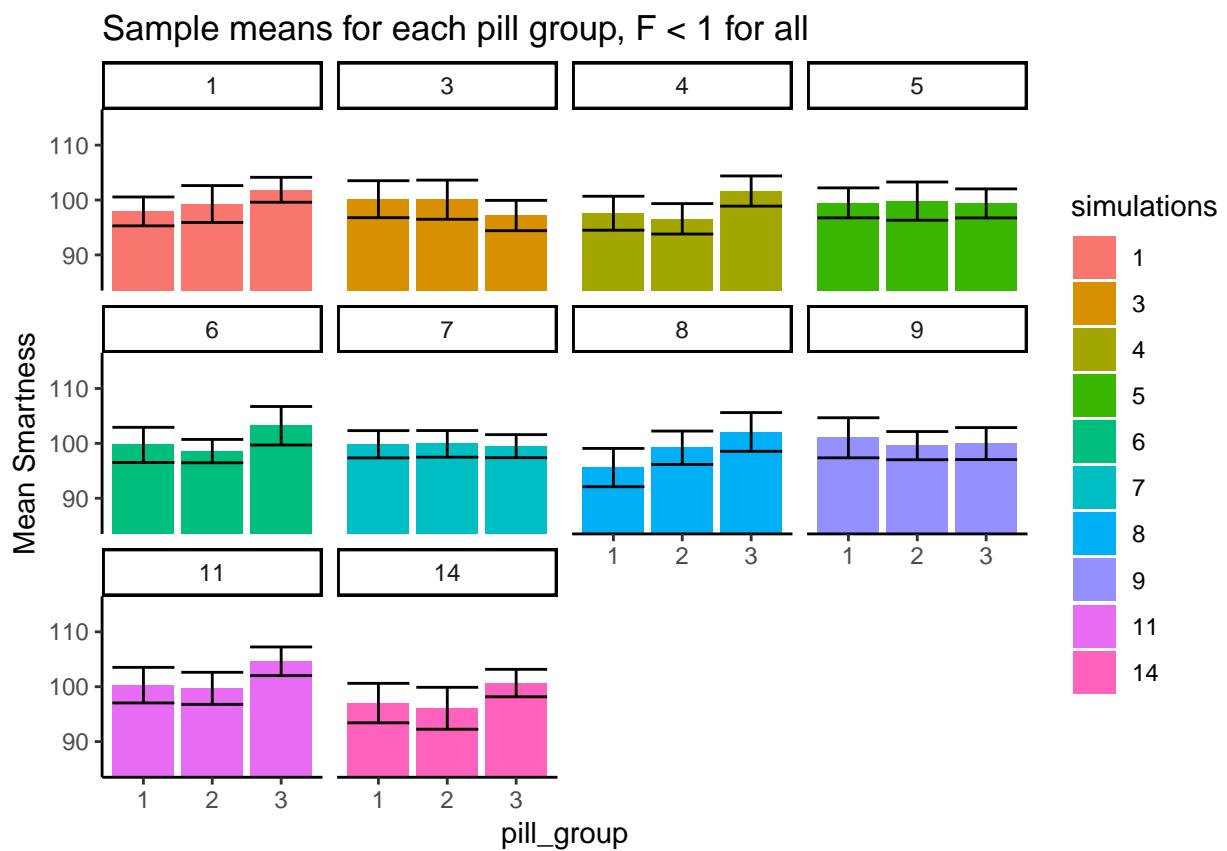


Figure 7.5: Different patterns of group means under the null (sampled from same distribution) when F is less than 1

7.3.2.5 What mean differences look like when $F > 3.35$

Earlier we found that the critical value for F in our situation was 3.35, this was the location on the F distribution where only 5% of F 's were 3.35 or greater. We would reject the hypothesis of no differences whenever F was greater than 3.35. In this case, whenever we did that, we would be making a type I error. That is because we are simulating the distribution of no differences (remember all of our sample means are coming from the exact same distribution). So, now we can take a look at what type I errors look like. In other words, we can run some simulations and look at the pattern in the means, only when F happens to be 3.35 or greater (this only happens 5% of the time, so we might have to let the computer simulate for a while). Let's see what that looks like:

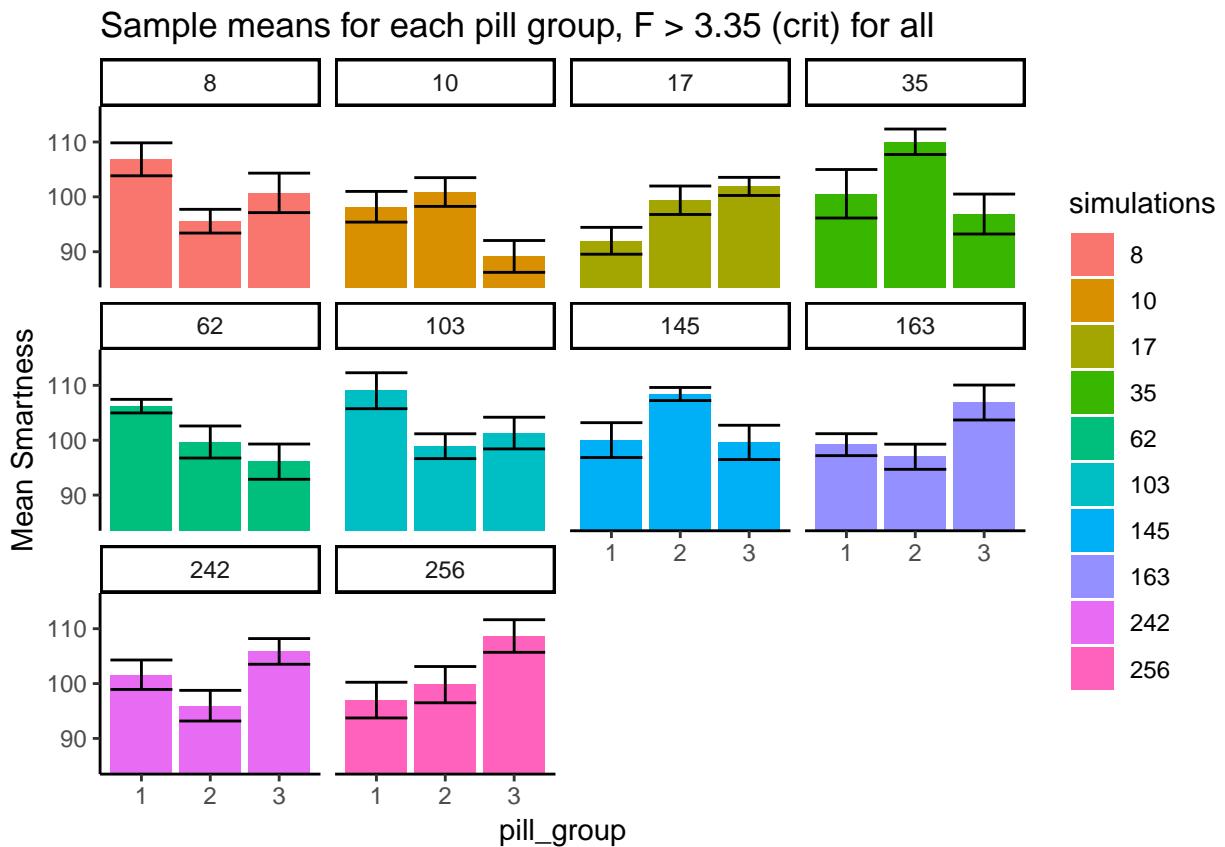


Figure 7.6: Different patterns of group means under the null when F is above critical value (these are all type I Errors)

The numbers in the panels now tell us which simulations actually produced F s that were greater than 3.35

What do you notice about the pattern of means inside each panel? Now, every single panel shows at least one mean that is different from the others. Specifically, the error bars for one mean do not overlap with the error bars for one or another mean. This is what mistakes looks like. These are all type I errors. They are insidious. When they happen to you by chance, the data really does appear to show a strong pattern, and your F -value is large, and your p -value is small! It is easy to be convinced by a type I error (it's the siren song of chance).

7.4 ANOVA on Real Data

We've covered many fundamentals about the ANOVA, how to calculate the necessary values to obtain an F -statistic, and how to interpret the F -statistic along with its associate p -value once we have one. In general, you will be conducting ANOVAs and playing with F s and p s using software that will automatically spit out the numbers for you. It's important that you understand what the numbers mean, that's why we've spent time on the concepts. We also recommend that you try to compute an ANOVA by hand at least once. It builds character, and let's you know that you know what you are doing with the numbers.

But, we've probably also lost the real thread of all this. The core thread is that when we run an experiment we use our inferential statistics, like ANOVA, to help us determine whether the differences we found are likely due to chance or not. In general, we like to find out that the differences that we find are not due to chance, but instead due to our manipulation.

So, we return to the application of the ANOVA to a real data set with a real question. This is the same one that you will be learning about in the lab. We give you a brief overview here so you know what to expect.

7.4.1 Tetris and bad memories

Yup, you read that right. The research you will learn about tests whether playing Tetris after watching a scary movie can help prevent you from having bad memories from the movie (James et al., 2015). Sometimes in life people have intrusive memories, and they think about things they'd rather not have to think about. This research looks at one method that could reduce the frequency of intrusive memories.

Here's what they did. Subjects watched a scary movie, then at the end of the week they reported how many intrusive memories about the movie they had. The mean number of intrusive memories was the measurement (the dependent variable). This was a between-subjects experiment with four groups. Each group of subjects received a different treatment following the scary movie. The question was whether any of these treatments would reduce the number of intrusive memories. All of these treatments occurred after watching the scary movie:

1. No-task control: These participants completed a 10-minute music filler task after watching the scary movie.
2. Reactivation + Tetris: These participants were shown a series of images from the trauma film to reactivate the traumatic memories (i.e., reactivation task). Then, participants played the video game Tetris for 12 minutes.
3. Tetris Only: These participants played Tetris for 12 minutes, but did not complete the reactivation task.
4. Reactivation Only: These participants completed the reactivation task, but did not play Tetris.

For reasons we elaborate on in the lab, the researchers hypothesized that the **Reactivation+Tetris** group would have fewer intrusive memories over the week than the other groups.

Let's look at the findings. Note you will learn how to do all of these steps in the lab. For now, we just show the findings and the ANOVA table. Then we walk through how to interpret it.

Ooooh, look at that. We did something fancy. You are looking at the data from the four groups. The height of each bar shows the mean intrusive memories for the week. The dots show the individual scores for each subject in each group (useful to see the spread of the data). The error bars show the standard errors of the mean.

What can we see here? Right away it looks like there is some support for the research hypothesis. The green bar, for the Reactivation + Tetris group had the lowest mean number of intrusive memories. Also, the error bar is not overlapping with any of the other error bars. This implies that the mean for the Reactivation + Tetris group is different from the means for the other groups. And, this difference is probably not very likely by chance.

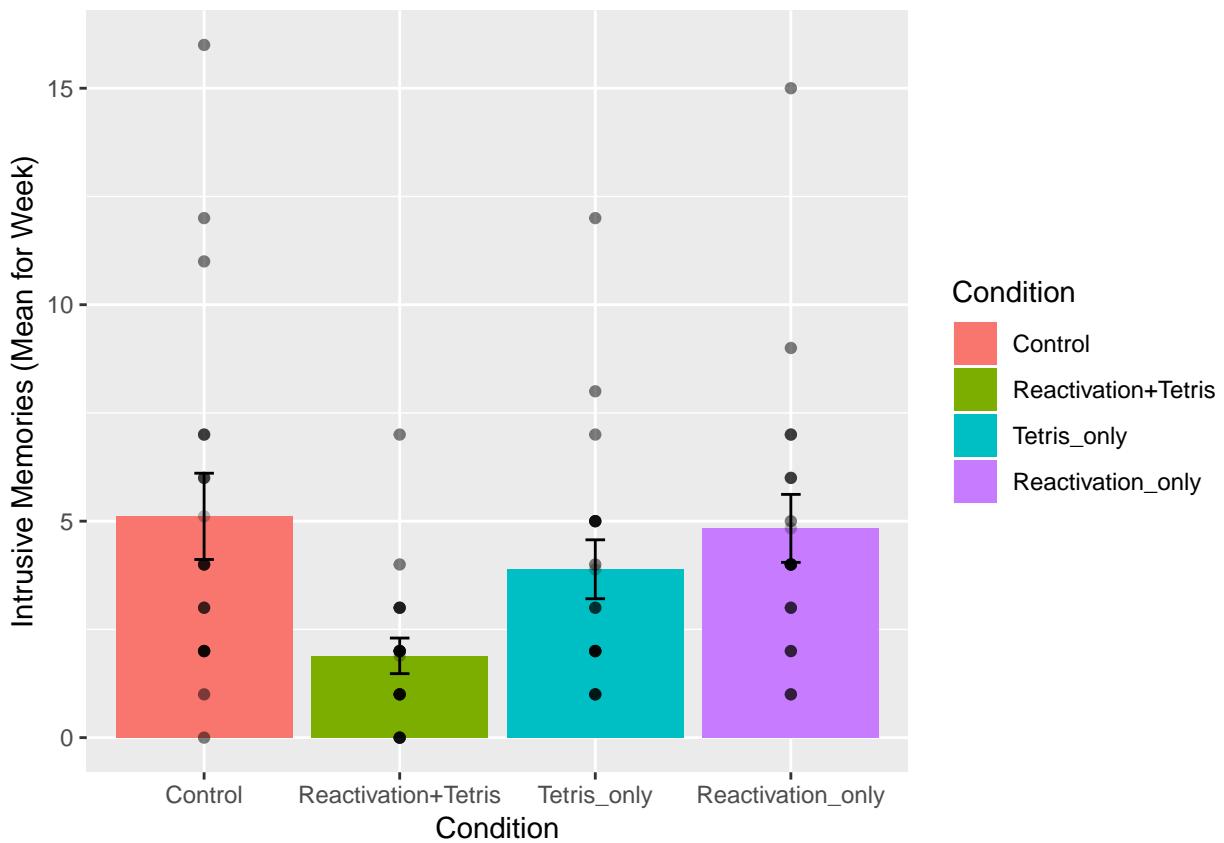


Figure 7.7: Mean number of intrusive memories per week as a function of experimental treatments

We can now conduct the ANOVA on the data to ask the omnibus question. If we get a an F -value with an associated p -value of less than .05 (the alpha criterion set by the authors), then we can reject the hypothesis of no differences. Let's see what happens:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Condition	3	114.8194	38.27315	3.794762	0.0140858
Residuals	68	685.8333	10.08578	NA	NA

We see the ANOVA table, it's up there. We could report the results from the ANOVA table like this:

There was a significant main effect of treatment condition, $F(3, 68) = 3.79$, $MSE = 10.08$, $p=0.014$.

We called this a significant effect because the p -value was less than 0.05. In other words, the F -value of 3.79 only happens 1.4% of the time when the null is true. Or, the differences we observed in the means only occur by random chance (sampling error) 1.4% of the time. Because chance rarely produces this kind of result, the researchers made the inference that chance DID NOT produce their differences, instead, they were inclined to conclude that the Reactivation + Tetris treatment really did cause a reduction in intrusive memories. That's pretty neat.

7.4.2 Comparing means after the ANOVA

Remember that the ANOVA is an omnibus test, it just tells us whether we can reject the idea that all of the means are the same. The F-test (synonym for ANOVA) that we just conducted suggested we could reject the hypothesis of no differences. As we discussed before, that must mean that there are some differences in the pattern of means.

Generally after conducting an ANOVA, researchers will conduct follow-up tests to compare differences between specific means. We will talk more about this practice throughout the textbook. There are many recommended practices for follow-up tests, and there is a lot of debate about what you should do. We are not going to wade into this debate right now. Instead we are going to point out that **you need to do something** to compare the means of interest after you conduct the ANOVA, because the ANOVA is just the beginning...It usually doesn't tell you want you want to know. You might wonder why bother conducting the ANOVA in the first place...Not a terrible question at all. A good question. You will see as we talk about more complicated designs, why ANOVAs are so useful. In the present example, they are just a common first step. There are required next steps, such as what we do next.

How can you compare the difference between two means, from a between-subjects design, to determine whether or not the difference you observed is likely or unlikely to be produced by chance? We covered this one already, it's the independent t -test. We'll do a couple t -tests, showing the process.

7.4.2.1 Control vs. Reactivation+Tetris

What we really want to know is if Reactivation+Tetris caused fewer intrusive memories...but compared to what? Well, if it did something, the Reactivation+Tetris group should have a smaller mean than the Control group. So, let's do that comparison:

```
## 
## Two Sample t-test
## 
## data: Days_One_to_Seven_Number_of_Intrusions by Condition
## t = 2.9893, df = 34, p-value = 0.005167
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.031592 5.412852
## sample estimates:
```

```
##          mean in group Control mean in group Reactivation+Tetris
##                                5.111111                           1.888889
```

We found that there was a significant difference between the control group ($M=5.11$) and Reactivation + Tetris group ($M=1.89$), $t(34) = 2.99$, $p=0.005$.

Above you just saw an example of reporting another t -test. This sentence does an OK job of telling the reader everything they want to know. It has the means for each group, and the important bits from the t -test.

More important, as we suspected the difference between the control and Reactivation + Tetris group was likely not due to chance.

7.4.2.2 Control vs. Tetris_only

Now we can really start wondering what caused the difference. Was it just playing Tetris? Does just playing Tetris reduce the number of intrusive memories during the week? Let's compare that to control:

```
##          mean in group Control mean in group Tetris_only
##                                5.111111                           3.888889
```

Here we did not find a significant difference. We found that no significant difference between the control group ($M=5.11$) and Tetris Only group ($M=3.89$), $t(34) = 2.99$, $p=0.318$.

So, it seems that not all of the differences between our means are large enough to be called statistically significant. In particular, the difference here, or larger, happens by chance 31.8% of the time.

You could go on doing more comparisons, between all of the different pairs of means. Each time conducting a t -test, and each time saying something more specific about the patterns across the means than you get to say with the omnibus test provided by the ANOVA.

Usually, it is the pattern of differences across the means that you as a researcher are primarily interested in understanding. Your theories will make predictions about how the pattern turns out (e.g., which specific means should be higher or lower and by how much). So, the practice of doing comparisons after an ANOVA is really important for establishing the patterns in the means.

7.5 ANOVA Summary

We have just finished a rather long introduction to the ANOVA, and the F -test. The next couple of chapters continue to explore properties of the ANOVA for different kinds of experimental designs. In general, the process to follow for all of the more complicated designs is very similar to what we did here, which boils down to two steps:

- 1) conduct the ANOVA on the data
- 2) conduct follow-up tests, looking at differences between particular means

So what's next...the ANOVA for repeated measures designs. See you in the next chapter.

Chapter 8

Repeated Measures ANOVA

This chapter introduces you to **repeated measures ANOVA**. Repeated measures ANOVAs are very common in Psychology, because psychologists often use repeated measures designs, and repeated measures ANOVAs are the appropriate test for making inferences about repeated measures designs.

Remember the paired sample t -test? We used that test to compare two means from a repeated measures design. Remember what a repeated measures design is? It's also called a within-subjects design. These designs involve measuring the same subject more than once. Specifically, at least once for every experimental condition. In the paired t -test example, we discussed a simple experiment with only two experimental conditions. There, each subject would contribute a measurement to level one and level two of the design.

However, paired-samples t -tests are limited to comparing two means. What if you had a design that had more than two experimental conditions? For example, perhaps your experiment had 3 levels for the independent variable, and each subject contributed data to each of the three levels?

This is starting to sounds like an ANOVA problem. ANOVAs are capable of evaluating whether there is a difference between any number of means, two or greater. So, we can use an ANOVA for our repeated measures design with three levels for the independent variable.

Great! So, what makes a repeated measures ANOVA different from the ANOVA we just talked about?

8.1 Repeated measures design

Let's use the exact same toy example from the previous chapter, but let's convert it to a repeated measures design.

Last time, we imagined we had some data in three groups, A, B, and C. The data looked like this:

groups	scores
A	20
A	11
A	2
B	6
B	2
B	7
C	2
C	11
C	2

The above table represents a between-subject design where each score involves a unique subject.

Let's change things up a tiny bit, and imagine we only had 3 subjects in total in the experiment. And, that each subject contributed data to the three levels of the independent variable, A, B, and C. Before we called the **IV groups**, because there were different groups of subjects. Let's change that to **conditions**, because now the same group of subjects participates in all three conditions. Here's the new table for a within-subjects (repeated measures) version of this experiment:

subjects	conditions	scores
1	A	20
2	A	11
3	A	2
1	B	6
2	B	2
3	B	7
1	C	2
2	C	11
3	C	2

8.2 Partitioning the Sums of Squares

Time to introduce a new name for an idea you learned about last chapter, it's called **partitioning the sums of squares**. Sometimes an obscure new name can be helpful for your understanding of what is going on. ANOVAs are all about partitioning the sums of squares. We already did some partitioning in the last chapter. What do we mean by partitioning?

Imagine you had a big empty house with no rooms in it. What would happen if you partitioned the house? What would you be doing? One way to partition the house is to split it up into different rooms. You can do this by adding new walls and making little rooms everywhere. That's what partitioning means, to split up.

The act of partitioning, or splitting up, is the core idea of ANOVA. To use the house analogy. Our total sums of squares (SS Total) is our big empty house. We want to split it up into little rooms. Before we partitioned SS Total using this formula:

$$SS_{\text{TOTAL}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

Remember, the SS_{Effect} was the variance we could attribute to the means of the different groups, and SS_{Error} was the leftover variance that we couldn't explain. SS_{Effect} and SS_{Error} are the partitions of SS_{TOTAL} , they are the little rooms.

In the between-subjects case above, we got to split SS_{TOTAL} into two parts. What is most interesting about the repeated-measures design, is that we get to split SS_{TOTAL} into three parts, there's one more partition. Can you guess what the new partition is? Hint: whenever we have a new way to calculate means in our design, we can always create a partition for those new means. What are the new means in the repeated measures design?

Here is the new idea for partitioning SS_{TOTAL} in a repeated-measures design:

$$SS_{\text{TOTAL}} = SS_{\text{Effect}} + SS_{\text{Subjects}} + SS_{\text{Error}}$$

We've added SS_{Subjects} as the new idea in the formula. What's the idea here? Well, because each subject was measured in each condition, we have a new set of means. These are the means for each subject, collapsed across the conditions. For example, subject 1 has a mean (mean of their scores in conditions A, B, and C); subject 2 has a mean (mean of their scores in conditions A, B, and C); and subject 3 has a mean (mean of their scores in conditions A, B, and C). There are three subject means, one for each subject, collapsed across the conditions. And, we can now estimate the portion of the total variance that is explained by these subject means.

We just showed you a “formula” to split up SS_{TOTAL} into three parts, but we called the formula an idea. We did that because the way we wrote the formula is a little bit misleading, and we need to clear something up. Before we clear the thing up, we will confuse you just a little bit. Be prepared to be confused a little bit.

First, we need to introduce you to some more terms. It turns out that different authors use different words to describe parts of the ANOVA. This can be really confusing. For example, we described the SS formula for a between subjects design like this:

$$SS_{TOTAL} = SS_{Effect} + SS_{Error}$$

However, the very same formula is often written differently, using the words between and within in place of effect and error, it looks like this:

$$SS_{TOTAL} = SS_{Between} + SS_{Within}$$

Whoa, hold on a minute. Haven’t we switched back to talking about a **between-subjects** ANOVA. YES! Then why are we using the word **within**, what does that mean? YES! We think this is very confusing for people. Here the word **within** has a special meaning. It **does not** refer to a within-subjects design. Let’s explain. First, $SS_{Between}$ (which we have been calling SS_{Effect}) refers to variation **between** the group means, that’s why it is called $SS_{Between}$. Second, and most important, SS_{Within} (which we have been calling SS_{Error}), refers to the leftover variation within each group mean. Specifically, it is the variation between each group mean and each score in the group. “AAGGH, you’ve just used the word between to describe within group variation!”. Yes! We feel your pain. Remember, for each group mean, every score is probably off a little bit from the mean. So, the scores within each group have some variation. This is the within group variation, and it is why the leftover error that we can’t explain is often called SS_{Within} .

OK. So why did we introduce this new confusing way of talking about things? Why can’t we just use SS_{Error} to talk about this instead of SS_{Within} , which you might (we do) find confusing. We’re getting there, but perhaps a picture will help to clear things up.

The figure lines up the partitioning of the Sums of Squares for both between-subjects and repeated-measures designs. In both designs, SS_{Total} is first split up into two pieces SS_{Effect} (between-groups) and SS_{Error} (within-groups). At this point, both ANOVAs are the same. In the repeated measures case we split the SS_{Error} (within-groups) into two more littler parts, which we call $SS_{Subjects}$ (error variation about the subject mean) and SS_{Error} (left-over variation we can’t explain).

So, when we earlier wrote the formula to split up SS in the repeated-measures design, we were kind of careless in defining what we actually meant by SS_{Error} , this was a little too vague:

$$SS_{TOTAL} = SS_{Effect} + SS_{Subjects} + SS_{Error}$$

The critical feature of the repeated-measures ANOVA, is that the SS_{Error} that we will later use to compute the MSE in the denominator for the F -value, is smaller in a repeated-measures design, compared to a between subjects design. This is because the SS_{Error} (within-groups) is split into two parts, $SS_{Subjects}$ (error variation about the subject mean) and SS_{Error} (left-over variation we can’t explain).

To make this more clear, we made another figure:

As we point out, the SS_{Error} (left-over) in the green circle will be a smaller number than the SS_{Error} (within-group). That’s because we are able to subtract out the $SS_{Subjects}$ part of the SS_{Error} (within-group). As we will see shortly, this can have the effect of producing larger F-values when using a repeated-measures design compared to a between-subjects design.

8.3 Calculating the RM ANOVA

Now that you are familiar with the concept of an ANOVA table (remember the table from last chapter where we reported all of the parts to calculate the F -value?), we can take a look at the things we need to find out

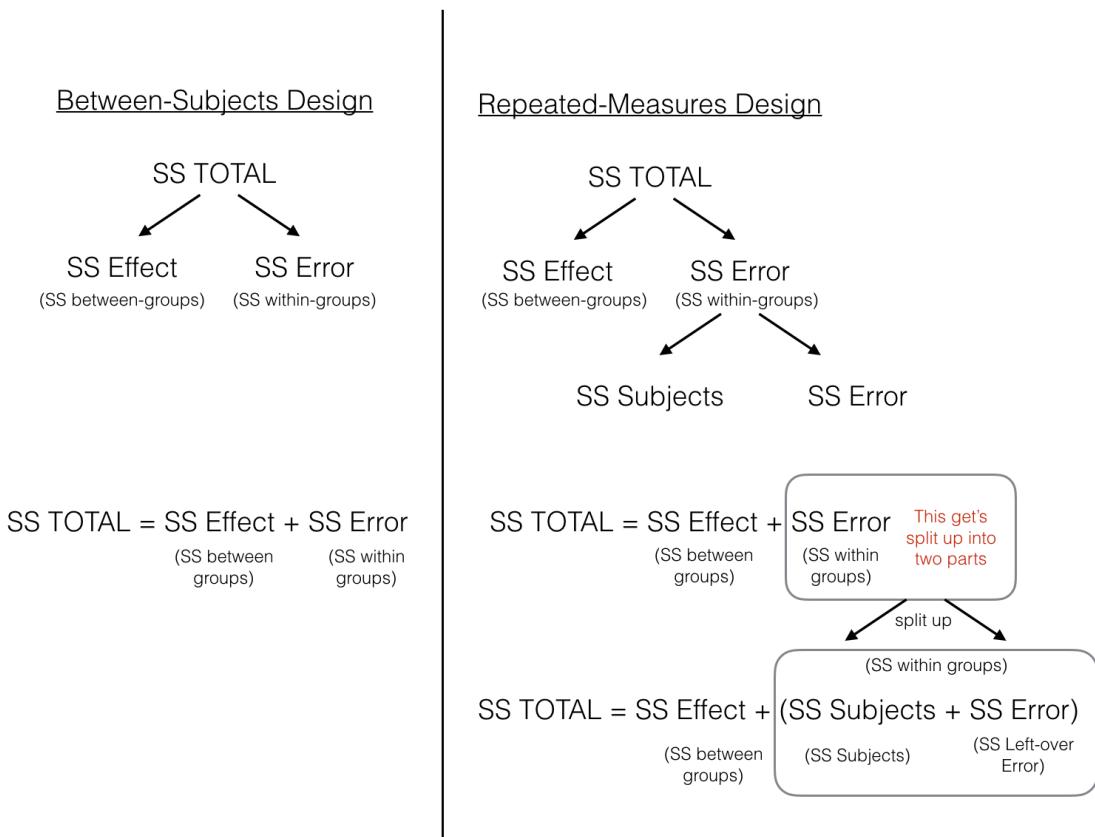


Figure 8.1: Illustration showing how the total sums of squares are partitioned differently for a between versus repeated-measures design

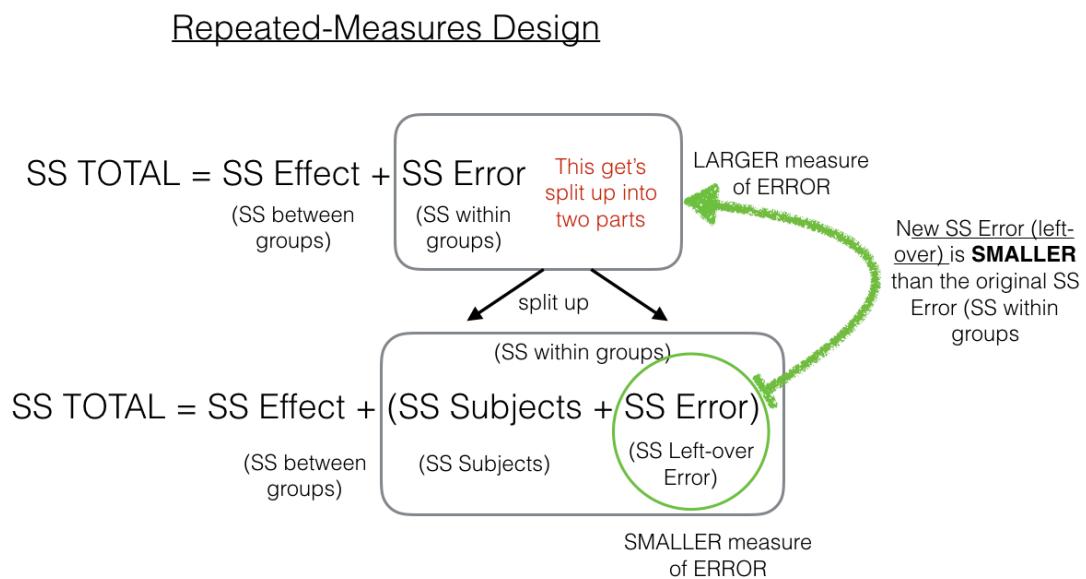


Figure 8.2: Close-up showing that the Error term is split into two parts in the repeated measures design

	df	SS	MSE	F	P
EFFECT	df1 effect = conditions-1	SS Effect= SS Total - SS Error (within-conditions)	MSE Effect = $\frac{SS \text{ Effect}}{df1 \text{ Effect}}$	$F = \frac{MSE \text{ Effect}}{MSE \text{ Error}}$	p = From Sampling distribution of $F(df1, df2)$
ERROR	df2 Effect = $(n-1) \times (conditions-1)$ n=# of subjects	SS Error (left-over)= SS Error (within-conditions) - SS Subjects	MSE Error (left-over) = $\frac{SS \text{ Error}}{df2 \text{ Error}}$		

Figure 8.3: Equations for computing the ANOVA table for a repeated-measures design

to make the ANOVA table. The figure below presents an abstract for the repeated-measures ANOVA table. It shows us all the thing we need to calculate to get the F -value for our data.

So, what we need to do is calculate all the SS es that we did before for the between-subjects ANOVA. That means the next three steps are identical to the ones you did before. In fact, I will just basically copy the next three steps to find SS_{TOTAL} , SS_{Effect} , and $SS_{\text{Error (within-conditions)}}$. After that we will talk about splitting up $SS_{\text{Error (within-conditions)}}$ into two parts, this is the new thing for this chapter. Here we go!

8.3.1 SS Total

The total sums of squares, or $SSTotal$ measures the total variation in a set of data. All we do is find the difference between each score and the grand mean, then we square the differences and add them all up.

subjects	conditions	scores	diff	diff_squared
1	A	20	13	169
2	A	11	4	16
3	A	2	-5	25
1	B	6	-1	1
2	B	2	-5	25
3	B	7	0	0
1	C	2	-5	25
2	C	11	4	16
3	C	2	-5	25
Sums		63	0	302
Means		7	0	33.55555555555556

The mean of all of the scores is called the **Grand Mean**. It's calculated in the table, the Grand Mean = 7.

We also calculated all of the difference scores **from the Grand Mean**. The difference scores are in the column titled `diff`. Next, we squared the difference scores, and those are in the next column called `diff_squared`.

When you add up all of the individual squared deviations (difference sscores) you get the sums of squares. That's why it's called the sums of squares (SS).

Now, we have the first part of our answer:

$$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

$$SS_{\text{total}} = 302 \text{ and}$$

$$302 = SS_{\text{Effect}} + SS_{\text{Error}}$$

8.3.2 SS Effect

SS_{Total} gave us a number representing all of the change in our data, how they all are different from the grand mean.

What we want to do next is estimate how much of the total change in the data might be due to the experimental manipulation. For example, if we ran an experiment that causes causes change in the measurement, then the means for each group will be different from other, and the scores in each group will be different from each. As a result, the manipulation forces change onto the numbers, and this will naturally mean that some part of the total variation in the numbers is caused by the manipulation.

The way to isolate the variation due to the manipulation (also called effect) is to look at the means in each group, and the calculate the difference scores between each group mean and the grand mean, and then the squared deviations to find the sum for SS_{Effect} .

Consider this table, showing the calculations for SS_{Effect} .

subjects	conditions	scores	means	diff	diff_squared
1	A	20	11	4	16
2	A	11	11	4	16
3	A	2	11	4	16
1	B	6	5	-2	4
2	B	2	5	-2	4
3	B	7	5	-2	4
1	C	2	5	-2	4
2	C	11	5	-2	4
3	C	2	5	-2	4
Sums		63	63	0	72
Means		7	7	0	8

Notice we created a new column called `means`, these are the means for each condition, A, B, and C.

SS_{Effect} represents the amount of variation that is caused by differences between the means. The `diff` column is the difference between each condition mean and the grand mean, so for the first row, we have $11 - 7 = 4$, and so on.

We found that $SS_{\text{Effect}} = 72$, this is the same as the ANOVA from the previous chapter

8.3.3 SS Error (within-conditions)

Great, we made it to SS Error. We already found SS Total, and SS Effect, so now we can solve for SS Error just like this:

$$SS_{\text{total}} = SS_{\text{Effect}} + SS_{\text{Error (within-conditions)}}$$

switching around:

$$\$ \text{SS_Error} = \text{SS_total} - \text{SS_Effect} \$$$

$$\$ \text{SS_Error (within conditions)} = 302 - 72 = 230 \$$$

Or, we could compute $SS_{\text{Error (within conditions)}}$ directly from the data as we did last time:

subjects	conditions	scores	means	diff	diff_squared
1	A	20	11	-9	81
2	A	11	11	0	0
3	A	2	11	9	81
1	B	6	5	-1	1
2	B	2	5	3	9
3	B	7	5	-2	4
1	C	2	5	3	9
2	C	11	5	-6	36
3	C	2	5	3	9
Sums		63	63	0	230
Means		7	7	0	25.55555555555556

When we compute $SS_{\text{Error (within conditions)}}$ directly, we find the difference between each score and the condition mean for that score. This gives us the remaining error variation around the condition mean, that the condition mean does not explain.

8.3.4 SS Subjects

Now we are ready to calculate new partition, called SS_{Subjects} . We first find the means for each subject. For subject 1, this is the mean of their scores across Conditions A, B, and C. The mean for subject 1 is 9.33 (repeating). Notice there is going to be some rounding error here, that's OK for now.

The **means** column now shows all of the subject means. We then find the difference between each subject mean and the grand mean. These deviations are shown in the **diff** column. Then we square the deviations, and sum them up.

subjects	conditions	scores	means	diff	diff_squared
1	A	20	9.33	2.33	5.4289
2	A	11	8	1	1
3	A	2	3.66	-3.34	11.1556
1	B	6	9.33	2.33	5.4289
2	B	2	8	1	1
3	B	7	3.66	-3.34	11.1556
1	C	2	9.33	2.33	5.4289
2	C	11	8	1	1
3	C	2	3.66	-3.34	11.1556
Sums		63	62.97	-0.0299999999999994	52.7535
Means		7	6.99666666666667	-0.0033333333333326	5.8615

We found that the sum of the squared deviations $SS_{\text{Subjects}} = 52.75$. Note again, this has some small rounding error because some of the subject means had repeating decimal places, and did not divide evenly.

We can see the effect of the rounding error if we look at the sum and mean in the `diff` column. We know these should be both zero, because the Grand mean is the balancing point in the data. The sum and mean are both very close to zero, but they are not zero because of rounding error.

8.3.5 SS Error (left-over)

Now we can do the last thing. Remember we wanted to split up the $SS_{\text{Error (within conditions)}}$ into two parts, SS_{Subjects} and $SS_{\text{Error (left-over)}}$. Because we have already calculate $SS_{\text{Error (within conditions)}}$ and SS_{Subjects} , we can solve for $SS_{\text{Error (left-over)}}$:

$$SS_{\text{Error (left-over)}} = SS_{\text{Error (within conditions)}} - SS_{\text{Subjects}}$$

$$SS_{\text{Error (left-over)}} = SS_{\text{Error (within conditions)}} - SS_{\text{Subjects}} = 230 - 52.75 = 177.25$$

8.3.6 Check our work

Before we continue to compute the MSEs and F-value for our data, let's quickly check our work. For example, we could have R compute the repeated measures ANOVA for us, and then we could look at the ANOVA table and see if we are on the right track so far.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	2	52.66667	26.33333	NA	NA
conditions	2	72.00000	36.00000	0.8120301	0.505848
Residuals	4	177.33333	44.33333	NA	NA

OK, looks good. We found the SS_{Effect} to be 72, and the SS for the conditions (same thing) in the table is also 72. We found the SS_{Subjects} to be 52.75, and the SS for the first residual (same thing) in the table is also 53.66 repeating. That's close, and our number is off because of rounding error. Finally, we found the $SS_{\text{Error (left-over)}}$ to be 177.25, and the SS for the bottom residuals in the table (same thing) in the table is 177.33 repeating, again close but slightly off due to rounding error.

We have finished our job of computing the sums of squares that we need in order to do the next steps, which include computing the MSEs for the effect and the error term. Once we do that, we can find the F-value, which is the ratio of the two MSEs.

Before we do that, you may have noticed that we solved for $SS_{\text{Error (left-over)}}$, rather than directly computing it from the data. In this chapter we are not going to show you the steps for doing this. We are not trying to hide anything from, instead it turns out these steps are related to another important idea in ANOVA. We discuss this idea, which is called an **interaction** in the next chapter, when we discuss **factorial** designs (designs with more than one independent variable).

8.3.7 Compute the MSEs

Calculating the MSEs (mean squared error) that we need for the F-value involves the same general steps as last time. We divide each SS by the degrees of freedom for the SS.

The degrees of freedom for SS_{Effect} are the same as before, the number of conditions - 1. We have three conditions, so the df is 2. Now we can compute the MSE_{Effect} .

$$MSE_{\text{Effect}} = \frac{SS_{\text{Effect}}}{df} = \frac{72}{2} = 36$$

The degrees of freedom for $SS_{\text{Error (left-over)}}$ are different than before, they are the (number of subjects - 1) multiplied by the (number of conditions - 1). We have 3 subjects and three conditions, so $(3 - 1) * (3 - 1) = 2 * 2 = 4$. You might be wondering why we are multiplying these numbers. Hold that thought for now and wait until the next chapter. Regardless, now we can compute the $MSE_{\text{Error (left-over)}}$.

$$MSE_{\text{Error (left-over)}} = \frac{SS_{\text{Error (left-over)}}}{df} = \frac{177.33}{4} = 44.33$$

8.3.8 Compute F

We just found the two MSEs that we need to compute F . We went through all of this to compute F for our data, so let's do it:

$$F = \frac{MSE_{\text{Effect}}}{MSE_{\text{Error (left-over)}}} = \frac{36}{44.33} = 0.812$$

And, there we have it!

8.3.9 p-value

We already conducted the repeated-measures ANOVA using R and reported the ANOVA. Here it is again. The table shows the p -value associated with our F -value.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	2	52.66667	26.33333	NA	NA
conditions	2	72.00000	36.00000	0.8120301	0.505848
Residuals	4	177.33333	44.33333	NA	NA

We might write up the results of our experiment and say that the main effect condition was not significant, $F(2,4) = 0.812$, $MSE = 44.33$, $p = 0.505$.

What does this statement mean? Remember, that the p -value represents the probability of getting the F value we observed or larger under the null (assuming that the samples come from the same distribution, the assumption of no differences). So, we know that an F -value of 0.812 or larger happens fairly often by chance (when there are no real differences), in fact it happens 50.5% of the time. As a result, we do not reject the idea that any differences in the means we have observed could have been produced by chance.

8.4 Things worth knowing

Repeated Measures ANOVAs have some special properties that are worth knowing about. The main special property is that the error term used to for the F -value (the MSE in the denominator) will always be smaller than the error term used for the F -value the ANOVA for a between-subjects design. We discussed this earlier. It is smaller, because we subtract out the error associated with the subject means.

This can have the consequence of generally making F -values in repeated measures designs larger than F -values in between-subjects designs. When the number in the bottom of the F formula is generally smaller, it will generally make the resulting ratio a larger number. That's what happens when you make the number in the bottom smaller.

Because big F values usually let us reject the idea that differences in our means are due to chance, the repeated-measures ANOVA becomes a more sensitive test of the differences (its F -values are usually larger).

At the same time, there is a trade-off here. The repeated measures ANOVA uses different degrees of freedom for the error term, and these are typically a smaller number of degrees of freedom. So, the F -distributions for the repeated measures and between-subjects designs are actually different F -distributions, because they have different degrees of freedom.

8.4.1 Repeated vs between-subjects ANOVA

Let's do a couple simulations to see some the differences between the ANOVA for a repeated measures design, and the ANOVA for a between-subjects design.

We will do the following.

1. Simulate a design with three conditions, A, B, and C
2. sample 10 scores into each condition from the same normal distribution (mean = 100, SD = 10)
3. We will include a subject factor for the repeated-measures version. Here there are 10 subjects, each contributing three scores, one each condition
4. For the between-subjects design there are 30 different subjects, each contributing one score in the condition they were assigned to (really the group).

We run 1000 simulated experiments for each design. We calculate the F for each experiment, for both the between and repeated measures designs. Here are the two sampling distributions of F for both designs.

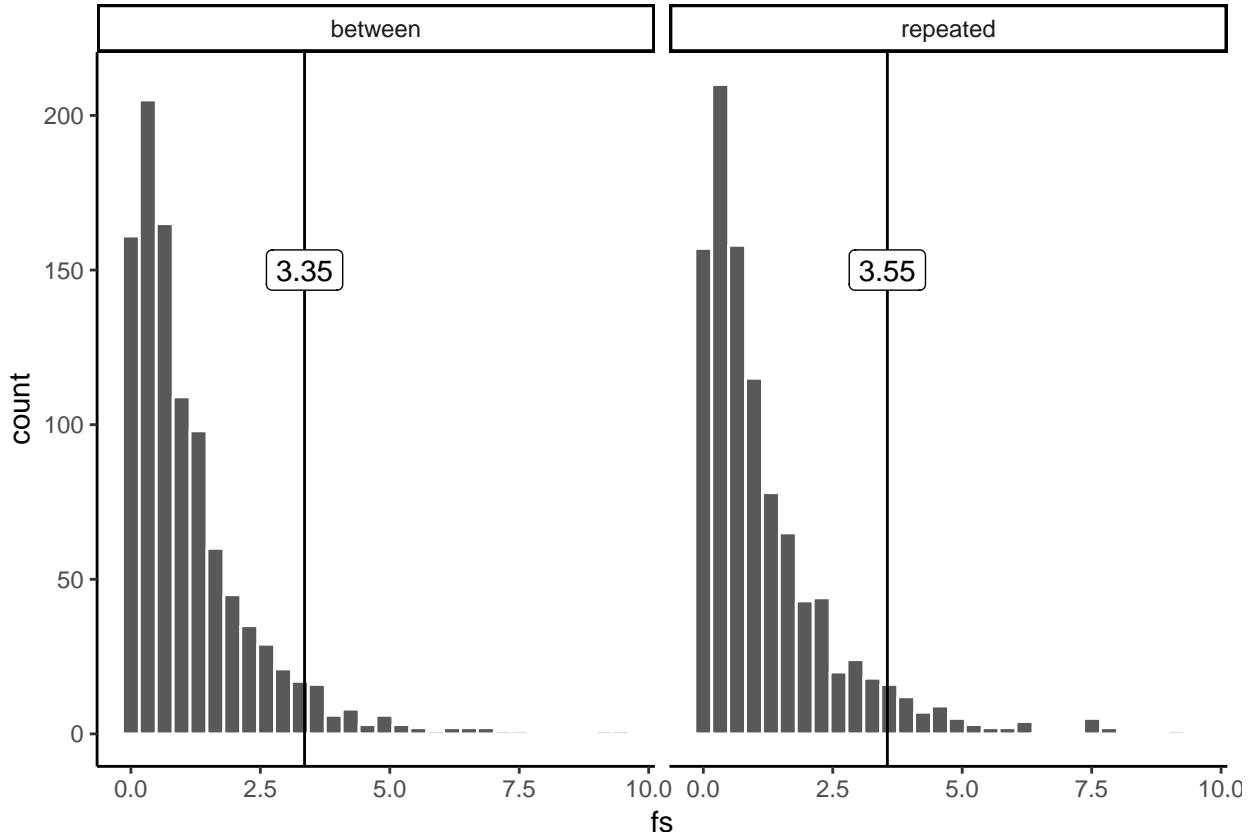


Figure 8.4: Comparing critical F values for a between and repeated measures design

These two F sampling distributions look pretty similar. However, they are subtly different. The between F distribution has degrees of freedom 2, and 27, for the numerator and denominator. There are 3 conditions, so $\text{df\$1} = 3-1 = 2$. There are 30 subjects, so $\text{df\$2} = 30-3 = 27$. The critical value, assuming an alpha of 0.05 is 3.35. This means F is 3.35 or larger 5% of the time under the null.

The repeated-measures F distribution has degrees of freedom 2, and 18, for the numerator and denominator. There are 3 conditions, so $\text{df\$1} = 3-1 = 2$. There are 10 subjects, so $\text{df\$2} = (10-1)(3-1) = 9 \times 2 = 18$. The critical value, assuming an alpha of 0.05 is 3.55. This means F is 3.55 or larger 5% of the time under the null.

The critical value for the repeated measures version is slightly higher. This is because when $\text{df\$2}$ (the denominator) is smaller, the F -distribution spreads out to the right a little bit. When it is skewed like this, we get some bigger F s a greater proportion of the time.

So, in order to detect a real difference, you need an F of 3.35 or greater in a between-subjects design, or an F of 3.55 or greater for a repeated-measures design. The catch here is that when there is a real difference

between the means, you will detect it more often with the repeated-measures design, even though you need a larger F (to pass the higher critical F -value for the repeated measures design).

8.4.2 repeated measures designs are more sensitive

To illustrate why repeated-measures designs are more sensitive, we will conduct another set of simulations.

We will do something slightly different this time. We will make sure that the scores for condition A, are always a little bit higher than the other scores. In other words, we will program in a real true difference. Specifically, the scores for condition will be sampled from a normal distribution with mean = 105, and SD = 10. This mean is 5 larger than the means for the other two conditions (still set to 100).

With a real difference in the means, we should now reject the hypothesis of no differences more often. We should find F values larger than the critical value more often. And, we should find p -values for each experiment that are smaller than .05 more often, those should occur more than 5% of the time.

To look at this we conduct 1000 experiments for each design, we conduct the ANOVA, then we save the p -value we obtained for each experiment. This is like asking how many times will we find a p -value less than 0.05, when there is a real difference (in this case an average of 5) between some of the means. We will plot histograms of the p -values:

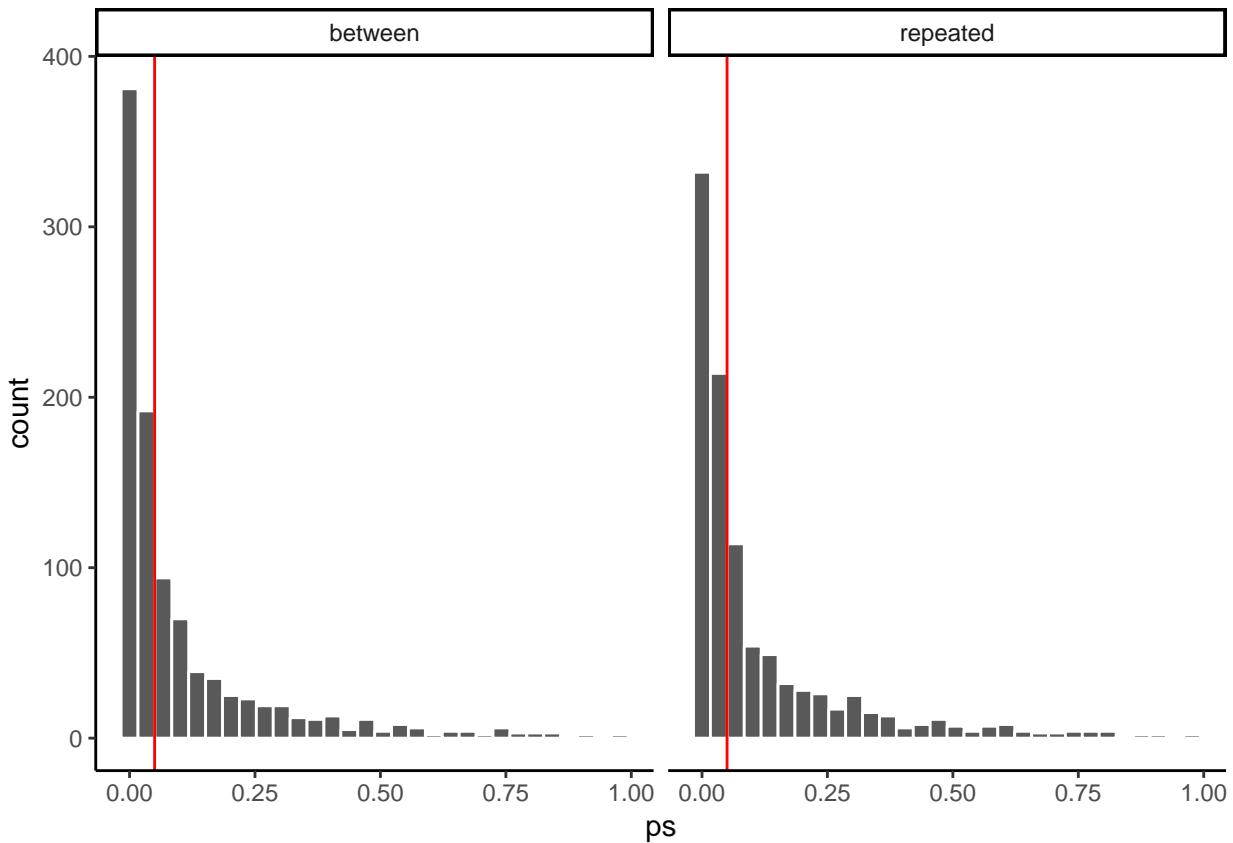


Figure 8.5: p -value distributions for a between and within-subjects ANOVA

Here we have two distributions of observed p -values for the simulations. The red line shows the location of 0.05. Overall, we can see that for both designs, we got a full range of p -values from 0 to 1. This means that many times we would not have rejected the hypothesis of no differences (even though we know there is a small difference). We would have rejected the null every time the p -value was less than 0.05.

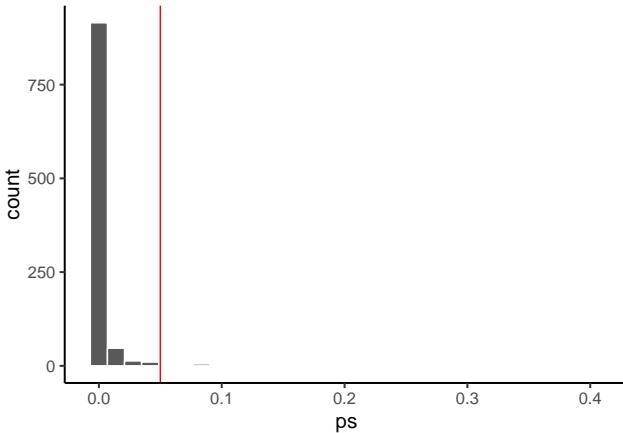


Figure 8.6: p-value distribution for within-subjects design with $n = 30$

For the between subject design, there were 570 experiments with a p less than 0.05, or 0.57 of experiments were “significant”, with $\alpha=.05$.

For the within subject design, there were 546 experiments with a p less than 0.05, or 0.546 of experiments were “significant”, with $\alpha=.05$.

OK, well, you still might not be impressed. In this case, the between-subjects design detected the true effect slightly more often than the repeated measures design. Both them were right around 55% of the time. Based on this, we could say the two designs are pretty comparable in their sensitivity, or ability to detect a true difference when there is one.

However, remember that the between-subjects design uses 30 subjects, and the repeated measures design only uses 10. We had to make a big investment to get our 30 subjects. And, we’re kind of unfairly comparing the between design (which is more sensitive because it has more subjects) with the repeated measures design that has fewer subjects.

What do you think would happen if we ran 30 subjects in the repeated measures design? Let’s find out. Here we redo the above, but this time only for the repeated measures design. We increase N from 10 to 30.

Wowsers! Look at that. When we ran 30 subjects in the repeated measures design almost all of the p -values were less than .05. There were 985 experiments with a p less than 0.05, or 0.985 of experiments were “significant”, with $\alpha=.05$. That’s huge! If we ran the repeated measures design, we would almost always detect the true difference when it is there. This is why the repeated measures design can be more sensitive than the between-subjects design.

8.5 Real Data

Let’s look at some real data from a published experiment that uses a repeated measures design. This is the same example that you will be using in the lab for repeated measures ANOVA. The data happen to be taken from a recent study conducted by Lawrence Behmer and myself, at Brooklyn College (Behmer and Crump, 2017).

We were interested in how people perform sequences of actions. One question is whether people learn individual parts of actions, or the whole larger pattern of a sequence of actions. We looked at these issues in a computer keyboard typing task. One of our questions was whether we would replicate some well known findings about how people type words and letters.

From prior work we knew that people type words way faster than random letters, but if you made the random letters a little bit more English-like, then people type those letter strings a little bit faster, but not

as slow as random string.

In the study, 38 participants sat in front of a computer and typed 5 letter strings one at a time. Sometimes the 5 letter made a word (Normal condition, TRUCK), sometimes they were completely random (Random Condition, JWYFG), and sometimes they followed patterns like you find in English (Bigram Condition, QUEND), but were not actual words. So, the independent variable for the typing material had three levels. We measured every single keystroke that participants made. This gave us a few different dependent measures. Let's take a look at the reaction times. This is how long it took for participants to start typing the first letter in the string.

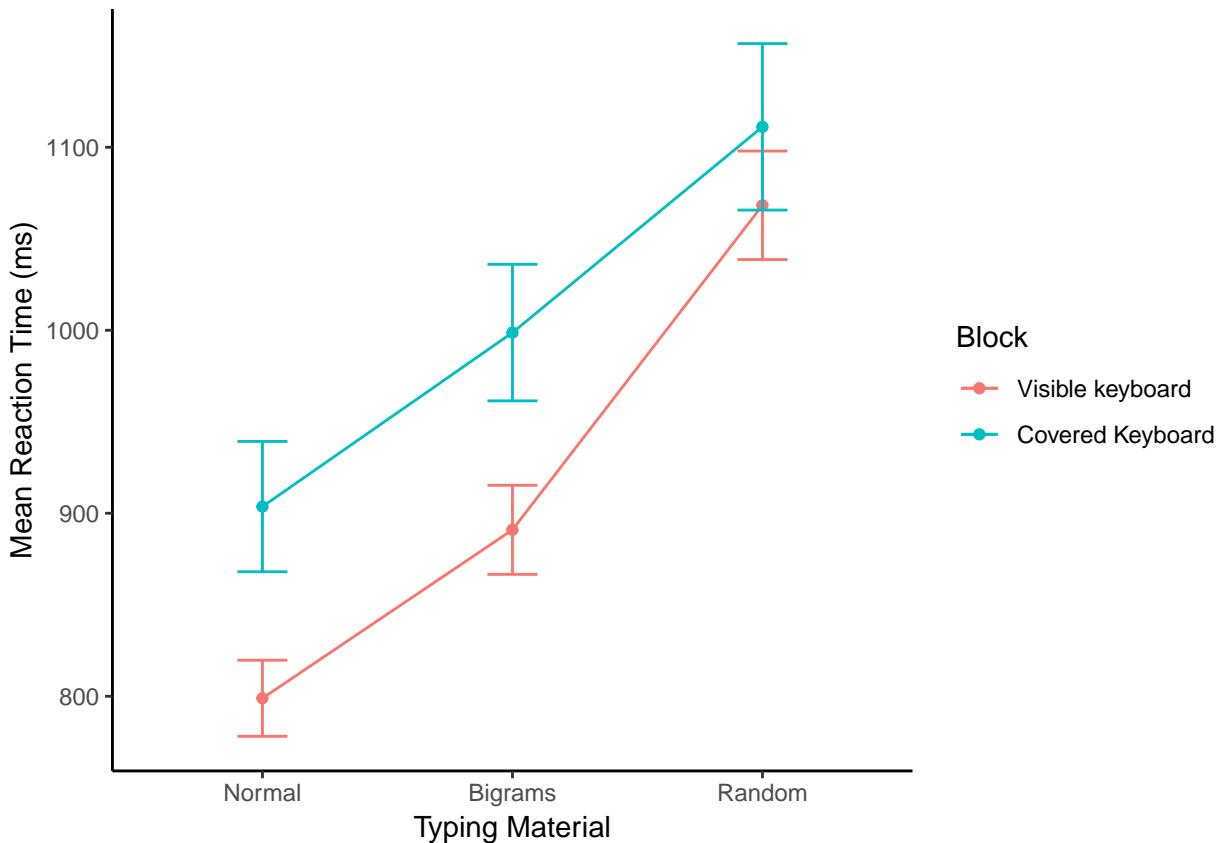


Figure 8.7: Results from Behmer & Crump (2017)

OK, I made a figure showing the mean reaction times for the different typing material conditions. You will notice that there are two sets of lines. That's because there was another manipulation I didn't tell you about. In one block of trials participants got to look at the keyboard while they typed, but in the other condition we covered up the keyboard so people had to type without looking. Finally, the error bars are standard error of the means.

Note, the use of error bars for repeated-measures designs is not very straightforward. In fact the standard error of the means that we have added here are not very meaningful for judging whether the differences between the means are likely not due to chance. They would be if this was a between-subjects design. We will update this textbook with a longer discussion of this issue, for now we will just live with these error bars.

For the purpose of this example, we will say, it sure looks like the previous finding replicated. For example, people started typing Normal words faster than Bigram strings (English-like), and they started typing random letters the most slowly of all. Just like prior research had found.

Let's focus only on the block of trials where participants were allowed to look at the keyboard while they

typed, that's the red line, for the “visible keyboard” block. We can see the means look different. Let's next ask, what is the likelihood that chance (random sampling error) could have produced these mean differences. To do that we run a repeated-measures ANOVA in R. Here is the ANOVA table.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	37	2452611.9	66286.808	NA	NA
Stimulus	2	1424914.0	712457.010	235.7342	0
Residuals1	74	223649.4	3022.289	NA	NA

Alright, we might report the results like this. There was a significant main effect of Stimulus type, $F(2, 74) = 235.73$, $MSE = 3022.289$, $p < 0.001$.

Notice a couple things. First, this is a huge F -value. It's 253! Notice also that the p-value is listed as 0. That doesn't mean there is zero chance of getting an F -value this big under the null. This is a rounding error. The true p-value is 0.00000000000000... The zeros keep going for a while. This means there is only a vanishingly small probability that these differences could have been produced by sampling error. So, we reject the idea that the differences between our means could be explained by chance. Instead, we are pretty confident, based on this evidence and previous work showing the same thing, that our experimental manipulation caused the difference. In other words, people really do type normal words faster than random letters, and they type English-like strings somewhere in the middle in terms of speed.

8.6 Summary

In this chapter you were introduced to the repeated-measures ANOVA. This analysis is appropriate for within-subjects or repeated measures designs. The main difference between the independent factor ANOVA and the repeated measures ANOVA, is the ability to partial out variance due to the individual subject means. This can often result in the repeated-measures ANOVA being more sensitive to true effects than the between-subjects ANOVA.

Chapter 9

Factorial ANOVA

We have arrived to the most complicated thing we are going to discuss in this class. Unfortunately, we have to warn you that you might find this next stuff a bit complicated. You might not, and that would be great! We will try our best to present the issues in a few different ways, so you have a few different tools to help you understand the issue.

What's this so very complicated issue? Well, the first part it isn't that complicated. For example, up until now we have been talking about experiments. Most every experiment has had two important bits, the independent variable (the manipulation), and the dependent variable (what we measure). In most cases, our independent variable has had two levels, or three or four; but, there has only been one independent variable.

What if you wanted to manipulate more than one independent variable? If you did that you would at least two independent variables, each with their own levels. The rest of the book is about designs with more than one independent variable, and the statistical tests we use to analyze those designs.

Let's go through some examples of designs so can see what we are talking about. We will be imagining experiments that are trying to improve students grades. So, the dependent variable will always be grade on a test.

1. 1 IV (two levels)

We would use a t-test for these designs, because they only have two levels.

- a. Time of day (Morning versus Afternoon): Do students do better on tests when they take them in the morning versus the afternoon? There is one IV (time of day), with two levels (Morning vs. Afternoon)
- b. Caffeine (some caffeine vs no caffeine): Do students do better on tests when they drink caffeine versus not drinking caffeine? There is one IV (caffeine), with two levels (some caffeine vs no caffeine)

2. 1 IV (three levels):

We would use an ANOVA for these designs because they have more than two levels

- a. Time of day (Morning, Afternoon, Night): Do students do better on tests when they take them in the morning, the afternoon, or at night? There is one IV (time of day), with three levels (Morning, Afternoon, and Night)
- b. Caffeine (1 coffee, 2 coffees, 3 coffees): Do students do better on tests when they drink 1 coffee, 2 coffees, or three coffees? There is one IV (caffeine), with three levels (1 coffee, 2 coffees, and 3 coffees)

3. 2 IVs, IV1 (two levels), IV2 (two levels)

We haven't talked about what kind of test to run for this design (hint it is called a factorial ANOVA)

- a. IV1 (Time of Day: Morning vs. Afternoon); IV2 (Caffeine: some caffeine vs. no caffeine): How does time of day and caffeine consumption influence student grades? We had students take tests in the

2x2 Design		IV 1		Time of Day	
		IV1: Level 1	IV1: Level 2		
IV 2	IV2: Level 1	dv	dv	Some Caffeine	dv
	IV2: Level 2	dv	dv	No Caffeine	dv

Figure 9.1: Structure of 2x2 factorial designs

morning or in the afternoon, with or without caffeine. There are two IVs (time of day & caffeine). IV1 (Time of day) has two levels (morning vs afternoon). IV2 (caffeine) has two levels (some caffeine vs. no caffeine)

OK, let's stop here for the moment. The first two designs both had one IV. The third design shows an example of a design with 2 IVs (time of day and caffeine), each with two levels. This is called a **2x2 Factorial Design**. It is called a **factorial** design, because the levels of each independent variable are fully crossed. This means that first each level of one IV, the levels of the other IV are also manipulated. "HOLD ON STOP PLEASE!" Yes, it seems as if we are starting to talk in the foreign language of statistics and research designs. We apologize for that. We'll keep mixing it up with some plain language, and some pictures.

9.1 Factorial basics

9.1.1 2x2 Designs

We've just started talking about a **2x2 Factorial design**. We said this means the IVs are crossed. To illustrate this, take a look at the following tables. We show an abstract version and a concrete version using time of day and caffeine as the two IVs, each with two levels in the design:

Let's talk about this crossing business. Here's what it means for the design. For the first level of Time of Day (morning), we measure test performance when some people drank caffeine and some did not. So, in the morning we manipulate whether or not caffeine is taken. Also, in the second level of the Time of Day (afternoon), we also manipulate caffeine. Some people drink or don't drink caffeine in the afternoon as well, and we collect measures of test performance in both conditions.

We could say the same thing, but talk from the point of view of the second IV. For example, when people drink caffeine, we test those people in the morning, and in the afternoon. So, time of day is manipulated for the people who drank caffeine. Also, when people do not drink caffeine, we test those people in the morning, and in the afternoon. So, time of day is manipulated for the people who did not drink caffeine.

Finally, each of the four squares representing a DV, is called a **condition**. So, we have 2 IVs, each with 2 levels, for a total of 4 conditions. This is why we call it a 2x2 design. $2 \times 2 = 4$. The notation tells us how to calculate the total number of conditions.

2x3 Design		IV 1		Time of Day	
		IV1: Level 1	IV1: Level 2		
IV 2	IV2: Level 1	dv	dv	1 coffee	dv
	IV2: Level 2	dv	dv	2 coffees	dv
	IV2: Level 3	dv	dv	3 coffees	dv

Figure 9.2: Structure of 2x3 factorial design

9.1.2 Factorial Notation

Anytime **all of the levels** of each IV in a design are fully crossed, so that they all occur for each level of every other IV, we can say the design is a **fully factorial** design.

We use a notation system to refer to these designs. The rules for notation are as follows. Each IV gets its own number. The number of levels in the IV is the number we use for the IV. Let's look at some examples:

2x2 = There are two IVs, the first IV has two levels, the second IV has 2 levels. There are a total of 4 conditions, $2 \times 2 = 4$.

2x3 = There are two IVs, the first IV has two levels, the second IV has three levels. There are a total of 6 conditions, $2 \times 3 = 6$.

3x2 = There are two IVs, the first IV has three levels, the second IV has two levels. There are a total of 6 conditions, $3 \times 2 = 6$.

4x4 = There are two IVs, the first IV has 4 levels, the second IV has 4 levels. There are a total of 16 conditions, $4 \times 4 = 16$.

2x3x2 = There are a total of three IVs. The first IV has 2 levels. The second IV has 3 levels. The third IV has 2 levels. There are a total of 12 conditions. $2 \times 3 \times 2 = 12$.

9.1.3 2 x 3 designs

Just for fun, let's illustrate a 2x3 design using the same kinds of tables we looked at before for the 2x2 design.

All we did was add another row for the second IV. It's a 2x3 design, so it should have 6 conditions. As you can see there are now 6 cells to measure the DV.

9.2 Purpose of Factorial Designs

Factorial designs let researchers manipulate more than one thing at once. This immediately makes things more complicated, because as you will see, there are many more details to keep track of. Why would researchers want to make things more complicated? Why would they want to manipulate more than one IV at a time.

Before we go on, let's clarify what we mean by manipulating more than one thing at once. When you have one IV in your design, by definition, you are manipulating only **one** thing. This might seem confusing at first, because the IV has more than one level, so it seems to have more than one manipulation. Consider manipulating the number of coffees that people drink before they do a test. We could have one IV (coffee), with three levels (1, 2, or 3 coffees). You might want to say we have three manipulations here, drinking 1, 2, or 3 coffees. But, the way we define manipulation is in terms of the IV. There is only one coffee IV. It does

have three levels. Nevertheless, we say you are only doing one coffee manipulation. The only thing you are manipulating is the amount of coffee. That's just one thing, so it's called one manipulation. To do another, second manipulation, you need to additionally manipulate something that is not coffee (like time of day in our previous example).

Returning to our question: why would researchers want to manipulate more than one thing in their experiment. The answer might be kind of obvious. They want to know if more than one thing causes change in the thing they are measuring! For example, if you are measuring people's happiness, you might assume that more than one thing causes happiness to change. If you wanted to track down how two things caused changes in happiness, then you might want to have two manipulations of two different IVs. This is not a wrong way to think about the reasons why researchers use factorial designs. They are often interested in questions like this. However, we think this is an unhelpful way to first learn about factorial designs.

We present a slightly different way of thinking about the usefulness of factorial designs, and we think it is so important, it gets its own section.

9.2.1 Factorials manipulate an effect of interest

Here is how researchers often use factorial designs to understand the causal influences behind the effects they are interested in measuring. Notice we didn't say the dependent variables they are measuring, we are now talking about something called effects. Effects are the change in a measure caused by a manipulation. You get an effect, any time one IV causes a change in a DV.

Here is an example. We will stick with this one example for a while, so pay attention... In fact, the example is about paying attention. Let's say you wanted to measure something like paying attention. You could something like this:

1. Pick a task for people to do that you can measure. For example, you can measure how well they perform the task. That will be the dependent measure
2. Pick a manipulation that you think will cause differences in paying attention. For example, we know that people can get distracted easily when there are distracting things around. You could have two levels for your manipulation: No distraction versus distraction.
3. Measure performance in the task under the two conditions
4. If your distraction manipulation changes how people perform the task, you may have successfully manipulated **how well people can pay attention** in your task.

9.2.2 Spot the difference

Let's elaborate this with another fake example. First, we pick a task. It's called **spot the difference**. You may have played this game before. You look at two pictures side-by-side, and then you locate as many differences as you can find. here is an example:

How many differences can you spot? When you look for the differences, it feels like you are doing something we would call "paying attention". If you pay attention to the clock tower, you will see that the hands on the clock are different. Ya! One difference spotted.

We could give people 30 seconds to find as many differences as they can. Then we give them another set of pictures and do it again. Every time we will measure how many differences they can spot. So, our measure of performance, our dependent variable, could be the mean number of differences spotted.



Figure 9.3: Spot the differences between the two images

9.2.3 Distraction manipulation

Now, let's think about a manipulation that might cause differences in how people pay attention. If people need to pay attention to spot differences, then presumably if we made it difficult to pay attention, people would spot less differences. What is a good way to distract people? I'm sure there are lots of ways to do this. How about we do the following:

1. No distraction condition: Here people do the task with no added distractions. They sit in front of a computer, in a quiet, distraction-free room, and find as many differences as they can for each pair of pictures
2. Distraction condition: Here we blast super loud ambulance sounds and fire alarms and heavy metal music while people attempt to spot differences. We also randomly turn the sounds on and off, and make them super-duper annoying and distracting. We make sure that the sounds aren't loud enough to do any physical damage to anybody's ear-drums. But, we want to make them loud enough to be super distracting. If you don't like this, we could also tickle people with a feather, or whisper silly things into their ears, or surround them by clowns, or whatever we want, it just has to be super distracting.

9.2.4 Distraction effect

If our distraction manipulation is super-distracting, then what should we expect to find when we compare spot-the-difference performance between the no-distraction and distraction conditions? We should find a difference!

If our manipulation works, then we should find that people find more differences when they are not distracted, and less differences when they are distracted. For example, the data might look something like this:

The figure shows a big difference in the mean number of difference spotted. People found 5 differences on average when they were distracted, and 10 differences when they were not distracted. We labelled the figure, "The distraction effect", because it shows a big effect of distraction. The effect of distraction is a mean

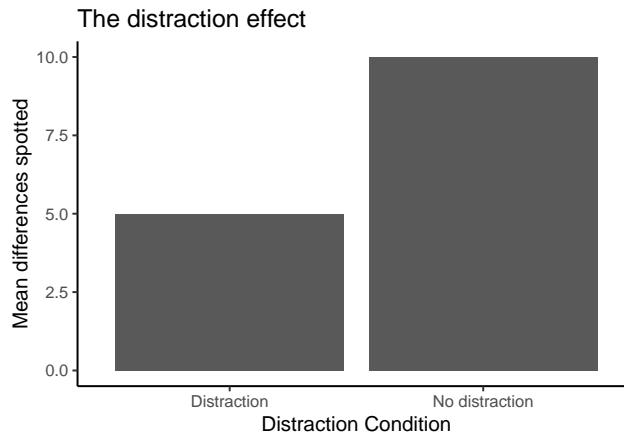


Figure 9.4: Example data from pretend experiment showing number of differences spotted in a distraction versus no distraction condition

of 5 spot the differences. It's the difference between performance in the Distraction and No-Distraction conditions. In general, it is very common to use the word **effect** to refer to the differences caused by the manipulation. We manipulated distraction, it caused a difference, so we call this the “distraction effect”.

9.2.5 Manipulating the Distraction effect

This is where factorial designs come in to play. We have done the hard work of finding an effect of interest, in this case the distraction effect. We think this distraction effect actually measures something about your ability to pay attention. For example, if you were the kind of person who had a small distraction effect (maybe you find 10 differences when you are not distracted, and 9 differences when you are distracted), that could mean you are very good at ignoring distracting things while you are paying attention. On the other hand, you could be the kind of person who had a big distraction effect (maybe you found 10 differences under no distraction, and only 1 difference when you were distracted); this could mean you are not very good at ignoring distracting things while you are paying attention.

Overall now, we are thinking of our distraction effect (the difference in performance between the two conditions) as the important thing we want to measure. We then might want to know how to make people better at ignoring distracting things. Or, we might want to know what makes people worse at ignoring things. In other words we want to find out what manipulations control the size of the distraction effect (make it bigger or smaller, or even flip around!).

Maybe there is a special drug that helps you ignore distracting things. People taking this drug should be less distracted, and if they took this drug while completing our task, they should have a smaller distraction effect compared to people not taking the drug.

Maybe rewarding people with money can help you pay attention and ignore distracting things better. People receiving 5 dollars every time they spot a difference might be able to focus more because of the reward, and they would show a smaller distraction effect in our task, compared to people who got no money for finding differences. Let's see what this would look like.

We are going to add a second IV to our task. The second IV will manipulate reward. In one condition, people will get 5 dollars for every difference they find (so they could leave the study with lots of money if they find lots of differences). In the other condition, people will get no money, but they will still have find differences. Remember, this will be a factorial design, so everybody will have to find differences when they are distracted and when they are not distracted.

The question we are now asking is: Will manipulating reward cause a change in the size of the distraction effect. We could predict that people receiving rewards will have a smaller distraction effect than people not

receiving rewards. If that happened, the data would look something like this:

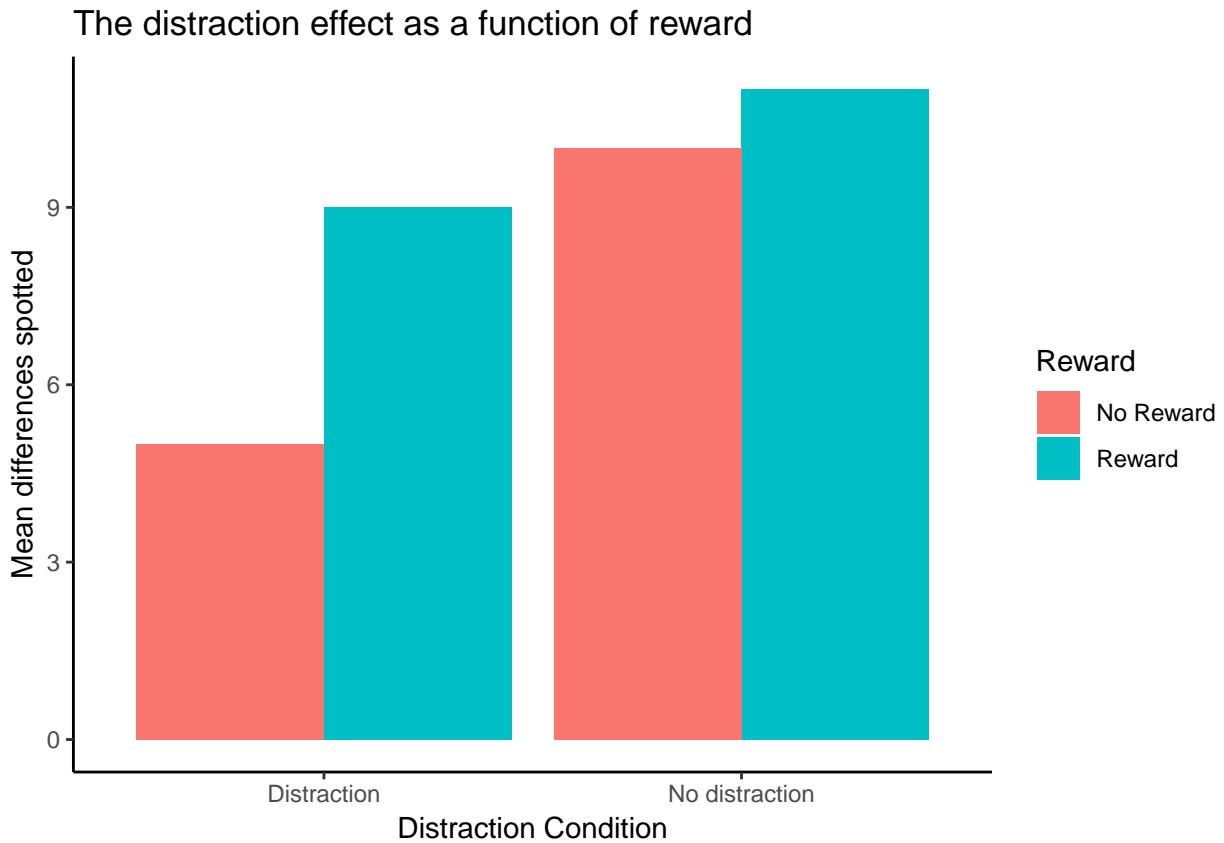


Figure 9.5: Example data showing how the distraction effect could be modulated by a reward manipulation. Distraction condition plotted on the x-axis, makes it more difficult to compare the changes in the distraction effect between reward conditions

I've just shown you a new kind of graph. I apologize right now for showing this to you first. It's more unhelpful than the next graph. What I did was keep the x-axis the same as before (to be consistent). So, we have distraction vs. no distraction on the x-axis. In the distraction condition, there are means for spot-the-difference performance in the no-reward (red), and reward (aqua) conditions. The same goes for the no-distraction condition, a red and an aqua bar for the no-reward and reward conditions. We can try to interpret this graph, but the next graph plots the same data in a different way, which makes it easier to see what we are talking about.

All we did was change the x-axis. Now the left side of the x-axis is for the no-reward condition, and the right side is for the reward condition. The red bar is for the distraction condition, and the aqua bar is for the no distraction condition. It is easier to see the distraction effect in this graph. The distraction effect is the difference in size between the red and aqua bars. For each reward condition, the red and aqua bars are right beside each other, so can see if there is a difference between them more easily, compared to the first graph.

No-Reward condition: In the no-reward condition people played spot the difference when they were distracted and when they were not distracted. This is a replication of our first fake study. We should expect to find the same pattern of results, and that's what the graph shows. There was a difference of 5. People found 5 differences when they were distracted and 10 when they were not distracted. So, there was a distraction effect of 5, same as we had last time.

Reward condition: In the reward condition people played spot the difference when they were distracted and when they were not distracted. Except, they got 5 dollars every time they spotted a difference. We

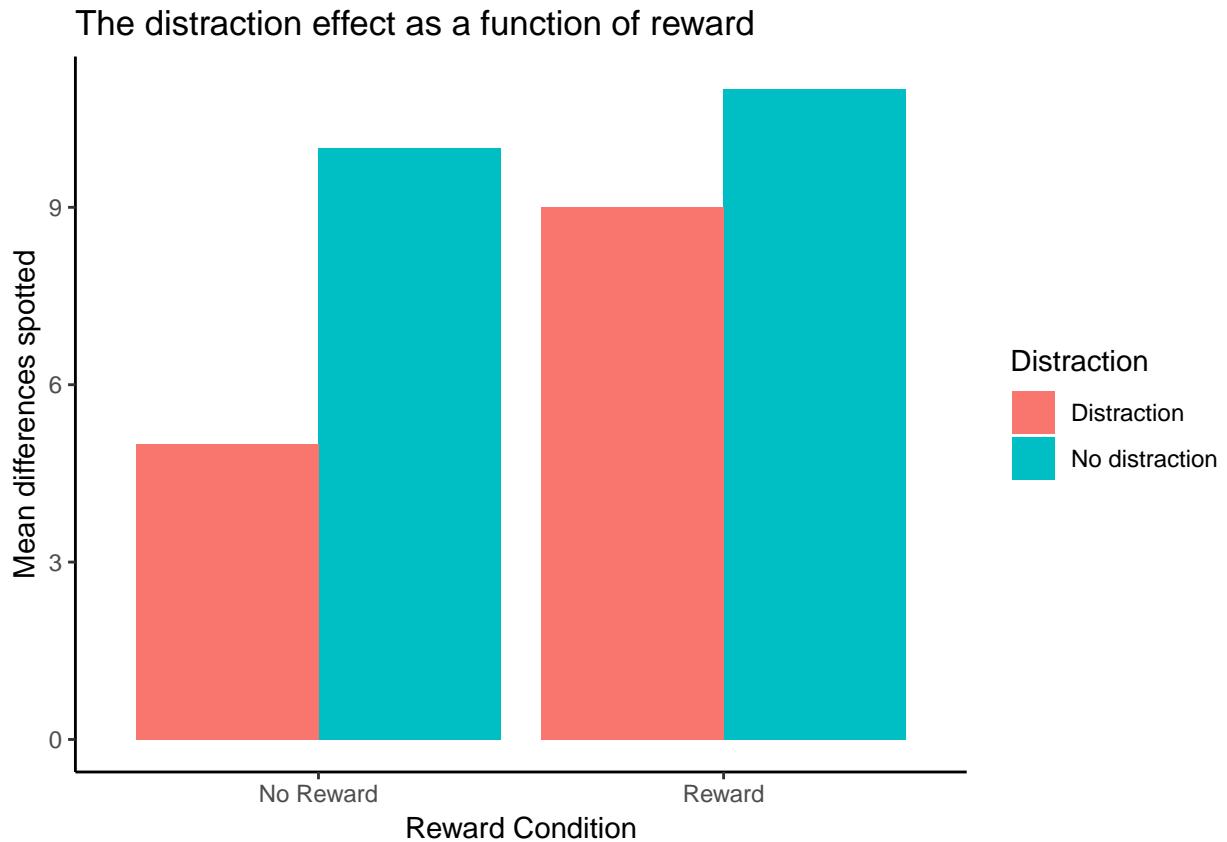


Figure 9.6: Example data showing how the distraction effect could be modulated by a reward manipulation. Reward condition plotted on the x-axis, makes it easier to compare the changes in the distraction effect between reward conditions

predicted this would cause people to pay more attention and do a better job of ignoring distracting things. The graph shows this is what happened. People found 9 differences when they were distracted and 11 when they were not distracted. So, there was a distraction effect of 2.

If we had conducted this study, we might have concluded that reward can manipulate the distraction effect. When there was no reward, the size of the distraction effect was 5. When there was reward, the size of the distraction effect was 2. So, the reward manipulation changed the size of the distraction effect by 3 ($5-2=3$).

This is our description of why factorial designs are so useful. They allow researchers to find out what kinds of manipulations can cause changes in the effects they measure. We measured the distraction effect, then we found that reward causes changes in the distraction effect. If we were trying to understand how paying attention works, we would then need to explain how it is that reward levels could causally change how people pay attention. We would have some evidence that reward does cause change in paying attention, and we would have to come up with some explanations, and then run more experiments to test whether those explanations hold water.

9.3 Graphing the means

In our example above we showed you two bar graphs of the very same means for our 2x2 design. Even though the graphs plot identical means, they look different, so they are more or less easy to interpret by looking at them. Results from 2x2 designs are also often plotted with line graphs. Those look different too. Here are four different graphs, using bars and lines to plot the very same means from before. We are showing you this so that you realize **how you graph your data** matters, and it makes it more or less easy for people to understand the results. Also, how the data is plotted matters for what you need to look at to interpret the results.

9.4 Knowing what you want to find out

When you conduct a design with more than one IV, you get more means to look at. As a result, there are more kinds of questions that you can ask of the data. Sometimes it turns out that the questions that you can ask, are not the ones that you want to ask, or have an interest in asking. Because you ran the design with more than one IV, you have the opportunity to ask these kinds of extra questions.

What kinds of new things are we talking about? Let's keep going with our distraction effect experiment. We have the first IV where we manipulated distraction. So, we could find the overall means in spot-the difference for the distraction vs. no-distraction conditions (that's two means). The second IV was reward. We could find the overall means in spot-the-difference performance for the reward vs. no-reward conditions (that's two more means). We could do what we already did, and look at the means for each combination, that is the mean for distraction/reward, distraction/no-reward, no-distraction/reward, and no-distraction/no-reward (that's four more means, if you're counting). There's even more. We could look at the mean distraction effect (the difference between distraction and no-distraction) for the reward condition, and the mean distraction effect for the no-reward condition (that's two more). I hope you see here that there are a lot of means to look. And they are all different means. Let's look at all of them together in one graph with four panels.

The purpose of showing all of these means is to orient you to your problem. If you conduct a 2x2 design (and this is the most simple factorial that you can conduct), you will get all of these means. You need to know what you want to know from the means. That is, you need to be able to connect the research question to the specific means you are interested in analyzing.

For example, in our example, the research question was whether reward would change the size of the distraction effect. The top left panel gives us some info about this question. We can see all of the condition means, and we can visually see that the distraction effect was larger in the No-reward compared to the reward

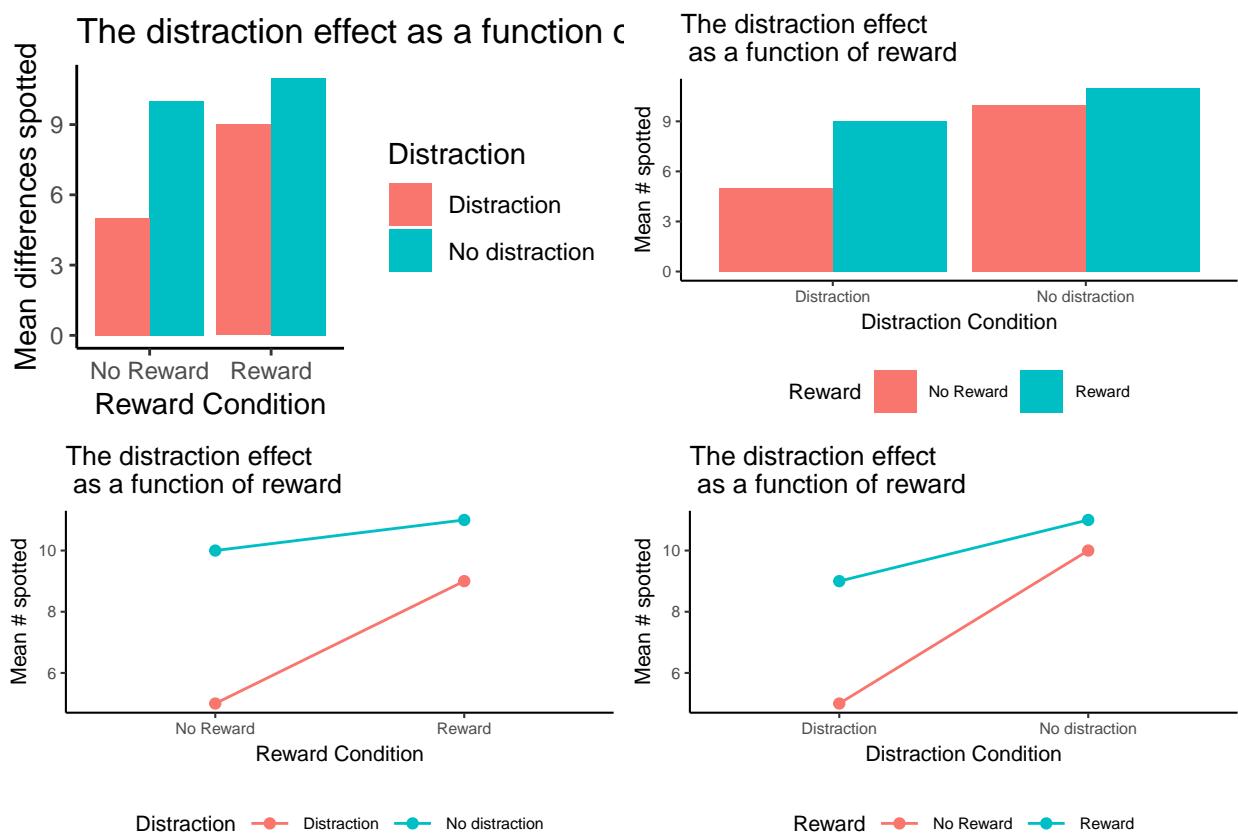


Figure 9.7: The same example means plotted using bar graphs or line graphs, and with Distraction or Reward on the x-axis

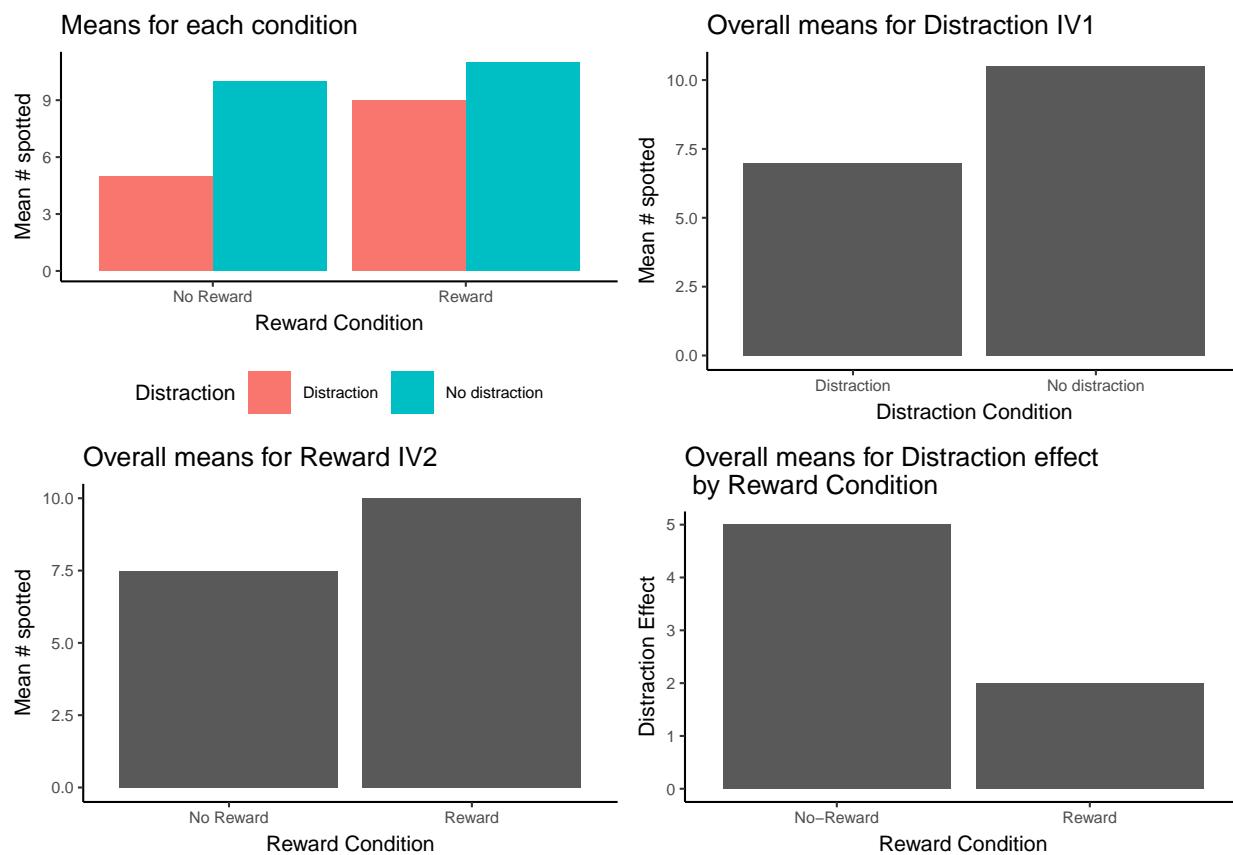


Figure 9.8: Each panel shows the mean for different effects in the design

condition. But, to “see” this, we need to do some visual subtraction. You need to look at the difference between the red and aqua bars for each of the reward and no-reward conditions.

Does the top right panel tell us about whether reward changed the size of the distraction effect? NO, it just shows that there was an overall distraction effect (this is called the **main effect** of distraction). **Main effects** are any differences between the levels of one independent variable.

Does the bottom left panel tell us about whether reward changed the size of the distraction effect? NO! it just shows that there was an overall reward effect, called the main effect of reward. People who were rewarded spotted a few more differences than the people who weren’t, but this doesn’t tell us if they were any less distracted.

Finally, how about the bottom left panel. Does this tell us about whether the reward changed the size of the distraction effect? YES! Notice, the y-axis is different for this panel. The y-axis here is labelled “Distraction Effect”. You are looking at two difference scores. The distraction effect in the no-reward condition ($10-5 = 5$), and the distraction effect in the Reward condition ($11-9 = 2$). These two bars are different as a function of reward. So, it looks like reward did produce a difference between the distraction effects! This was the whole point of the fake study. It is these means that were most important for answering the question of the study. As a very last point, this panel contains what we call an **interaction**. We explain this in the next section.

Pro tip: Make sure you know what you want to know from your means before you run the study, otherwise you will just have way too many means, and you won’t know what they mean.

9.5 Simple analysis of 2x2 repeated measures design

Normally in a chapter about factorial designs we would introduce you to Factorial ANOVAs, which are totally a thing. We will introduce you to them soon. But, before we do that, we are going to show you how to analyze a 2x2 repeated measures ANOVA design with paired-samples t-tests. This is probably something you won’t do very often. However, it turns out the answers you get from this method are the same ones you would get from an ANOVA.

Admittedly, if you found the explanation of ANOVA complicated, it will just appear even more complicated for factorial designs. So, our purpose here is to delay the complication, and show you with t-tests what it is that the Factorial ANOVA is doing. More important, when you do the analysis with t-tests, you have to be very careful to make all of the comparisons in the right way. As a result, you will get some experience learning how to know what it is you want to know from factorial designs. Once you know what you want to know, you can use the ANOVA to find out the answers, and then you will also know what answers to look for after you run the ANOVA. Isn’t new knowledge fun!

The first thing we need to do is define **main effects** and **interactions**. Whenever you conduct a Factorial design, you will also have the opportunity to analyze **main effects** and **interactions**. However, the number of **main effects** and **interactions** you get to analyse depends on the number of IVs in the design.

9.5.1 Main effects

Formally, main effects are the mean differences for a single Independent variable. There is always one main effect for each IV. A 2x2 design has 2 IVs, so there are two main effects. In our example, there is one main effect for distraction, and one main effect for reward. We will often ask if the main effect of some IV is significant. This refers to a statistical question: Were the differences between the means for that IV likely or unlikely to be caused by chance (sampling error).

If you had a 2x2x2 design, you would measure three main effects, one for each IV. If you had a 3x3x3 design, you would still only have 3 IVs, so you would have three main effects.

9.5.2 Interaction

We find that the interaction concept is one of the most confusing concepts for factorial designs. Formally, we might say an interaction occurs whenever the effect of one IV has an influence on the size of the effect for another IV. That's probably not very helpful. In more concrete terms, using our example, we found that the reward IV had an effect on the size of the distraction effect. The distraction effect was larger when there was no-reward, and it was smaller when there was a reward. So, there was an interaction.

We might also say an interaction occurs when the difference between the differences are different! Yikes. Let's explain. There was a difference in spot-the-difference performance between the distraction and no-distraction condition, this is called the distraction effect (it is a difference measure). The reward manipulation changed the size of the distraction effect, that means there was difference in the size of the distraction effect. The distraction effect is itself a measure of differences. So, we did find that the difference (in the distraction effect) between the differences (the two measures of the distraction effect between the reward conditions) were different. When you start to write down explanations of what interactions are, you find out why they come across as complicated. We'll leave our definition of interaction like this for now. Don't worry, we'll go through lots of examples to help firm up this concept for you.

The number of interactions in the design also depend on the number of IVs. For a 2x2 design there is only 1 interaction. The interaction between IV1 and IV2. This occurs when the effect of say IV2 (whether there is a difference between the levels of IV2) changes across the levels of IV1. We could write this in reverse, and ask if the effect of IV1 (whether there is a difference between the levels of IV1) changes across the levels of IV2. However, just because we can write this two ways, does not mean there are two interactions. We'll see in a bit, that no matter how do the calculation to see if the difference scores—measure of effect for one IV—change across the levels of the other IV, we always get the same answer. That is why there is only one interaction for a 2x2. Similarly, there is only one interaction for a 3x3, because there again we only have two IVs (each with three levels). Only when we get up to designs with more than 2 IVs, do we find more possible interactions. A design with three IVs, has four interactions. If the IVs are labelled A, B, and C, then we have three 2-way interactions (AB, AC, and BC), and one three-way interaction (ABC). We hold off on this stuff for much later

9.5.3 Looking at the data

It is most helpful to see some data in order to understand how we will analyze it. Let's imagine we ran our fake attention study. We will have five people in the study, and they will participate in all conditions, so it will be a fully repeated-measures design. The data could look like this:

subject	No Reward		Reward	
	No Distraction	Distraction	No Distraction	Distraction
	A	B	C	D
1	10	5	12	9
2	8	4	13	8
3	11	3	14	10
4	9	4	11	11
5	10	2	13	12

Note:

Number of differences spotted for each subject in each condition.

9.5.4 Main effect of Distraction

The main effect of distraction compares the overall means for all scores in the no-distraction and distraction conditions, collapsing over the reward conditions.

All Conditions							
	No Reward		Reward		Distraction Means		Distraction Effect
	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	Difference
subject	A	B	C	D	AC	BD	AC.minus.BD
1	10	5	12	9	11	7	4
2	8	4	13	8	10.5	6	4.5
3	11	3	14	10	12.5	6.5	6
4	9	4	11	11	10	7.5	2.5
5	10	2	13	12	11.5	7	4.5
Means					11.1	6.8	4.3

Figure 9.9: Computing the main effect of distraction

The yellow columns show the no-distraction scores for each subject. The blue columns show the distraction scores for each subject.

The overall means for each subject, for the two distraction conditions are shown to the right. For example, subject 1 had a 10 and 12 in the no-distraction condition, so their mean is 11.

We are interested in the main effect of distraction. This is the difference between the AC column (average of subject scores in the no-distraction condition) and the BD column (average of the subject scores in the distraction condition). These differences for each subject are shown in the last green column. The overall means, averaging over subjects are in the bottom green row.

Just looking at the means, we can see there was a main effect of Distraction, the mean for the no-distraction condition was 11.1, and the mean for the distraction condition was 6.8. The size of the main effect was 4.3 (the difference between 11.1 and 6.8).

Now, what if we wanted to know if this main effect of distraction (the difference of 4.3) could have been caused by chance, or sampling error. You could do two things. You could run a paired samples *t*-test between the mean no-distraction scores for each subject (column AC) and the mean distraction scores for each subject (column BD). Or, you could run a one-sample *t*-test on the difference scores column, testing against a mean difference of 0. Either way you will get the same answer.

Here's the paired samples version:

```
##
## Paired t-test
##
## data: AC and BD
## t = 7.6615, df = 4, p-value = 0.00156
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
## 2.741724 5.858276
## sample estimates:
## mean of the differences
## 4.3
```

Here's the one sample version:

```
##
## One Sample t-test
##
## data: AC - BD
## t = 7.6615, df = 4, p-value = 0.00156
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 2.741724 5.858276
## sample estimates:
## mean of x
## 4.3
```

If we were to write-up our results for the main effect of distraction we could say something like this:

The main effect of distraction was significant, $t(4) = 7.66$, $p = 0.001$. The mean number of differences spotted was higher in the no-distraction condition ($M = 11.1$) than the distraction condition ($M = 6.8$).

9.5.5 Main effect of Reward

The main effect of reward compares the overall means for all scores in the no-reward and reward conditions, collapsing over the reward conditions.

The yellow columns show the no-reward scores for each subject. The blue columns show the reward scores for each subject.

The overall means for each subject, for the two reward conditions are shown to the right. For example, subject 1 had a 10 and 5 in the no-reward condition, so their mean is 7.5.

We are interested in the main effect of reward. This is the difference between the AB column (average of subject scores in the no-reward condition) and the CD column (average of the subject scores in the reward condition). These differences for each subject are shown in the last green column. The overall means, averaging over subjects are in the bottom green row.

Just looking at the means, we can see there was a main effect of reward. The mean number of differences spotted was 11.3 in the reward condition, and 6.6 in the no-reward condition. So, the size of the main effect of reward was 4.7.

Is a difference of this size likely or unlikely due to chance? We could conduct a paired-samples t -test on the AB vs. CD means, or a one-sample t -test on the difference scores. They both give the same answer:

Here's the paired samples version:

```
##
## Paired t-test
##
## data: CD and AB
## t = 8.3742, df = 4, p-value = 0.001112
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 3.141724 6.258276
## sample estimates:
## mean of the differences
```

All Conditions							
	No Reward		Reward		Reward Means		Reward Effect
	No Distraction	Distraction	No Distraction	Distraction	No Reward	Reward	Difference
subject	A	B	C	D	AB	CD	CD.minus.AB
1	10	5	12	9	7.5	10.5	3
2	8	4	13	8	6	10.5	4.5
3	11	3	14	10	7	12	5
4	9	4	11	11	6.5	11	4.5
5	10	2	13	12	6	12.5	6.5
Means					6.6	11.3	4.7

Figure 9.10: Computing the main effect of reward

```
##          4.7
```

Here's the one sample version:

```
##
##  One Sample t-test
##
## data: CD - AB
## t = 8.3742, df = 4, p-value = 0.001112
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  3.141724 6.258276
## sample estimates:
## mean of x
##        4.7
```

If we were to write-up our results for the main effect of reward we could say something like this:

The main effect of reward was significant, $t(4) = 8.37$, $p = 0.001$. The mean number of differences spotted was higher in the reward condition ($M = 11.3$) than the no-reward condition ($M = 6.6$).

9.5.6 Interaction between Distraction and Reward

Now we are ready to look at the interaction. Remember, the whole point of this fake study was what? Can you remember?

Here's a reminder. We wanted to know if giving rewards versus not would change the size of the distraction effect.

All Conditions							
subject	No Reward		Reward		Distraction Effects		Interaction Effect
	No Distraction	Distraction	No Distraction	Distraction	No Reward	Reward	Difference
1	10	5	12	9	5	3	2
2	8	4	13	8	4	5	-1
3	11	3	14	10	8	4	4
4	9	4	11	11	5	0	5
5	10	2	13	12	8	1	7
Means				6	2.6	3.4	

Figure 9.11: Computing the interaction between distraction and reward

Notice, neither the main effect of distraction, or the main effect of reward, which we just went through the process of computing, answers this question.

In order to answer the question we need to do two things. First, compute distraction effect for each subject when they were in the no-reward condition. Second, compute the distraction effect for each subject when they were in the reward condition.

Then, we can compare the two distraction effects and see if they are different. The comparison between the two distraction effects is what we call the **interaction effect**. Remember, this is a difference between two difference scores. We first get the difference scores for the distraction effects in the no-reward and reward conditions. Then we find the difference scores between the two distraction effects. This difference of differences is the interaction effect (green column in the table)

The mean distraction effects in the no-reward (6) and reward (2.6) conditions were different. This difference is the interaction effect. The size of the interaction effect was 3.4.

How can we test whether the interaction effect was likely or unlikely due to chance? We could run another paired-sample *t*-test between the two distraction effect measures for each subject, or a one sample *t*-test on the green column (representing the difference between the differences). Both of these *t*-tests will give the same results:

Here's the paired samples version:

```
##  
## Paired t-test
```

```
##
## data: A_B and C_D
## t = 2.493, df = 4, p-value = 0.06727
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.3865663 7.1865663
## sample estimates:
## mean of the differences
## 3.4
```

Here's the one sample version:

```
##
## One Sample t-test
##
## data: A_B - C_D
## t = 2.493, df = 4, p-value = 0.06727
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -0.3865663 7.1865663
## sample estimates:
## mean of x
## 3.4
```

Oh look, the interaction was not significant. At least, if we had set our alpha criterion to 0.05, it would not have met that criteria. We could write up the results like this. The two-way interaction between between distraction and reward was not significant, $t(4) = 2.493, p = 0.067$.

Often times when a result is “not significant” according to the alpha criteria, the pattern among the means is not described further. One reason for this practice is that the researcher is treating the means as if they are not different (because there was an above alpha probability that the observed idfferences were due to chance). If they are not different, then there is no pattern to report.

There are differences in opinion among reasonable and expert statisticians on what should or should not be reported. Let's say we wanted to report the observed mean differences, we would write something like this:

The two-way interaction between between distraction and reward was not significant, $t(4) = 2.493, p = 0.067$. The mean distraction effect in the no-reward condition was 6 and the mean distraction effect in the reward condition was 2.6.

9.5.7 Writing it all up

We have completed an analysis of a 2x2 repeated measures design using paired-samples t -tests. Here is what a full write-up of the results could look like.

The main effect of distraction was significant, $t(4) = 7.66, p = 0.001$. The mean number of differences spotted was higher in the no-distraction condition ($M = 11.1$) than the distraction condition ($M = 6.8$).

The main effect of reward was significant, $t(4) = 8.37, p = 0.001$. The mean number of differences spotted was higher in the reward condition ($M = 11.3$) than the no-reward condition ($M = 6.6$).

The two-way interaction between between distraction and reward was not significant, $t(4) = 2.493, p = 0.067$. The mean distraction effect in the no-reward condition was 6 and the mean distraction effect in the reward condition was 2.6.

Interim Summary. We went through this exercise to show you how to break up the data into individual comparisons of interest. Generally speaking, a 2x2 repeated measures design would not be anlayzed with three paired-samples t -test. This is because it is more convenient to use the repeated measures ANOVA for

this task. We will do this in a moment to show you that they give the same results. And, by the same results, what we will show is that the p -values for each main effect, and the interaction, are the same. The ANOVA will give us F -values rather than t values. It turns out that in this situation, the F -values are related to the t values. In fact, $t^2 = F$.

9.5.8 2x2 Repeated Measures ANOVA

We just showed how a 2x2 repeated measures design can be analyzed using paired-sampled t -tests. We broke up the analysis into three parts. The main effect for distraction, the main effect for reward, and the 2-way interaction between distraction and reward. We claimed the results of the paired-samples t -test analysis would mirror what we would find if we conducted the analysis using an ANOVA. Let's show that the results are the same. Here are the results from the 2x2 repeated-measures ANOVA, using the `aov` function in R.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	4	3.70	0.925	NA	NA
Distraction	1	92.45	92.450	58.698413	0.0015600
Residuals	4	6.30	1.575	NA	NA
Reward	1	110.45	110.450	70.126984	0.0011122
Residuals1	4	6.30	1.575	NA	NA
Distraction:Reward	1	14.45	14.450	6.215054	0.0672681
Residuals	4	9.30	2.325	NA	NA

Let's compare these results with the paired-samples t -tests.

Main effect of Distraction: Using the paired samples t -test, we found $t(4) = 7.6615$, $p=0.00156$. Using the ANOVA we found, $F(1,4) = 58.69$, $p=0.00156$. See, the p -values are the same, and $t^2 = 7.6615^2 = 58.69 = F$.

Main effect of Reward: Using the paired samples t -test, we found $t(4) = 8.3742$, $p=0.001112$. Using the ANOVA we found, $F(1,4) = 70.126$, $p=0.001112$. See, the p -values are the same, and $t^2 = 8.3742^2 = 70.12 = F$.

Interaction effect: Using the paired samples t -test, we found $t(4) = 2.493$, $p=0.06727$. Using the ANOVA we found, $F(1,4) = 6.215$, $p=0.06727$. See, the p -values are the same, and $t^2 = 2.493^2 = 6.215 = F$.

There you have it. The results from a 2x2 repeated measures ANOVA are the same as you would get if you used paired-samples t -tests for the main effects and interactions.

9.6 2x2 Between-subjects ANOVA

You must be wondering how to calculate a 2x2 ANOVA. We haven't discussed this yet. We've only shown you that you don't have to do it when the design is a 2x2 repeated measures design (note this is a special case).

We are now going to work through some examples of calculating the ANOVA table for 2x2 designs. We will start with the between-subjects ANOVA for 2x2 designs. We do essentially the same thing that we did before (in the other ANOVAs), and the only new thing is to show how to compute the interaction effect.

Remember the logic of the ANOVA is to partition the variance into different parts. The SS formula for the between-subjects 2x2 ANOVA looks like this:

$$SS_{\text{Total}} = SS_{\text{Effect IV1}} + SS_{\text{Effect IV2}} + SS_{\text{Effect IV1xIV2}} + SS_{\text{Error}}$$

In the following sections we use tables to show the calculation of each SS. We use the same example as before with the exception that **we are turning this into a between-subjects design**. There are now 5 different subjects in each condition, for a total of 20 subjects. As a result, we remove the subjects column.

All Conditions				Difference from Grand Mean				Squared Differences			
No Reward		Reward		No Reward		Reward		No Reward		Reward	
No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction
A	B	C	D	A-GrandM	B-GrandM	C-GrandM	D-GrandM	(A-GrandM)^2	(B-GrandM)^2	(C-GrandM)^2	(D-GrandM)^2
10	5	12	9	1.05	-3.95	3.05	0.05	1.1025	15.6025	9.3025	0.0025
8	4	13	8	-0.95	-4.95	4.05	-0.95	0.9025	24.5025	16.4025	0.9025
11	3	14	10	2.05	-5.95	5.05	1.05	4.2025	35.4025	25.5025	1.1025
9	4	11	11	0.05	-4.95	2.05	2.05	0.0025	24.5025	4.2025	4.2025
10	2	13	12	1.05	-6.95	4.05	3.05	1.1025	48.3025	16.4025	9.3025
Means	9.6	3.6	12.6	10							
Grand Mean	8.95										
sums						Sums		7.3125	148.3125	71.8125	15.5125
SS Total						SS Total		242.95			

Figure 9.12: Computing SS total

9.6.1 SS Total

We calculate the grand mean (mean of all of the score). Then, we calculate the differences between each score and the grand mean. We square the difference scores, and sum them up. That is SS_{Total} , reported in the bottom yellow row.

9.6.2 SS Distraction

We need to compute the SS for the main effect for distraction. We calculate the grand mean (mean of all of the scores). Then, we calculate the means for the two distraction conditions. Then we treat each score as if it was the mean for it's respective distraction condition. We find the differences between each distraction condition mean and the grand mean. Then we square the differences and sum them up. That is $SS_{Distraction}$, reported in the bottom yellow row.

These tables are a lot to look at! Notice here, that we first found the grand mean (8.95). Then we found the mean for all the scores in the no-distraction condition (columns A and C), that was 11.1. All of the difference scores for the no-distraction condition are $11.1 - 8.95 = 2.15$. We also found the mean for the scores in the distraction condition (columns B and D), that was 6.8. So, all of the difference scores are $6.8 - 8.95 = -2.15$. Remember, means are the balancing point in the data, this is why the difference scores are +2.15 and -2.15. The grand mean 8.95 is in between the two condition means (11.1 and 6.8), by a difference of 2.15.

9.6.3 SS Reward

We need to compute the SS for the main effect for reward. We calculate the grand mean (mean of all of the scores). Then, we calculate the means for the two reward conditions. Then we treat each score as if it was the mean for it's respective reward condition. We find the differences between each reward condition mean and the grand mean. Then we square the differences and sum them up. That is SS_{Reward} , reported in the bottom yellow row.

All Conditions				Distraction Mean - GM				Squared Differences			
No Reward		Reward		No Reward		Reward		No Reward		Reward	
No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction
A	B	C	D	NDM-GM A	DM-GM B	NDM-GM C	DM-GM D	(NDM-GM)^2 A	(DM-GM)^2 B	(NDM-GM)^2 C	(DM-GM)^2 D
10	5	12	9	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
8	4	13	8	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
11	3	14	10	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
9	4	11	11	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
10	2	13	12	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
Means	9.6	3.6	12.6	10							
Grand Mean	8.95	No Distraction	11.1	Distraction	6.8						
sums						Sums		23.1125	23.1125	23.1125	23.1125
SS Distraction						SS Distraction		92.45			

Figure 9.13: Computing the main effect of distraction

All Conditions				Reward Mean - GM				Squared Differences			
No Reward		Reward		No Reward		Reward		No Reward		Reward	
No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction
A	B	C	D	NRM-GM A	NRM-GM B	RM-GM C	RM-GM D	(NRM-GM)^2 A	(NRM-GM)^2 B	(RM-GM)^2 C	(RM-GM)^2 D
10	5	12	9	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
8	4	13	8	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
11	3	14	10	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
9	4	11	11	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
10	2	13	12	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
Means	9.6	3.6	12.6	10							
Grand Mean	8.95	No Reward	6.6	Reward	11.3						
sums						Sums		27.6125	27.6125	27.6125	27.6125
SS Reward						SS Reward		110.45			

Figure 9.14: Computing the main effect of reward

Now we treat each no-reward score as the mean for the no-reward condition (6.6), and subtract it from the grand mean (8.95), to get -2.35. Then, we treat each reward score as the mean for the reward condition (11.3), and subtract it from the grand mean (8.95), to get +2.35. Then we square the differences and sum them up.

9.6.4 SS Distraction by Reward

We need to compute the SS for the interaction effect between distraction and reward. This is the new thing that we do in an ANOVA with more than one IV. How do we calculate the variation explained by the interaction?

The heart of the question is something like this. Do the individual means for each of the four conditions do something a little bit different than the group means for both of the independent variables.

For example, consider the overall mean for all of the scores in the no reward group, we found that to be 6.6. Now, was the mean for each no-reward group in the whole design a 6.6? For example, in the no-distraction group, was the mean for column A (the no-reward condition in that group) also 6.6? The answer is no, it was 9.6. How about the distraction group? Was the mean for the reward condition in the distraction group (column B) 6.6? No, it was 3.6. The mean of 9.6 and 3.6 is 6.6. If there was no hint of an interaction, we would expect that the means for the reward condition in both levels of the distraction group would be the same, they would both be 6.6. However, when there is an interaction, the means for the reward group will depend on the levels of the group from another IV. In this case, it looks like there is an interaction because the means are different from 6.6, they are 9.6 and 3.6 for the no-distraction and distraction conditions. This is extra-variance that is not explained by the mean for the reward condition. We want to capture this extra variance and sum it up. Then we will have measure of the portion of the variance that is due to the interaction between the reward and distraction conditions.

What we will do is this. We will find the four condition means. Then we will see how much additional variation they explain beyond the group means for reward and distraction. To do this we treat each score as the condition mean for that score. Then we subtract the mean for the distraction group, and the mean for the reward group, and then we add the grand mean. This gives us the unique variation that is due to the interaction. We could also say that we are subtracting each condition mean from the grand mean, and then adding back in the distraction mean and the reward mean, that would amount to the same thing, and perhaps make more sense.

Here is a formula to describe the process for each score:

$$\bar{X}_{\text{condition}} = \bar{X}_{\text{IV1}} - \bar{X}_{\text{IV2}} + \bar{X}_{\text{Grand Mean}}$$

Or we could write it this way:

$$\bar{X}_{\text{condition}} = \bar{X}_{\text{Grand Mean}} + \bar{X}_{\text{IV1}} + \bar{X}_{\text{IV2}}$$

When you look at the following table, we apply this formula to the calculation of each of the differences scores. We then square the difference scores, and sum them up to get $SS_{\text{Interaction}}$, which is reported in the bottom yellow row.

9.6.5 SS Error

The last thing we need to find is the SS Error. We can solve for that because we found everything else in this formula:

$$SS_{\text{Total}} = SS_{\text{Effect IV1}} + SS_{\text{Effect IV2}} + SS_{\text{Effect IV1xIV2}} + SS_{\text{Error}}$$

Even though this textbook meant to explain things in a step by step way, we guess you are tired from watching us work out the 2x2 ANOVA by hand. You and me both, making these tables was a lot of work.

All Conditions				Interaction Differences				Squared Differences			
No Reward		Reward		No Reward		Reward		No Reward		Reward	
No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction
A	B	C	D	A-ND-NR+GM	B-D-NR+GM	C-ND-R+GM	D-D-R+GM	(A-ND-NR+GM)^2 A	(B-D-NR+GM)^2 B	(C-ND-R+GM)^2 C	(D-D-R+GM)^2 D
10	5	12	9	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
8	4	13	8	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
11	3	14	10	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
9	4	11	11	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
10	2	13	12	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
Means	9.6	3.6	12.6	10							
Grand Mean	8.95										
sums							Sums	3.6125	3.6125	3.6125	3.6125
SS Interaction							SS Interaction	14.45			

Figure 9.15: Computing the interaction between distraction and reward

We have already shown you how to compute the SS for error before, so we will not do the full example here. Instead, we solve for SS Error using the numbers we have already obtained.

$$\text{SS Error} = \text{SS Total} - \text{SS Effect IV1} - \text{SS Effect IV2} - \text{SS Effect IV1xIV2}$$

$$\text{SS Error} = 242.95 - 92.45 - 110.45 - 14.45 = 25.6$$

9.6.6 Check your work

We are going to skip the part where we divide the SSes by their dfs to find the MSEs so that we can compute the three F -values. Instead, if we have done the calculations of the SSes correctly, they should be same as what we would get if we used R to calculate the SSes. Let's make R do the work, and then compare to check our work.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Distraction	1	92.45	92.45	57.78125	0.0000011
Reward	1	110.45	110.45	69.03125	0.0000003
Distraction:Reward	1	14.45	14.45	9.03125	0.0083879
Residuals	16	25.60	1.60	NA	NA

A quick look through the column Sum Sq shows that we did our work by hand correctly. Congratulations to us! Note, this is not the same results as we had before with the repeated measures ANOVA. We conducted a between-subjects design, so we did not get to further partition the SS error into a part due to subject variation and a left-over part. We also gained degrees of freedom in the error term. It turns out with this specific set of data, we find p-values of less than 0.05 for all effects (main effects and the interaction, which was not less than 0.05 using the same data, but treating it as a repeated-measures design)

9.7 Fireside chat

Sometimes it's good to get together around a fire and have a chat. Let's pretend we're sitting around a fire.

It's been a long day. A long couple of weeks and months since we started this course on statistics. We just went through the most complicated things we have done so far. This is a long chapter. What should we do next?

Here's a couple of options. We could work through, by hand, more and more ANOVAs. Do you want to do that? I don't, making these tables isn't too bad, but it takes a lot of time. It's really good to see everything that we do laid bare in the table form a few times. We've done that already. It's really good for you to attempt to calculate an ANOVA by hand at least once in your life. It builds character. It helps you know that you know what you are doing, and what the ANOVA is doing. We can't make you do this, we can only make the suggestion. If we keep doing these by hand, it is not good for us, and it is not you doing them by hand. So, what are the other options.

The other options are to work at a slightly higher level. We will discuss some research designs, and the ANOVAs that are appropriate for their analysis. We will conduct the ANOVAs using R, and print out the ANOVA tables. This is what you do in the lab, and what most researchers do. They use software most of the time to make the computer do the work. Because of this, it is most important that you know what the software is doing. You can make mistakes when telling software what to do, so you need to be able to check the software's work so you know when the software is giving you wrong answers. All of these skills are built up over time through the process of analyzing different data sets. So, for the remainder of our discussion on ANOVAs we stick to that higher level. No more monster tables of SSes. You are welcome.

9.8 Real Data

Let's go through the process of looking at a 2x2 factorial design in the wild. This will be the very same data that you will analyze in the lab for factorial designs.

9.8.1 Stand at attention

Do you pay more attention when you are sitting or standing? This was the kind of research question the researchers were asking in the study we will look at. In fact, the general question and design is very similar to our fake study idea that we used to explain factorial designs in this chapter.

The paper we look at is called "Stand by your Stroop: Standing up enhances selective attention and cognitive control" (Rosenbaum et al., 2017). This paper asked whether sitting versus standing would influence a measure of selective attention, the ability to ignore distracting information.

They used a classic test of selective attention, called the Stroop effect. You may already know what the Stroop effect is. In a typical Stroop experiment, subjects name the color of words as fast as they can. The trick is that sometimes the color of the word is the same as the name of the word, and sometimes it is not. Here are some examples:

Congruent trials occur when the color and word match. So, the correct answers for each of the congruent stimuli shown would be to say, red, green, blue and yellow. Incongruent trials occur when the color and word mismatch. The correct answers for each of the incongruent stimuli would be: blue, yellow, red, green.

The Stroop effect is an example of a well-known phenomena. What happens is that people are faster to name the color of the congruent items compared to the color of the incongruent items. This difference (incongruent reaction time - congruent reaction time) is called the Stroop effect.

Many researchers argue that the Stroop effect measures something about selective attention, the ability to ignore distracting information. In this case, the target information that you need to pay attention to is the color, not the word. For each item, the word is potentially distracting, it is not information that you are supposed to respond to. However, it seems that most people can't help but notice the word, and their performance in the color-naming task is subsequently influenced by the presence of the distracting word.

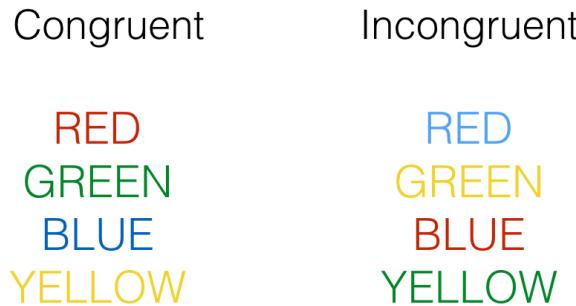


Figure 9.16: Examples of congruent and incongruent Stroop stimuli. The task is to name the color, not the word

People who are good at ignoring the distracting words should have small Stroop effects. They will ignore the word, and it won't influence them very much for either congruent or incongruent trials. As a result, the difference in performance (the Stroop effect) should be fairly small (if you have “good” selective attention in this task). People who are bad at ignoring the distracting words should have big Stroop effects. They will not ignore the words, causing them to be relatively fast when the word helps, and relatively slow when the word mismatches. As a result, they will show a difference in performance between the incongruent and congruent conditions.

If we take the size of the Stroop effect as a measure of selective attention, we can then start wondering what sorts of things improve selective attention (e.g., that make the Stroop effect smaller), and what kinds of things impair selective attention (e.g., make the Stroop effect bigger).

The research question of this study was to ask whether standing up improves selective attention compared to sitting down. They predicted smaller Stroop effects when people were standing up and doing the task, compared to when they were sitting down and doing the task.

The design of the study was a 2x2 repeated-measures design. The first IV was congruency (congruent vs incongruent). The second IV was posture (sitting vs. standing). The DV was reaction time to name the word.

9.8.2 Plot the data

They had subjects perform many individual trials responding to single Stroop stimuli, both congruent and incongruent. And they had subjects stand up sometimes and do it, and sit-down sometimes and do it. Here is a graph of what they found:

The figure shows the means. We can see that Stroop effects were observed in both the sitting position and the standing position. In the sitting position, mean congruent RTs were shorter than mean incongruent RTs (the red bar is lower than the aqua bar). The same general pattern is observed for the standing position. However, it does look as if the Stroop effect is slightly smaller in the stand condition: the difference between the red and aqua bars is slightly smaller compared to the difference when people were sitting.

9.8.3 Conduct the ANOVA

Let's conduct a 2x2 repeated measures ANOVA on the data to evaluate whether the differences in the means are likely or unlikely to be due to chance. The ANOVA will give us main effects for congruency and posture (the two IVs), as well as one interaction effect to evaluate (congruency X posture). Remember, the interaction effect tells us whether the congruency effect changes across the levels of the posture manipulation.

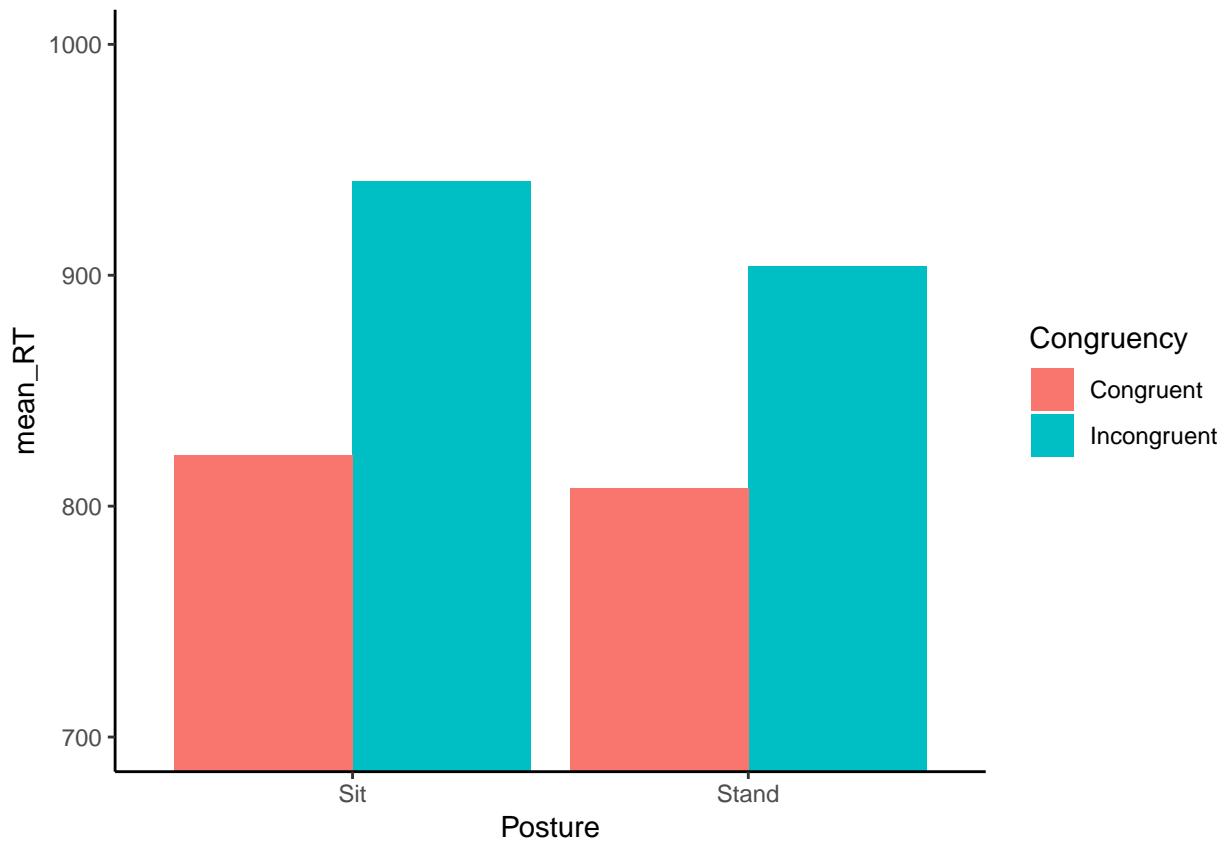


Figure 9.17: Means from Rosenbaum et al (2017)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	49	2250738.636	45933.4416	NA	NA
Congruency	1	576821.635	576821.6349	342.452244	0.0000000
Residuals	49	82534.895	1684.3856	NA	NA
Posture	1	32303.453	32303.4534	7.329876	0.0093104
Residuals1	49	215947.614	4407.0942	NA	NA
Congruency:Posture	1	6560.339	6560.3389	8.964444	0.0043060
Residuals	49	35859.069	731.8177	NA	NA

9.8.4 Main effect of Congruency

Let's talk about each aspect of the ANOVA table, one step at a time. First, we see that there was a significant main effect of congruency, $F(1, 49) = 342.45, p < 0.001$. The F value is extremely large, and the p -value is so small it reads as a zero. This F -value basically never happens by sampling error. We can be very confident that the overall mean difference between congruent and incongruent RTs was not caused by sampling error.

What were the overall mean differences between mean RTs in the congruent and incongruent conditions? We would have to look at those means to find out. Here's a table:

mean_rt	sd	SEM
868.6454	126.8237	8.967789

The table shows the mean RTs, standard deviation (sd), and standard error of the mean for each condition. These means show that there was a Stroop effect. Mean incongruent RTs were slower (larger number in milliseconds) than mean congruent RTs. The main effect of congruency is important for establishing that the researchers were able to measure the Stroop effect. However, the main effect of congruency does not say whether the size of the Stroop effect changed between the levels of the posture variable. So, this main effect was not particularly important for answering the specific question posed by the study.

9.8.5 Main effect of Posture

There was also a main effect of posture, $F(1, 49) = 7.329, p = 0.009$.

Let's look at the overall means for the sitting and standing conditions and see what this is all about:

mean_rt	sd	SEM
868.6454	126.8237	8.967789

Remember, the posture main effect collapses over the means in the congruency condition. We are not measuring a Stroop effect here. We are measuring a general effect of sitting vs standing on overall reaction time. The table shows that people were a little faster overall when they were standing, compared to when they were sitting.

Again, the main effect of posture was not the primary effect of interest. The authors weren't interested if people are in general faster when they stand. They wanted to know if their selective attention would improve when they stand vs when they sit. They were most interested in whether the size of the Stroop effect (difference between incongruent and congruent performance) would be smaller when people stand, compared to when they sit. To answer this question, we need to look at the interaction effect.

9.8.6 Congruency X Posture Interaction

Last, there was a significant congruency X posture interaction, $F(1, 49) = 8.96, p = 0.004$.

With this information, and by looking at the figure, we can get a pretty good idea of what this means. We know the size of the Stroop effect must have been different between the standing and sitting conditions, otherwise we would have gotten a smaller F -value and a much larger p -value.

We can see from the figure the direction of this difference, but let's look at the table to see the numbers more clearly.

mean_rt	sd	SEM
868.6454	126.8237	8.967789

In the sitting condition the Stroop effect was roughly $941 - 822 = 119$ ms.

In the standing condition the Stroop effect was roughly $904 - 808 = 96$ ms.

So, the Stroop effect was $119 - 96 = 23$ ms smaller when people were standing. This is a pretty small effect in terms of the amount of time reduced, but even though it is small, a difference even this big was not very likely to be due to chance.

9.8.7 What does it all mean?

Based on this research there appears to be some support for the following logic chain. First, the researchers can say that standing up reduces the size of a person's Stroop effect. Fine, what could that mean? Well, if the Stroop effect is an index of selective attention, then it could mean that standing up is one way to improve your ability to selectively focus and ignore distracting information. The actual size of the benefit is fairly small, so the real-world implications are not that clear. Nevertheless, maybe the next time you lose your keys, you should stand up and look for them, rather than sitting down and not look for them.

9.9 Factorial summary

We have introduced you to factorial designs, which are simply designs with more than one IV. The special property of factorial designs is that all of the levels of each IV need to be crossed with the other IVs.

We showed you how to analyse a repeated measures 2x2 design with paired samples-tests, and what an ANOVA table would look like if you did this in R. We also went through, by hand, the task of calculating an ANOVA table for a 2x2 between subjects design.

The main point we want you take away is that factorial designs are extremely useful for determining things that cause effects to change. Generally a researcher measures an effect of interest (their IV 1). Then, they want to know what makes that effect get bigger or smaller. They want to exert experimental control over their effect. For example, they might have a theory that says doing X should make the effect bigger, but doing Y should make it smaller. They can test these theories using factorial designs, and manipulating X or Y as a second independent variable.

In a factorial design each IV will have its own main effect. Sometimes the main effect themselves are what the researcher is interested in measures. But more often, it is the interaction effect that is most relevant. The interaction can test whether the effect of IV1 changes between the levels of IV2. When it does, researchers can infer that their second manipulation (IV2) causes change in their effect of interest. These changes are then documented and used to test underlying causal theories about the effects of interest.

Chapter 10

More On Factorial Designs

We are going to do a couple things in this chapter. The most important thing we do is give you more exposure to factorial designs. The second thing we do is show that you can mix it up with ANOVA. You already know that you can have more than one IV. And, you know that research designs can be between-subjects or within-subjects (repeated-measures). When you have more than one IV, they can all be between-subjects variables, they can all be within-subject repeated measures, or they can be a mix: say one between-subject variable and one within-subject variable. You can use ANOVA to analyze all of these kinds of designs. You always get one main effect for each IV, and a number of interactions, or just one, depending on the number of IVs.

10.1 Looking at main effects and interactions

Designs with multiple factors are very common. When you read a research article you will often see graphs that show the results from designs with multiple factors. It would be good for you if you were comfortable interpreting the meaning of those results. The skill here is to be able to look at a graph and see the pattern of main effects and interactions. This skill is important, because the patterns in the data can quickly become very complicated looking, especially when there are more than two independent variables, with more than two levels.

10.1.1 2x2 designs

Let's take the case of 2x2 designs. There will always be the possibility of two main effects and one interaction. You will always be able to compare the means for each main effect and interaction. If the appropriate means are different then there is a main effect or interaction. Here's the thing, there are a bunch of ways all of this can turn out. Check out the ways, there are 8 of them:

1. no IV1 main effect, no IV2 main effect, no interaction
2. IV1 main effect, no IV2 main effect, no interaction
3. IV1 main effect, no IV2 main effect, interaction
4. IV1 main effect, IV2 main effect, no interaction
5. IV1 main effect, IV2 main effect, interaction
6. no IV1 main effect, IV2 main effect, no interaction
7. no IV1 main effect, IV2 main effect, interaction
8. no IV1 main effect, no IV2 main effect, interaction

OK, so if you run a 2x2, any of these 8 general patterns could occur in your data. That's a lot to keep track of isn't. As you develop your skills in examining graphs that plot means, you should be able to look

at the graph and visually guesstimate if there is, or is not, a main effect or interaction. You will need you inferential statistics to tell you for sure, but it is worth knowing how to know see the patterns.

In this section we show you some example patterns so that you can get some practice looking at the patterns. First, in bar graph form. Note, we used the following labels for the graph:

- 1 = there was a main effect for IV1.
- ~ 1 = there was not a main effect for IV1
- 2 = there was a main effect for IV2
- ~ 2 = there was not a main effect of IV2
- 1×2 = there was an interaction ' $\sim 1 \times 2$ ' = there was not an interaction

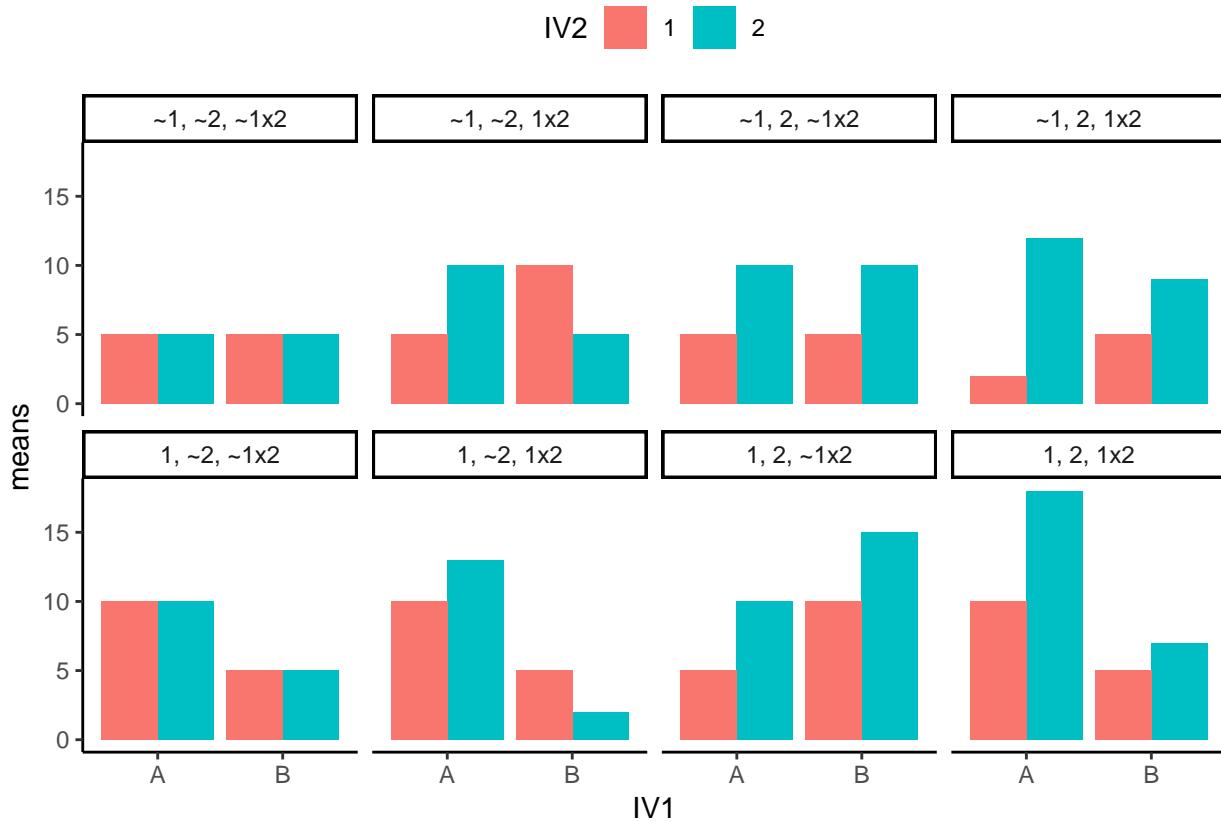


Figure 10.1: 8 Example patterns for means for each of the possible kinds of general outcomes in a 2x2 design

Next, we show you the same thing in line graph form:

You might find the line graphs easier to interpret. Whenever the lines cross, or would cross if they kept going, you have a possibility of an interaction. Whenever the lines are parallel, there can't be an interaction. When both of the points on the A side are higher or lower than both of the points on the B side, then you have a main effect for IV1 (A vs B). Whenever the green line is above or below the red line, then you have a main effect for IV2 (1 vs. 2). We know this is complicated. You should see what all the possibilities look like when we start adding more levels or more IVs. It gets nuts. Because of this nuttiness, it is often good practice to make your research designs simple (as few IVs and levels as possible to test your question). That way it will be easier to interpret your data. Whenever you see that someone ran a 4x3x7x2 design, your head should spin. It's just too complicated.

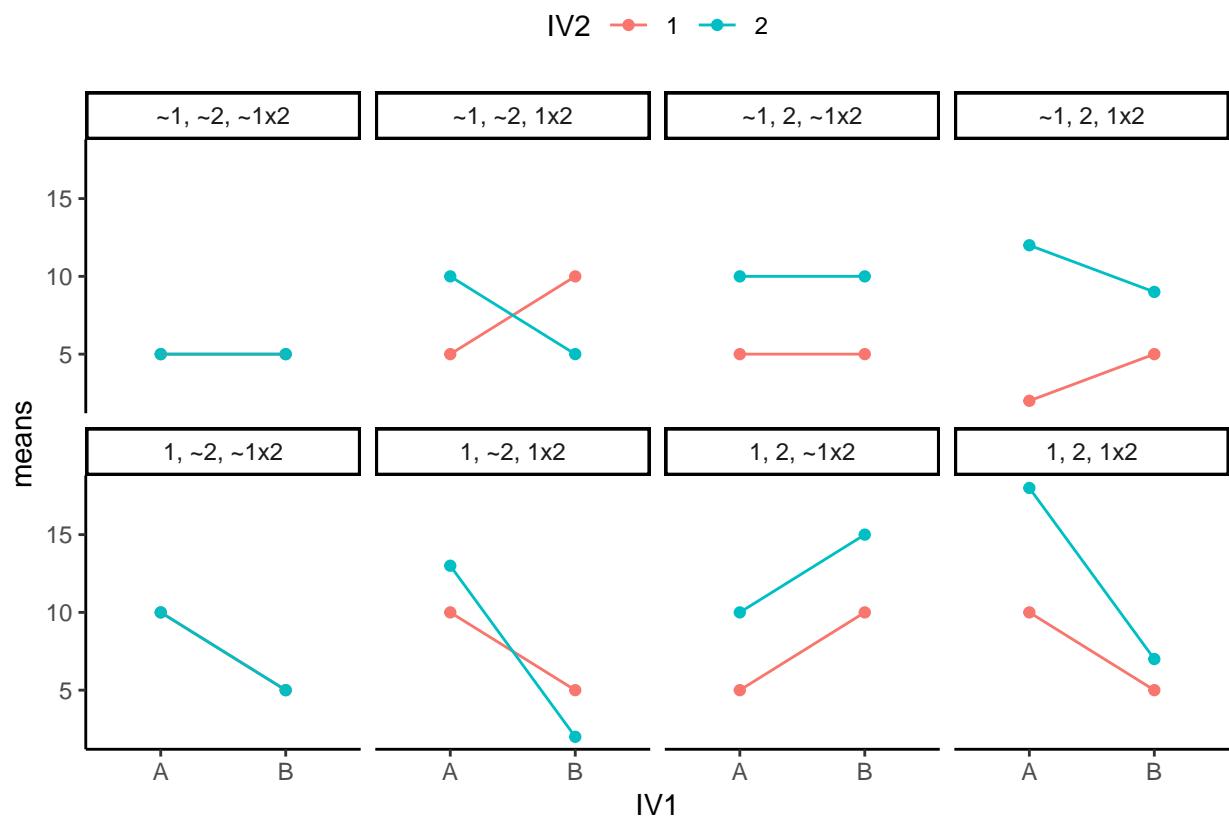


Figure 10.2: Line graphs showing 8 possible general outcomes for a 2x2 design

10.2 Interpreting main effects and interactions

The interpretation of main effects and interactions can get tricky. Consider the concept of a main effect. This is the idea that a particular IV has a consistent effect. For example, drinking 5 cups of coffee makes you more awake compared to not drinking 5 cups of coffee. The main effect of drinking 5 cups of coffee vs not drinking coffee will generally be true across the levels of other IVs in our life. For example, let's say you conducted an experiment testing whether the effect of drinking 5 cups of coffee vs not, changes depending on whether you are in your house or in a car. Perhaps the situation matters? No, probably not so much. You will probably still be more awake in your house, or your car, after having 5 cups of coffee, compared to if you hadn't.

The coffee example is a reasonably good example of a consistent main effect. Another silly kind of example might be the main effect of shoes on your height. For example, if your IV was wearing shoes or not, and your DV was height, then we could expect to find a main effect of wearing shoes on your measurement of height. When you wear shoes, you will become taller compared to when you don't wear shoes. Wearing shoes adds to your total height. In fact, it's hard to imagine how the effect of wearing shoes on your total height would ever interact with other kinds of variables. You will be always be that extra bit taller wearing shoes. Indeed, if there was another manipulation that could cause an interaction that would truly be strange. For example, imagine if the effect of being inside a bodega or outside a bodega interacted with the effect of wearing shoes on your height. That could mean that shoes make you taller when you are outside a bodega, but when you step inside, your shoes make you shorter...but, obviously this is just totally ridiculous. That's correct, it is often ridiculous to expect that one IV will have an influence on the effect of another, especially when there is no good reason.

The summary here is that it is convenient to think of main effects as a consistent influence of one manipulation. However, when an interaction is observed, this messes up the consistency of the main effect. That is the very definition of an interaction. It means that some main effect is **not** behaving consistently across different situations. Indeed, whenever we find an interaction, sometimes we can question whether or not there really is a general consistent effect of some manipulation, or instead whether that effect only happens in specific situations.

For this reason, you will often see that researchers report their findings this way:

"We found a main effect of X, BUT, this main effect was qualified by an interaction between X and Y".

Notice the big **BUT**. Why is it there? The sentence points out that before they talk about the main effect, they need to first talk about the interaction, which is making the main effect behave inconsistently. In other words, the interpretation of the main effect depends on the interaction, the two things have to be thought of together to make sense of them.

Here are two examples to help you make sense of these issues:

10.2.1 A consistent main effect and an interaction

There is a main effect of IV2: the level 1 means (red points and bar) are both lower than the level 2 means (aqua points and bar). There is also an interaction. The size of the difference between the red and aqua points in the A condition (left) is bigger than the size of the difference in the B condition.

How would we interpret this? We could say there WAS a main effect of IV2, BUT it was qualified by an IV1 x IV2 interaction.

What's the qualification? The size of the IV2 effect changed as a function of the levels of IV1. It was big for level A, and small for level B of IV1.

What does the qualification mean for the main effect? Well, first it means the main effect can be changed by the other IV. That's important to know. Does it also mean that the main effect is not a real main effect because there was an interaction? Not really, there is a generally consistent effect of IV2. The

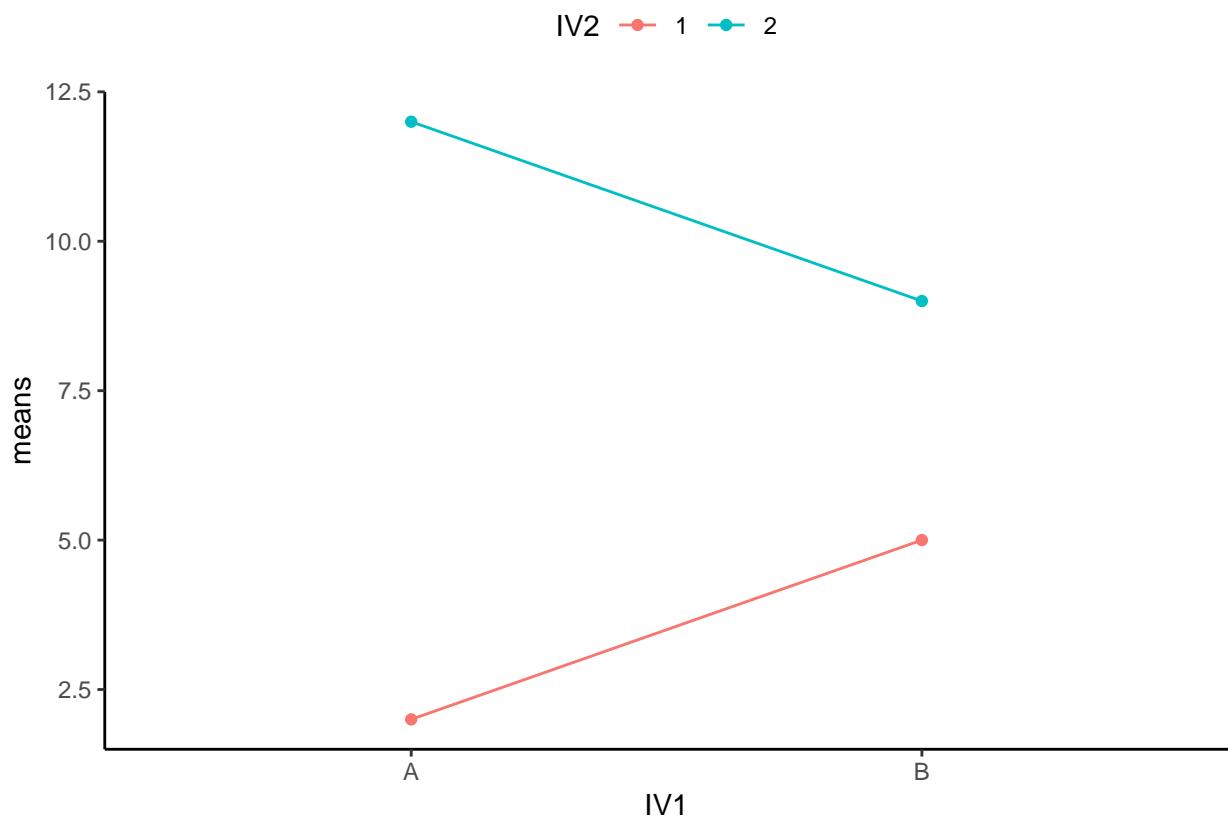


Figure 10.3: Example means showing a generally consistent main effect along with an interaction

green points are above the red points in all cases. Whatever IV2 is doing, it seems to work in at least a couple situations, even if the other IV also causes some change to the influence.

10.2.2 An inconsistent main effect and an interaction

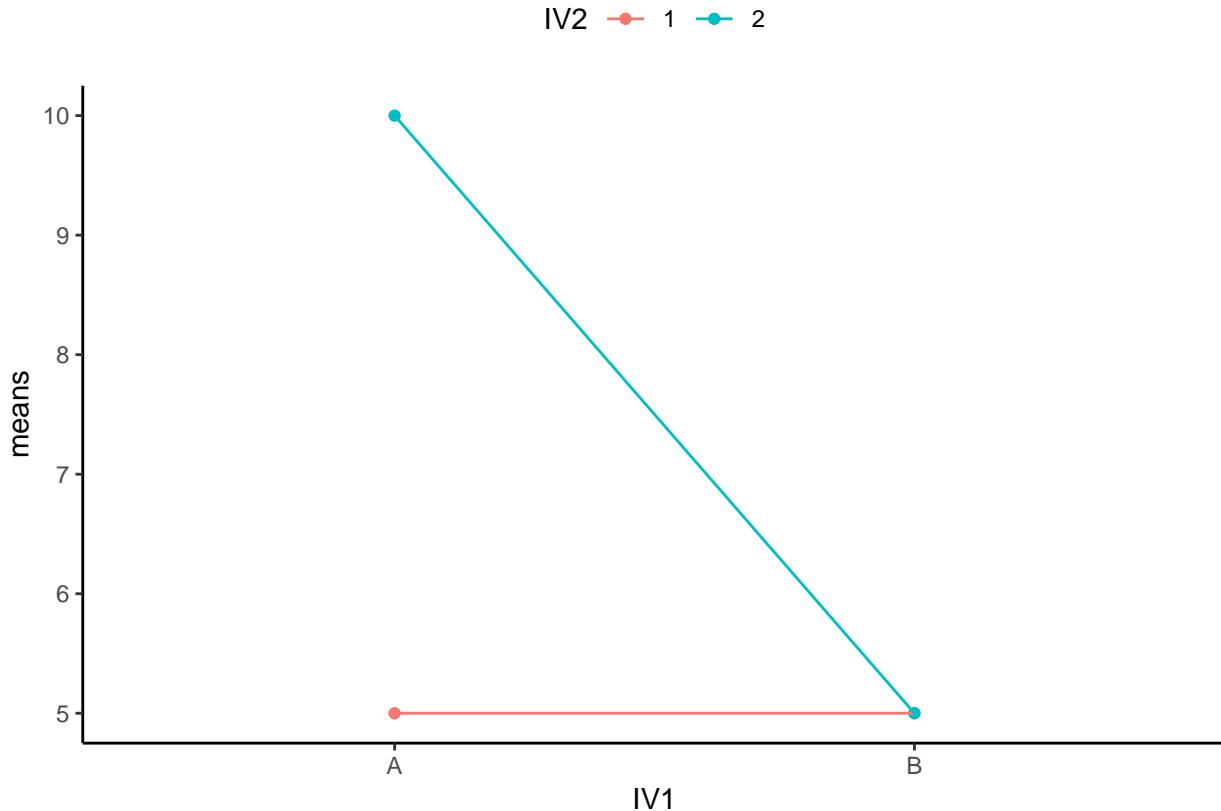


Figure 10.4: Example data showing how an interaction exists, and a main effect does not, even though the means for the main effect may show a difference

This figure shows another 2x2 design. You should see an interaction here straight away. The difference between the aqua and red points in condition A (left two dots) is huge, and there is 0 difference between them in condition B. Is there an interaction? Yes!

Are there any main effects here? With data like this, sometimes an ANOVA will suggest that you do have significant main effects. For example, what is the mean difference between level 1 and 2 of IV2? That is the average of the green points ($(10+5)/2 = 15/2 = 7.5$) compared to the average of the red points (5). There will be a difference of 2.5 for the main effect (7.5 vs. 5).

Starting to see the issue here? From the perspective of the main effect (which collapses over everything and ignores the interaction), there is an overall effect of 2.5. In other words, level 2 adds 2.5 in general compared to level 1. However, we can see from the graph that IV2 does not do anything in general. It does not add 2.5s everywhere. It adds 5 in condition A, and nothing in condition B. It only does one thing in one condition.

What is happening here is that a “main effect” is produced by the process of averaging over a clear interaction.

How would we interpret this? We might have to say there was a main effect of IV2, BUT we would definitely say it was qualified by an IV1 x IV2 interaction.

What's the qualification? The size of the IV2 effect completely changes as a function of the levels of IV1. It was big for level A, and nonexistent for level B of IV1.

What does the qualification mean for the main effect? In this case, we might doubt whether there is a main effect of IV2 at all. It could turn out that IV2 does not have a general influence over the DV all of the time, it may only do something in very specific circumstances, in combination with the presence of other factors.

10.3 Mixed Designs

Throughout this book we keep reminding you that research designs can take different forms. The manipulations can be between-subjects (different subjects in each group), or within-subjects (everybody contributes data in all conditions). If you have more than one manipulation, you can have a mixed design when one of your IVs is between-subjects and one of the other ones is within-subjects.

The only “trick” to these designs is to use the appropriate error terms to construct the F-values for each effect. Effects that have a within-subjects repeated measure (IV) use different error terms than effects that only have a between-subject IV. In principle, you could run an ANOVA with any number of IVs, and any of them good be between or within-subjects variables.

Because this is an introductory textbook, we leave out a full discussion on mixed designs. What we are leaving out are the formulas to construct ANOVA tables that show how to use the correct error terms for each effect. There are many good more advanced textbooks that discuss these issues in much more depth. And, these things can all be Googled. This is a bit of a cop-out on our part, and we may return to fill in this section at some point in the future (or perhaps someone else will add a chapter about this).

In the lab manual, you will learn how to conduct a mixed design ANOVA using software. Generally speaking, the software takes care of the problem of using the correct error terms to construct the ANOVA table.

10.4 More complicated designs

Up until now we have focus on the simplest case for factorial designs, the 2x2 design, with two IVs, each with 2 levels. It is worth spending some time looking at a few more complicated designs and how to interpret them.

10.4.1 2x3 design

In a 2x3 design there are two IVs. IV1 has two levels, and IV2 has three levels. Typically, there would be one DV. Let's talk about the main effects and interaction for this design.

First, let's make the design concrete. Let's imagine we are running a memory experiment. We give people some words to remember, and then test them to see how many they can correctly remember. Our DV is proportion correct. We know that people forget things over time. Our first IV will be time of test, immediate vs. 1 week. The time of test IV will produce a forgetting effect. Generally, people will have a higher proportion correct on an immediate test of their memory for things they just saw, compared to testing a week later.

We might be interested in manipulations that reduce the amount of forgetting that happens over the week. The second IV could be many things. Let's make it the number of time people got to study the items before the memory test, once, twice or three times. We call IV2 the repetition manipulation.

We might expect data that looks like this:

The figure shows some pretend means in all conditions. Let's talk about the main effects and interaction.

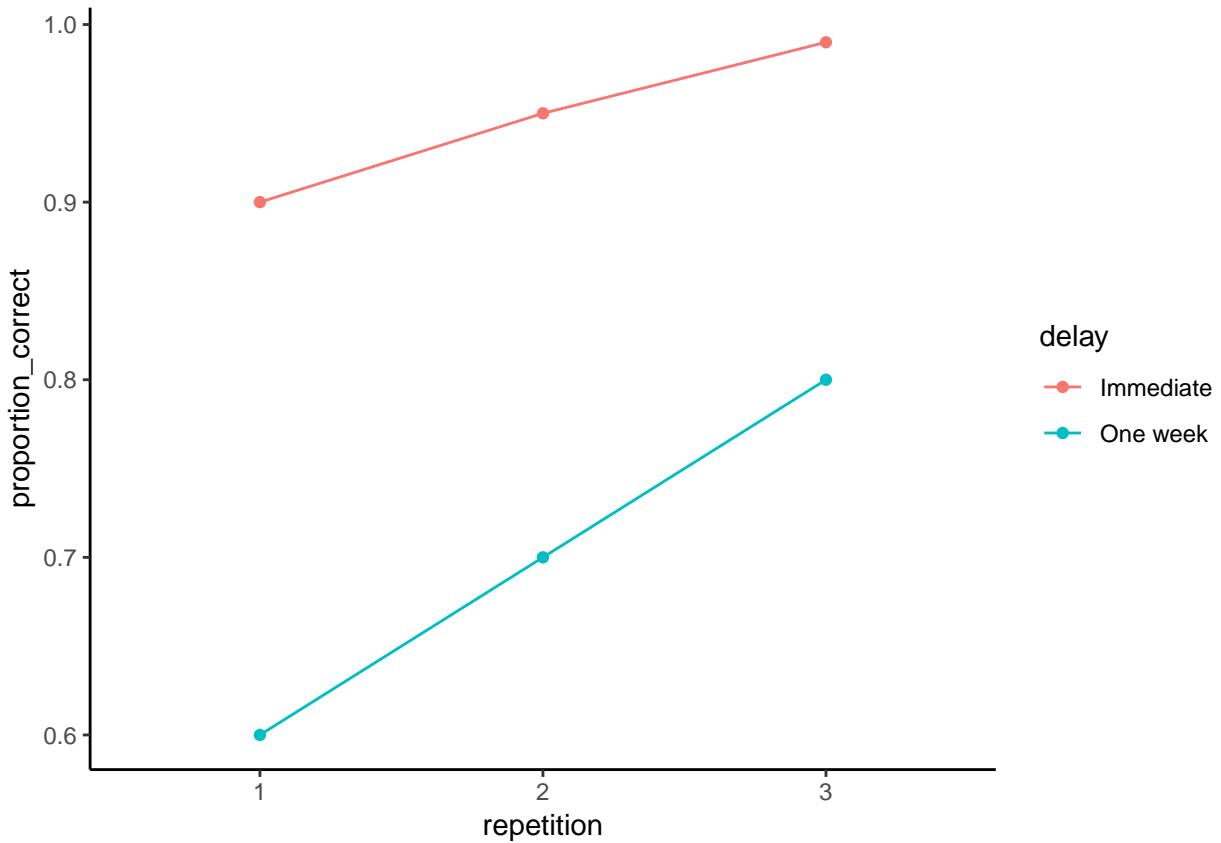


Figure 10.5: Example means for a 2x3 factorial design

First, the main effect of delay (time of test) is very obvious, the red line is way above the aqua line. Proportion correct on the memory test is always higher when the memory test is taken immediately compared to after one week.

Second, the main effect of repetition seems to be clearly present. The more times people saw the items in the memory test (once, twice, or three times), the more they remembered, as measured by increasingly higher proportion correct as a function of number of repetitions.

Is there an interaction? Yes, there is. Remember, an interaction occurs when the effect of one IV depends on the levels of another. The delay IV measures the forgetting effect. Does the size of the forgetting effect change across the levels of the repetition variable? Yes it does. With one repetition the forgetting effect is $.9 - .6 = .4$. With two repetitions, the forgetting effect is a little bit smaller, and with three, the repetition is even smaller still. So, the size of the forgetting effect changes as a function of the levels of the repetition IV. There is evidence in the means for an interaction. You would have to conduct an inferential test on the interaction term to see if these differences were likely or unlikely to be due to sampling error.

If there was no interaction, and say, no main effect of repetition, we would see something like this:

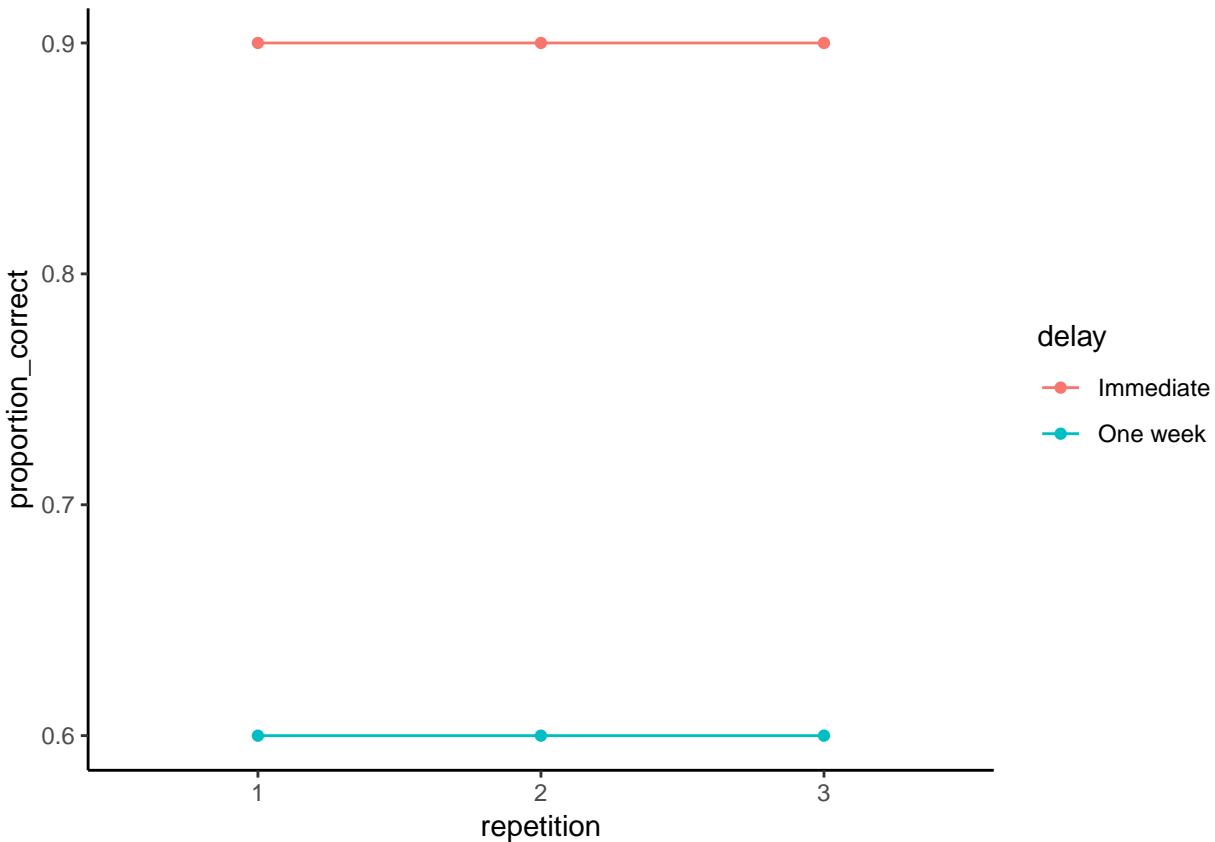


Figure 10.6: Example means for a 2x3 design when there is only one main effect

What would you say about the interaction if you saw something like this:

The correct answer is that there is evidence in the means for an interaction. Remember, we are measuring the forgetting effect (effect of delay) three times. The forgetting effect is the same for repetition condition 1 and 2, but it is much smaller for repetition condition 3. The size of the forgetting effect depends on the levels of the repetition IV, so here again there is an interaction.

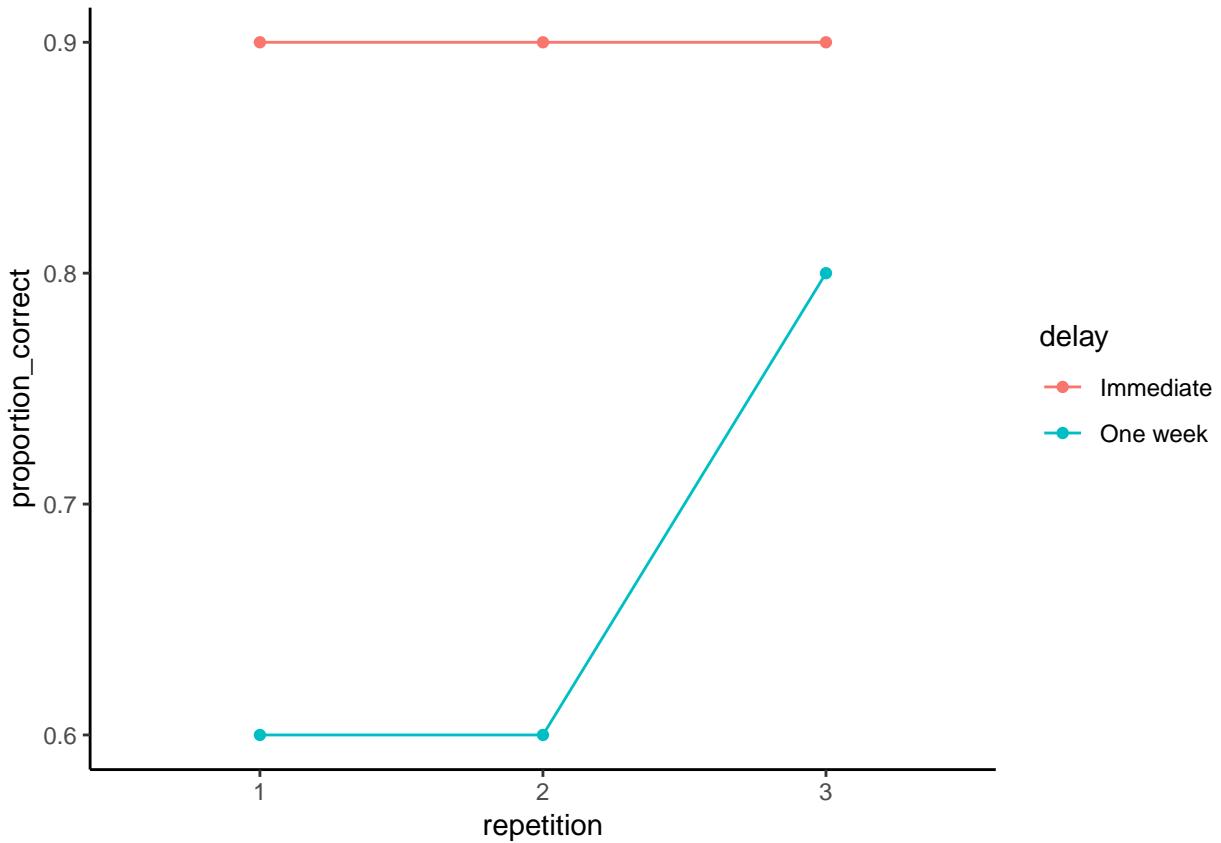


Figure 10.7: Example means for a 2x3 design showing another pattern that produces an interaction

10.4.2 2x2x2 designs

Let's take it up a notch and look at a 2x2x2 design. Here there are three IVs, each with 2 levels each. There are three main effects, three two-way interactions, and one 2-way interaction.

We will use the same example as before but add an additional manipulation of the kind of material that is to be remembered. For example, we could present words during an encoding phase either visually or spoken (auditory) over headphones.

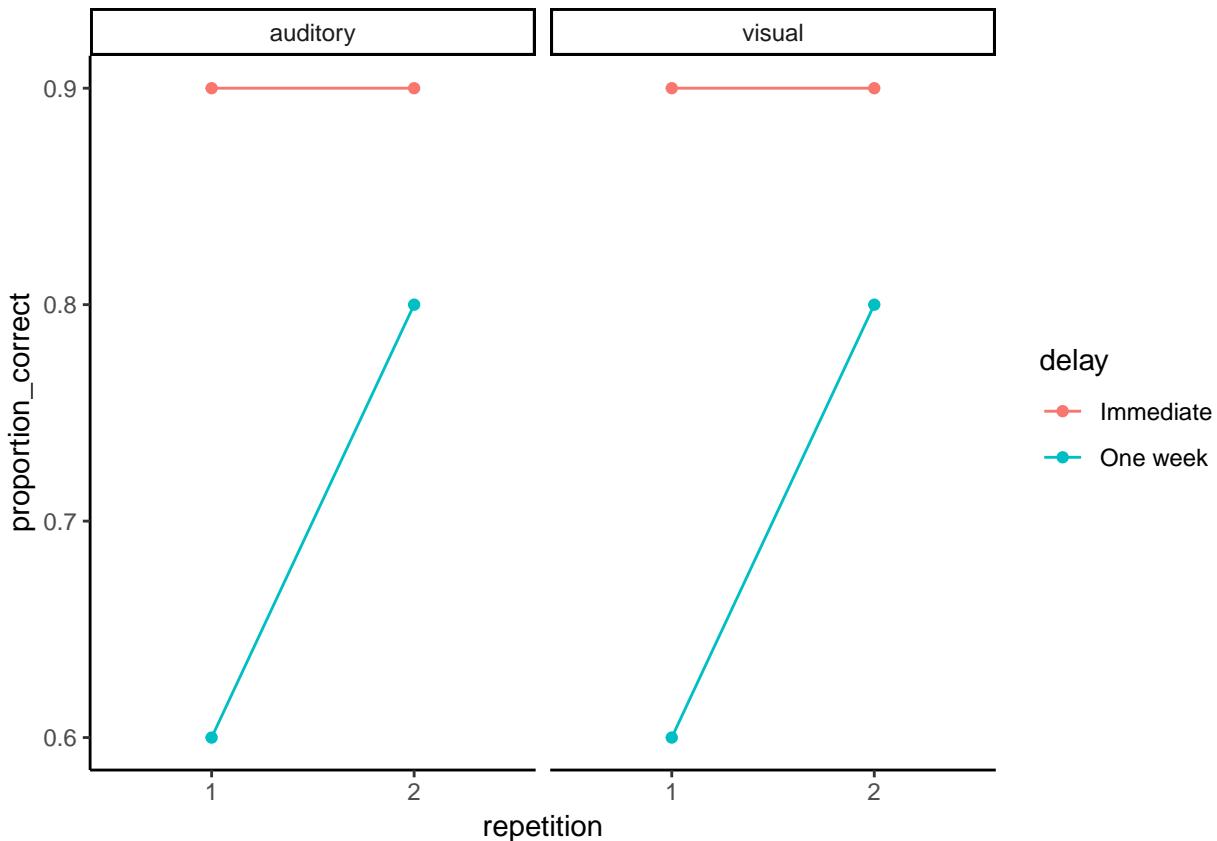


Figure 10.8: Example means from a 2x2x2 design with no three-way interaction

Now we have two panels one for auditory and one for visual. You can think of the 2x2x2, as two 2x2s, one for auditory and one for visual. What's the take home from this example data? We can see that the graphs for auditory and visual are the same. They both show a 2x2 interaction between delay and repetition. People forgot more things across the week when they studied the material once, compared to when they studied the material twice. There is a main effect of delay, there is a main effect of repetition, there is no main effect of modality, and there is no three-way interaction.

What is a three-way interaction anyway? That would occur if there was a difference between the 2x2 interactions. For example, consider the next pattern of results.

We are looking at a 3-way interaction between modality, repetition and delay. What is going on here? These results would be very strange, here is an interpretation.

For auditory stimuli, we see that there is a small forgetting effect when people studied things once, but the forgetting effect gets bigger if they studies things twice. A pattern like this would generally be very strange, usually people would do better if they got to review the material twice.

The visual stimuli show a different pattern. Here, the forgetting effect is large when studying visual things

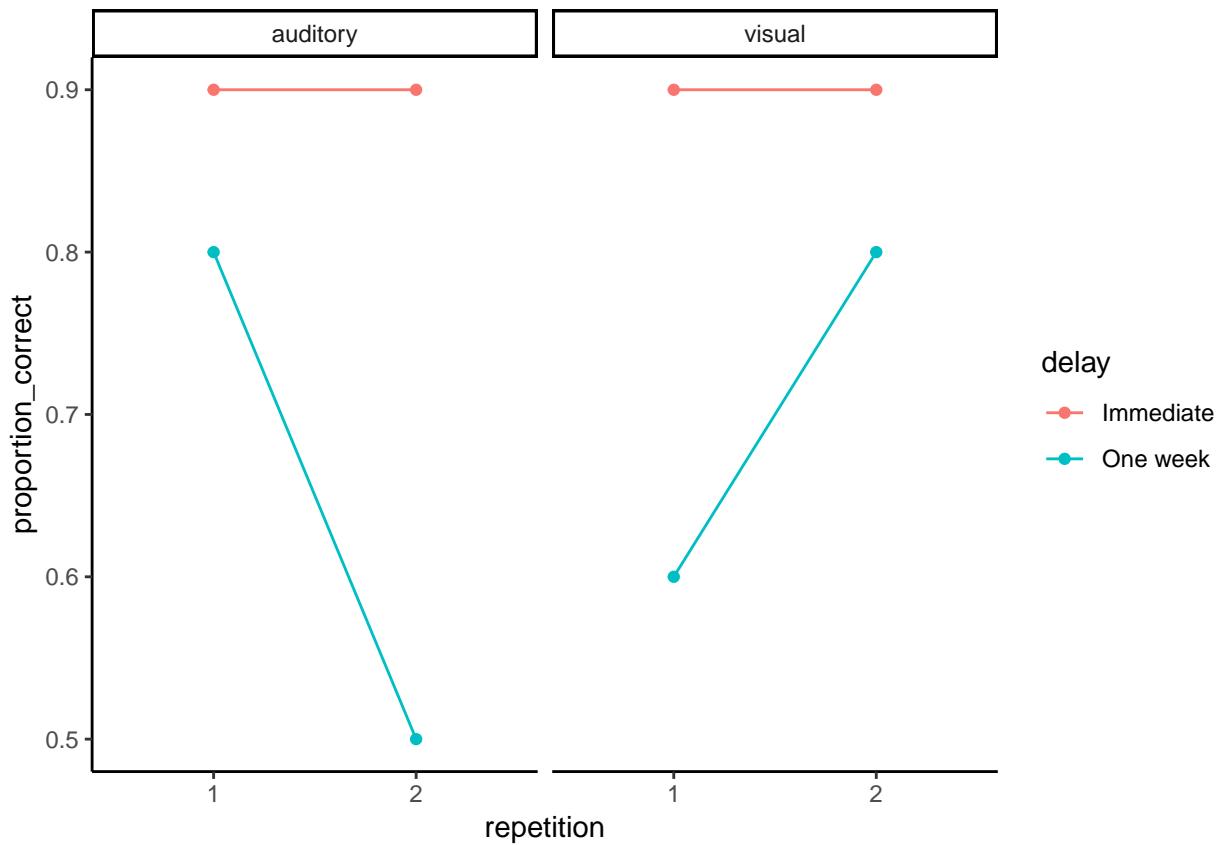


Figure 10.9: Example means from a $2 \times 2 \times 2$ design with a three-way interaction

once, and it gets smaller when studying visual things twice.

We see that there is an interaction between delay (the forgetting effect) and repetition for the auditory stimuli; BUT, this interaction effect is **different** from the interaction effect we see for the visual stimuli. The 2x2 interaction for the auditory stimuli is **different** from the 2x2 interaction for the visual stimuli. In other words, there is an interaction between the two interactions, as a result there is a three-way interaction, called a 2x2x2 interaction.

We will note a general pattern here. Imagine you had a 2x2x2x2 design. That would have a 4-way interaction. What would that mean? It would mean that the pattern of the 2x2x2 interaction changes across the levels of the 4th IV. If two three-way interactions are different, then there is a four-way interaction.

Chapter 11

Simulating Data

You may have noticed that throughout this book so far we have analyzed a lot of fake data. We used R to simulate pretend numbers, and then we analyzed those numbers. We also, from time to time, loaded in some “real” data, and analyzed that. In your labs each week, you have been analyzing a lot of real data. You might be thinking that the simulations we ran were just for educational purposes, to show you how things work. That’s partly true, that’s one reason we ran so many simulations. At the same time, conducting simulations to understand how data behaves is a legitimate branch of statistics. There are some problems out there where we don’t have really good analytic math formulas to tell us the correct answer, so we create and run simulations to approximate the answer.

I’m going to say something mildly controversial right now: If you can’t simulate your data, then you probably don’t really understand your data or how to analyze it. Perhaps, this is too bold of a statement. There are many researchers out there who have never simulated their data, and it might be too much to claim that they don’t really understand their data because they didn’t simulate. Perhaps. There are also many students who have taken statistics classes, and learned how to press some buttons, or copy some code, to analyze some real data; but, who never learned how to run simulations. Perhaps my statement applies more to those students, who I believe would benefit greatly from learning some simulation tricks.

11.1 Reasons to simulate

There are many good reasons to learn simulation techniques, here are some:

1. You force yourself to consider the details of your design, how many subjects, how many conditions, how many observations per condition per subject, and how you will store and represent the data to describe all of these details when you run the experiment
2. You force yourself to consider the kinds of numbers you will be collecting. Specifically, the distributional properties of those numbers. You will have to make decisions about the distributions that you sample from in your simulation, and thinking about this issue helps you better understand your own data when you get it.
3. You learn a bit of computer programming, and this is a very useful general skill that you can build upon to do many things.
4. You can make reasonable and informed assumptions about how your experiment might turn out, and then use the results of your simulation to choose parameters for your design (such as number of subjects, number of observations per condition and subject) that will improve the sensitivity of your design to detect the effects you are interested in measuring.

5. You can even run simulations on the data that you collect to learn more about how it behaves, and to do other kinds of advanced statistics that we don't discuss in this book.
6. You get to improve your intuitions about how data behaves when you measure it. You can test your intuitions by running simulations, and you can learn things you didn't know to begin with. Simulations can be highly informative.
7. When you simulate data in advance of collecting real data, you can work out exactly what kinds of tests you are planning to perform, and you will have already written your analysis code, so it will be ready and waiting for you as soon as you collect the data

OK, so that's just a few reasons why simulations are useful.

11.2 Simulation overview

The basic idea here is actually pretty simple. You make some assumptions about how many subjects will be in your design (set N), you make some assumptions about the distributions that you will be sampling your scores from, then you use R to fabricate fake data according to the parameters you set. Once you build some simulated data, you can conduct a statistical analysis that you would be planning to run on the real data. Then you can see what happens. More importantly, you can repeat the above process many times. This is similar to conducting a replication of your experiment to see if you find the same thing, only you make the computer replicate your simulation 1000s of times. This way you can see how your simulated experiment would turn out over the long run. For example, you might find that the experiment you are planning to run will only produce a "significant" result 25% of the time, that's not very good. Your simulation might also tell you that if you increase your N by say 25, that could really help, and your new experiment with N=25 might succeed 90% of the time. That's information worth knowing.

Before we go into more simulation details, let's just run a quick one. We'll do an independent samples *t*-test. Imagine we have a study with N=10 in each group. There are two groups. We are measuring heart rate. Let's say we know that heart rate is on average 100 beats per minute with a standard deviation of 7. We are going to measure heart rate in condition A where nothing happens, and we are going to measure heart rate in condition B while they watch a scary movie. We think the scary movie might increase heart rate by 5 beats per minute. Let's run a simulation of this:

```
group_A <- rnorm(10,100,7)
group_B <- rnorm(10,105, 7)
t.test(group_A,group_B,var.equal = TRUE)

##
## Two Sample t-test
##
## data: group_A and group_B
## t = -2.7101, df = 18, p-value = 0.01434
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -14.61915 -1.85109
## sample estimates:
## mean of x mean of y
## 95.88966 104.12478
```

We sampled 10 scores from a normal distribution for each group. We changed the mean for group_b to 105, because we were thinking their heart rate would be 5 more than group A. We ran one *t*-test, and we got a result. This result tells us what happens for this one simulation.

We could learn more by repeating the simulation 1000 times, saving the *p*-values from each replication, and then finding out how many of our 1000 simulated experiments give us a significant result:

```

save_ps<-length(1000)
for(i in 1:1000){
  group_A <- rnorm(10,100,7)
  group_B <- rnorm(10,105, 7)
  t_results <- t.test(group_A,group_B,var.equal = TRUE)
  save_ps[i] <- t_results$p.value
}

prop_p<-length(save_ps[save_ps<0.05])/1000
print(prop_p)

## [1] 0.333

```

Now this is more interesting. We found that 33.3% of simulated experiments had a p -value less than 0.05. That's not very good. If you were going to collect data in this kind of experiment, and you made the correct assumptions about the mean and standard deviation of the distribution, and you made the correct assumption about the size of difference between the groups, you would be planning to run an experiment that would not work-out most of the time.

What happens if we increase the number of subject to 50 in each group?

```

save_ps<-length(1000)
for(i in 1:1000){
  group_A <- rnorm(50,100,7)
  group_B <- rnorm(50,105, 7)
  t_results <- t.test(group_A,group_B,var.equal = TRUE)
  save_ps[i] <- t_results$p.value
}

prop_p<-length(save_ps[save_ps<0.05])/1000
print(prop_p)

## [1] 0.953

```

Ooh, look, almost all of the experiments are significant now. So, it would be better to use 50 subjects per group than 10 per group according to this simulation.

Of course, you might already be wondering so many different kinds of things. How can we plausibly know the parameters for the distribution we are sampling from? Isn't this all just guess work? We'll discuss some of these issues as we move forward in this chapter.

IV1	IV2	means	grand_m	IV1_m	IV2_m	IV1xIV2_m
A	1	6	5	-2	3	0
A	2	0	5	-2	-3	0
B	1	10	5	2	3	0
B	2	4	5	2	-3	0

Great, we get to look at the data in two ways. First, look at the table. The four means are listed in the means column. Each number is the sum of the grand mean, the IV1 mean, the IV2 mean, and the in the interaction mean. So, for example, $6 = 5 + (-2) + 3 + 0$, and $0 = 5 + (-2) + (-3) + 0$. The other rows add up to 10 and 4.

We can also work in reverse. What is the grand mean of our scores?

$(6 + 0 + 10 + 4)/4 = 20/4 = 5$, that's the same number as listed in the `grand_m` column.

Next, you can see from the graph that there is a main effect for IV1, both of the points for group A are lower than the points for group B. How big is this main effect? We should be able to figure this out from the table. The table has -2 and +2 for the `IV1_m`. These numbers represent deviations from the grand mean,

caused by the IV1 manipulation. The difference between -2 and 2 is 4, so the size of the main effect should be 4. If this is true, we should find that the difference between the means for Group A and B is also 4.

- Group A mean = $(6+0)/2 = 3$
- Group B mean = $(10+4)/2 = 7$

Well, $7-3 = 4$, which is the difference between the means for IV1. We could do the same thing for IV2, which also shows a main effect. We can see that the main effect must be a difference of 6, that the difference between -3 and 3 in the table.

Finally, notice that the interaction column at the end is all 0s. As a result, there is no interaction. You can see this in the graph. The lines are parallel, no interaction.

11.2.1 What are we doing here?

When you make predictions for how your data might turn out, one way to do it is just to write down what you think the mean would be for each condition. For example, I could have predicted that the means for the above would be 6, 0, 10, and 4, for each of the four conditions.

What we are doing with the GLM approach is making these same predictions, but doing them in terms of the breakdown of the effects we are analyzing for later in the ANOVA. Consider the steps like this:

1. What is your global prediction for all the data, ignoring all of the conditions in your experiment? This is a prediction about the grand mean that you will get. Start your GLM by predicting the grand mean.
2. What is your prediction for the effect of the first IV? This prediction is really that the mean for each group will be different from the grand mean. The grand mean will always be in the center of the data, so your predictions for the main effect of IV1 must deviate from the center. If I want to predict a total difference of 4, then half of th

->

11.3 Simulating t-tests

We've already seen some code for simulating a *t*-test 1000 times, saving the *p*-values, and then calculating the proportion of simulations that are significant ($p < 0.05$). It looked like this:

```
save_ps<-length(1000)
for(i in 1:1000){
  group_A <- rnorm(50,100,7)
  group_B <- rnorm(50,105, 7)
  t_results <- t.test(group_A,group_B,var.equal = TRUE)
  save_ps[i] <- t_results$p.value
}

prop_p<-length(save_ps[save_ps<0.05])/1000
print(prop_p)

## [1] 0.934
```

You could play around with that, and it would be very useful I think. Is there anything else that we can do that would be more useful? Sure there is. With the above simulation, you have to change N or the mean difference each time to see how proportion of significant experiments turns out. It would be nice to look at a graph where we could vary the number of subjects, and the size of the mean difference. That's what the next simulation does. This kind of simulation can make your computer do some hard work depending on how many simulations you run. To make my computer do less work, we will only run 100 simulations for each

parameter. But, what we will do is vary the number of subjects from 10 to 50 (steps of 10), and vary the size of the effect from 0 to 20 in steps of 4.

```

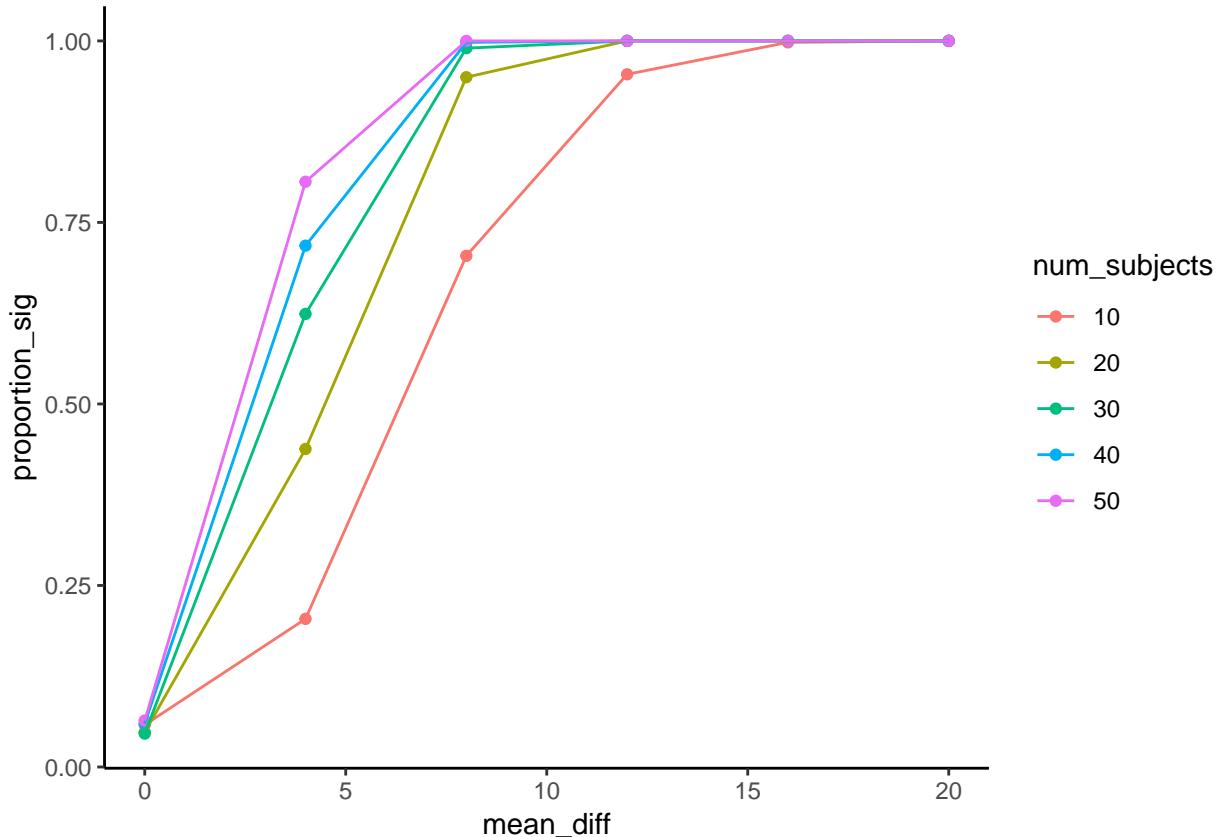
num_sims      <-500
N            <-c(10,20,30,40,50)
mean_difference <-c(0,4,8,12,16,20)
save_ps<-length(num_sims)

all_df<-data.frame()
for(diff in mean_difference){
  for (j in N){
    for(i in 1:num_sims){
      group_A <- rnorm(j,100,7)
      group_B <- rnorm(j,100+diff, 7)
      t_results <- t.test(group_A,group_B,var.equal = TRUE)
      save_ps[i] <- t_results$p.value
    }
    sim_df <-data.frame(save_ps,
                          num_subjects=as.factor(rep(j,num_sims)),
                          mean_diff =rep(diff,num_sims))
    all_df <- rbind(all_df,sim_df)
  }
}

plot_df <- all_df %>%
  dplyr::group_by(num_subjects,mean_diff) %>%
  dplyr::summarise(proportion_sig = length(save_ps[save_ps<0.05])/num_sims)

ggplot(plot_df, aes(x=mean_diff,
                    y=proportion_sig,
                    group=num_subjects,
                    color=num_subjects)+
  geom_point()+
  geom_line()+
  theme_classic()

```



A graph like this is very helpful to look at. Generally, before we run an experiment, we might not have a very good idea of the size of the effect that our manipulation might cause. Will it be a mean difference of 0 (no effect), or 5, or 10, or 20? If you are doing something new, you just might not have a good idea about this. You would know in general that bigger effects are easier to detect. You would be able to detect smaller and smaller effects if you ran more and more subjects. When you run this kind of simulation, you can vary the possible mean differences and the number of the subjects at the same time, and then see what happens.

When the mean difference is 0, we should get an average of 5%, or (0.05 proportion) experiments being significant. This is what we expect by chance, and it doesn't matter how many subjects we run. When there is no difference, we will reject the null 5% of the time (these would all be type 1 errors).

How about when there is a difference of 4? This a pretty small effect. If we only run 10 subjects in each group, we can see that less than 25% of simulated experiments would show significant results. If we wanted a higher chance of success to measure an effect of this size, then we should go up to 40-50 subjects, that would get us around 75% success rates. If that's not good enough for you (25% failures remember, that's still a lot), then re-run the simulation with even more subjects.

Another thing worth pointing out is that if the mean difference is bigger than about 12.5, you can see that all of the designs produce significant outcomes nearly 100% of the time. If you knew this, perhaps you would simply run 10-20 subjects in your experiment, rather than 50. After all, 10-20 is just fine for detecting the effect, and 50 subjects might be a waste of resources (both yours and your participants).

11.4 Simulating one-factor ANOVAs

The following builds simulated data for a one-factor ANOVA, appropriate for a between subjects design. We build the data frame containing a column for the group factor levels, and a column for the DV. Then, we run the ANOVA and print it out.

```
N <- 10
groups <- rep(c("A", "B", "C"), each=10)
DV <- c(rnorm(100, 10, 15),    # means for group A
        rnorm(100, 10, 15),    # means for group B
        rnorm(100, 20, 15)     # means for group C
        )
sim_df<-data.frame(groups,DV)

aov_results <- summary(aov(DV~groups, sim_df))

library(xtable)
knitr::kable(xtable(aov_results))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
groups	2	250.7596	125.3798	0.4670805	0.6272893
Residuals	297	79724.5916	268.4330	NA	NA

In this next example, we simulate the same design 100 times, save the *p*-values, and determine the proportion of significant simulations.

```
N <- 10

save_p<-length(100)
for(i in 1:100){
  groups <- rep(c("A", "B", "C"), each=10)
  DV <- c(rnorm(100, 10, 15),    # means for group A
          rnorm(100, 10, 15),    # means for group B
          rnorm(100, 20, 15)     # means for group C
          )
  sim_df<-data.frame(groups,DV)

  aov_results <- summary(aov(DV~groups, sim_df))
  save_p[i]<-aov_results[[1]]$`Pr(>F)`[1]
}

length(save_p[save_p<0.05])/100

## [1] 0.07
```

11.5 Other resources

OK, It's a tuesday, the summer is almost over. I've spent most of this summer (2018) writing this textbook, because we are using it this Fall 2018. Because I am running out of time, I need to finish this and make sure everything is in place for the course to work. As a result, I am not going to finish this chapter right now. The nice thing about this book, is that I (and other people) can fill things in over time. We have shown a few examples of data-simulation, so that's at least something.

If you want to see more examples, I suggest you check out this chapter:

<https://crumplab.github.io/programmingforpsych/simulating-and-analyzing-data-in-r.html#simulating-data-for-multi-fact>

This section will get longer as I find more resources to add, and hopefully the entire chapter will get longer as I add in more examples over time.

Chapter 12

Thinking about answering questions with data

You might be happy that this is the last chapter (so far) of this textbook. At this point we are in the last weeks of our introductory statistics course. It's called "introductory" for a reason. We have covered far less about statistics than we have covered. There's just too much out there to cover in one short semester. In this chapter we acknowledge some of the things we haven't yet covered, and treat them as things that you should think about. If there is one take home message that we want to get across to you, it's that when you ask questions with data, you should be able to **justify** how you answer those questions.

12.1 Effect-size and power

If you already know something about statistics while you were reading this book, you might have noticed that we neglected to discuss the topic of effect-size, and we barely talked about statistical power. We will talk a little bit about these things here.

First, it is worth pointing out that over the years, at least in Psychology, many societies and journals have made recommendations about how researchers should report their statistical analyses. Among the recommendations is that measures of "effect size" should be reported. Similarly, many journals now require that researchers report an "a priori" power-analysis (the recommendation is this should be done before the data is collected). Because these recommendations are so prevalent, it is worth discussing what these ideas refer to. At the same time, the meaning of effect-size and power somewhat depend on your "philosophical" bent, and these two ideas can become completely meaningless depending on how you think of statistics. For these complicating reasons we have suspended our discussion of the topic until now.

The question or practice of using measures of effect size and conducting power-analyses are also good examples of the more general need to think about what you are doing. If you are going to report effect size, and conduct power analyses, these activities should not be done blindly because someone else recommends that you do them, these activities and other suitable ones should be done as a part of justifying what you are doing. It is a part of thinking about how to make your data answer questions for you.

12.1.1 Chance vs. real effects

Let's rehash something we've said over and over again. First, researchers are interested in whether their manipulation causes a change in their measurement. If it does, they can become confident that they have uncovered a causal force (the manipulation). However, we know that differences in the measure between experimental conditions can arise by chance alone, just by sampling error. In fact, we can create pictures

that show us the window of chance for a given statistic, these tells us roughly the range and likelihoods of getting various differences just by chance. With these windows in hand, we can then determine whether the differences we found in some data that we collected were likely or unlikely to be due to chance. We also learned that sample-size plays a big role in the shape of the chance window. Small samples give chance a large opportunity make big differences. Large samples give chance a small opportunity to make big differences. The general lesson up to this point has been, design an experiment with a large enough sample to detect the effect of interest. If your design isn't well formed, you could easily be measuring noise, and your differences could be caused by sampling error. Generally speaking, this is still a very good lesson: better designs produce better data; and you can't fix a broken design with statistics.

There is clearly another thing that can determine whether or not your differences are due to chance. That is the effect itself. If the manipulation does cause a change, then there is an effect, and that effect is a real one. Effects refer to differences in the measurement between experimental conditions. The thing about effects is that they can be big or small, they have a size.

For example, you can think of a manipulation in terms of the size of its hammer. A strong manipulation is like a jack-hammer: it is loud, it produces a big effect, it creates huge differences. A medium manipulation is like regular hammer: it works, you can hear it, it drives a nail into wood, but it doesn't destroy concrete like a jack-hammer, it produces a reliable effect. A small manipulation is like tapping something with a pencil: it does something, you can barely hear it, and only in a quiet room, it doesn't do a good job of driving a nail into wood, and it does nothing to concrete, it produces tiny, unreliable effects. Finally, a really small effect would be hammering something with a feather, it leaves almost no mark and does nothing that is obviously perceptible to nails or pavement. The lesson is, if you want to break up concrete, use a jack-hammer; or, if you want to measure your effect, make your manipulation stronger (like a jack-hammer) so it produces a bigger difference.

12.1.2 Effect size: concrete vs. abstract notions

Generally speaking, the big concept of effect size, is simply how big the differences are, that's it. However, the bigness or smallness of effects quickly becomes a little bit complicated. On the one hand, the raw difference in the means can be very meaningful. Let's say we are measuring performance on a final exam, and we are testing whether or not a miracle drug can make you do better on the test. Let's say taking the drug makes you do 5% better on the test, compared to not taking the drug. You know what 5% means, that's basically a whole letter grade. Pretty good. An effect-size of 25% would be even better right! Lot's of measures have a concrete quality to them, and we often want to the size of the effect expressed in terms of the original measure.

Let's talk about concrete measures some more. How about learning a musical instrument. Let's say it takes 10,000 hours to become an expert piano, violin, or guitar player. And, let's say you found something online that says that using their method, you will learn the instrument in less time than normal. That is a claim about the effect size of their method. You would want to know how big the effect is right? For example, the effect-size could be 10 hours. That would mean it would take you 9,980 hours to become an expert (that's a whole 10 hours less). If I knew the effect-size was so tiny, I wouldn't bother with their new method. But, if the effect size was say 1,000 hours, that's a pretty big deal, that's 10% less (still doesn't seem like much, but saving 1,000 hours seems like a lot).

Just as often as we have concrete measures that are readily interpretable, Psychology often produces measures that are extremely difficult to interpret. For example, questionnaire measures often have no concrete meaning, and only an abstract statistical meaning. If you wanted to know whether a manipulation caused people to more or less happy, and you used to questionnaire to measure happiness, you might find that people were 50 happy in condition 1, and 60 happy in condition 2, that's a difference of 10 happy units. But how much is 10? Is that a big or small difference? It's not immediately obvious. What is the solution here? A common solution is to provide a standardized measure of the difference, like a z-score. For example, if a difference of 10 reflected a shift of one standard deviation that would be useful to know, and that would be a sizeable

shift. If the difference was only a .1 shift in terms of standard deviation, then the difference of 10 wouldn't be very large. We elaborate on this idea next in describing cohen's d.

12.1.3 Cohen's d

Let's look a few distributions to firm up some ideas about effect-size. In the graph below you will see four panels. The first panel (0) represents the null distribution of no differences. This is the idea that your manipulation (A vs. B) doesn't do anything at all, as a result when you measure scores in conditions A and B, you are effectively sampling scores from the very same overall distribution. The panel shows the distribution as green for condition B, but the red one for condition A is identical and drawn underneath (it's invisible). There is 0 difference between these distributions, so it represent a null effect.

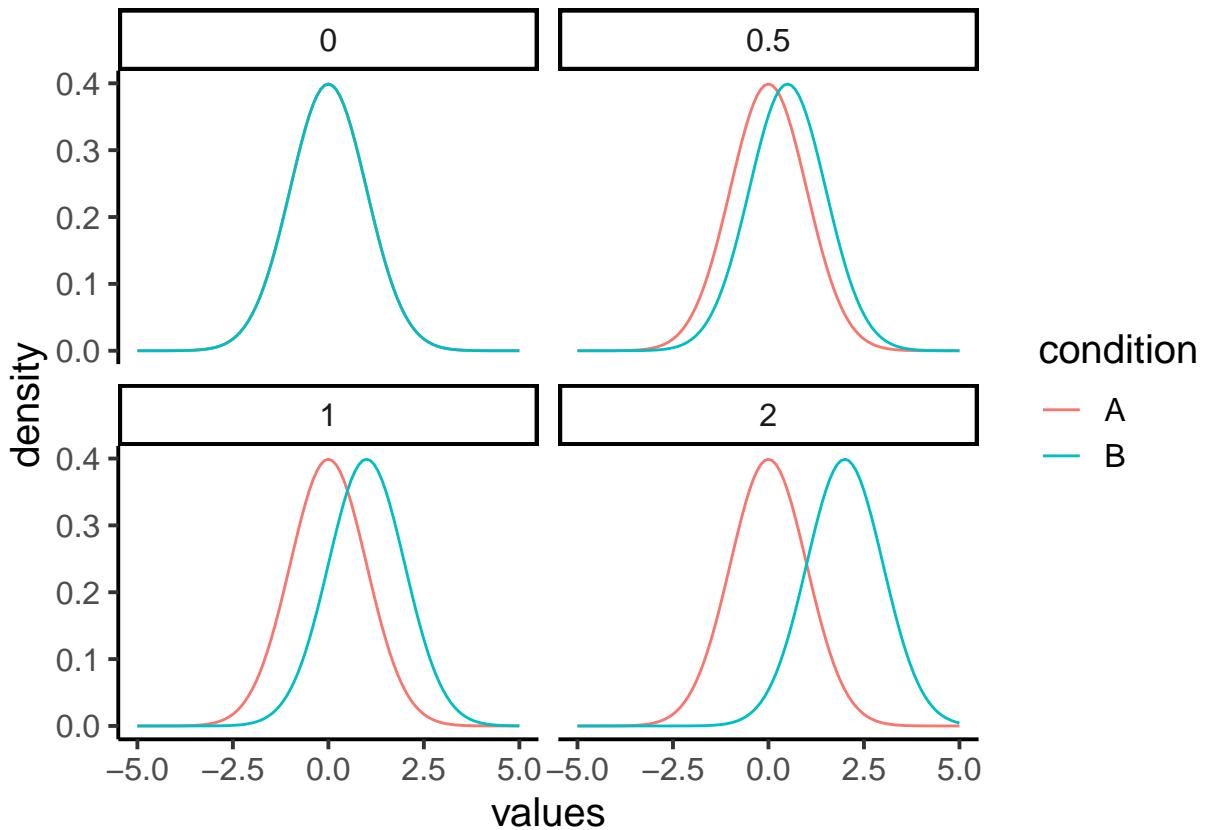


Figure 12.1: Each panel shows hypothetical distributions for two conditions. As the effect-size increases, the difference between the distributions become larger

The remaining panels are hypothetical examples of what a true effect could look like, when your manipulation actually causes a difference. For example, if condition A is a control group, and condition B is a treatment group, we are looking at three cases where the treatment manipulation causes a positive shift in the mean of distribution. We are using normal curves with mean =0 and sd =1 for this demonstration, so a shift of .5 is a shift of half of a standard deviation. A shift of 1 is a shift of 1 standard deviation, and a shift of 2 is a shift of 2 standard deviations. We could draw many more examples showing even bigger shifts, or shifts that go in the other direction.

Let's look at another example, but this time we'll use some concrete measurements. Let's say we are looking at final exam performance, so our numbers are grade percentages. Let's also say that we know the mean on the test is 65%, with a standard deviation of 5%. Group A could be a control that just takes the test,

Group B could receive some “educational” manipulation designed to improve the test score. These graphs then show us some hypotheses about what the manipulation may or may not be doing.

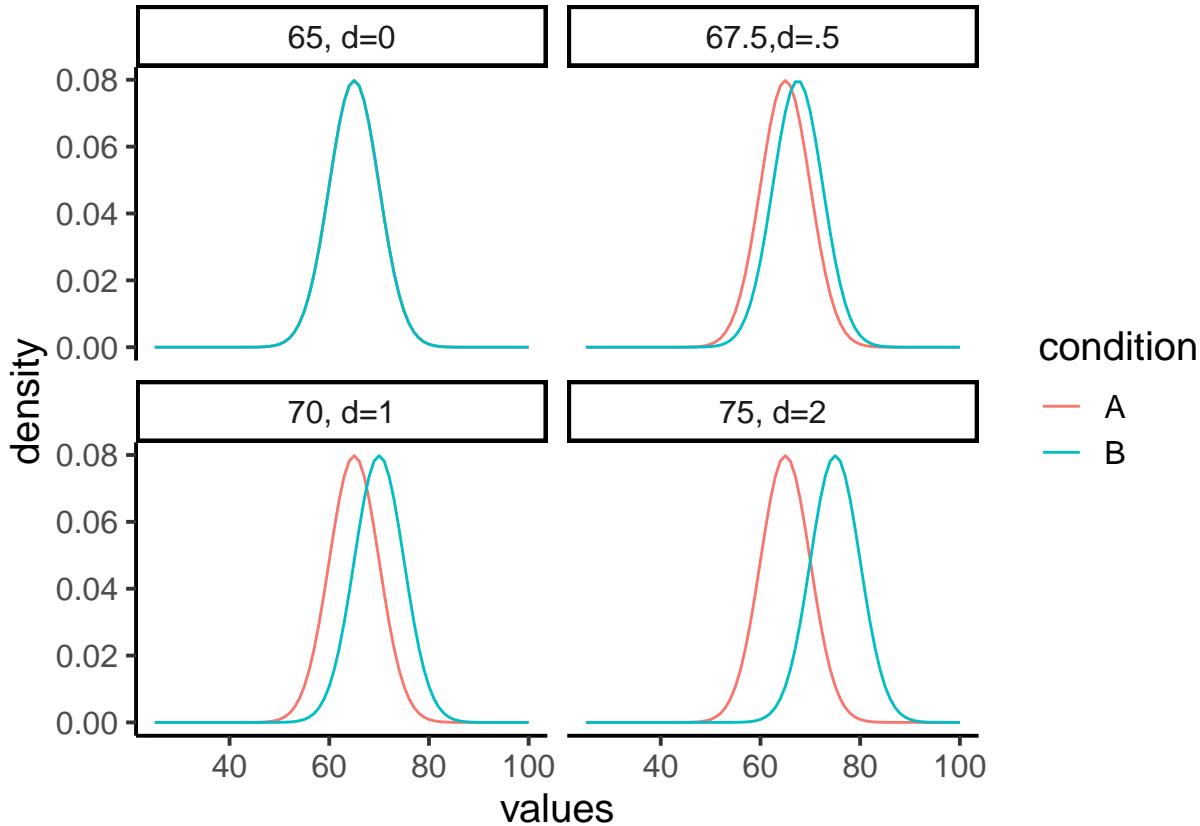


Figure 12.2: Each panel shows hypothetical distributions for two conditions. As the effect-size increases, the difference between the distributions become larger

The first panel shows that both condition A and B will sample test scores from the same distribution (mean =65, with 0 effect). The other panels show shifted mean for condition B (the treatment that is supposed to increase test performance). So, the treatment could increase the test performance by 2.5% (mean 67.5, .5 sd shift), or by 5% (mean 70, 1 sd shift), or by 10% (mean 75%, 2 sd shift), or by any other amount. In terms of our previous metaphor, a shift of 2 standard deviations is more like jack-hammer in terms of size, and a shift of .5 standard deviations is more like using a pencil. The thing about research, is we often have no clue about whether our manipulation will produce a big or small effect, that's why we are conducting the research.

You might have noticed that the letter d appears in the above figure. Why is that? Jacob Cohen (Cohen, 1988) used the letter d in defining the effect-size for this situation, and now everyone calls it Cohen's d . The formula for Cohen's d is:

$$d = \frac{\text{mean for condition 1} - \text{mean for condition 2}}{\text{population standard deviation}}$$

If you notice, this is just a kind of z-score. It is a way to standardize the mean difference in terms of the population standard deviation.

It is also worth noting again that this measure of effect-size is entirely hypothetical for most purposes. In general, researchers do not know the population standard deviation, they can only guess at it, or estimate it from the sample. The same goes for means, in the formula these are hypothetical mean differences in two population distributions. In practice, researchers do not know these values, they guess at them from their samples.

Before discussing why the concept of effect-size can be useful, we note that Cohen's d is useful for understanding abstract measures. For example, when you don't know what a difference of 10 or 20 means as a raw score, you can standardize the difference by the sample standard deviation, then you know roughly how big the effect is in terms of standard units. If you thought a 20 was big, but it turned out to be only 1/10th of a standard deviation, then you would know the effect is actually quite small with respect to the overall variability in the data.

12.2 Power

When there is a true effect out there to measure, you want to make sure your design is sensitive enough to detect the effect, otherwise what's the point. We've already talked about the idea that an effect can have different sizes. The next idea is that your design can be more less sensitive in its ability to reliably measure the effect. We have discussed this general idea many times already in the textbook, for example we know that we will be more likely to detect "significant" effects (when there are real differences) when we increase our sample-size. Here, we will talk about the idea of design sensitivity in terms of the concept of power. Interestingly, the concept of power is a somewhat limited concept, in that it only exists as a concept within some philosophies of statistics.

12.2.1 A digression about hypothesis testing

In particular, the concept of power falls out of the Neyman-Pearson concept of null vs. alternative hypothesis testing. Up to this point, we have largely avoided this terminology. This is perhaps a disservice in that the Neyman-Pearson ideas are by now the most common and widespread, and in the opinion of some of us, they are also the most widely misunderstood and abused idea, which is why we have avoided these ideas until now.

What we have been mainly doing is talking about hypothesis testing from the Fisherian (Sir Ronald Fisher, the ANOVA guy) perspective. This is a basic perspective that we think can't be easily ignored. It is also quite limited. The basic idea is this:

1. We know that chance can cause some differences when we measure something between experimental conditions.
2. We want to rule out the possibility that the difference that we observed can not be due to chance
3. We construct large N designs that permit us to do this when a real effect is observed, such that we can confidently say that big differences that we find are so big (well outside the chance window) that it is highly implausible that chance alone could have produced.
4. The final conclusion is that chance was extremely unlikely to have produced the differences. We then infer that something else, like the manipulation, must have caused the difference.
5. We don't say anything else about the something else.
6. We either reject the null distribution as an explanation (that chance couldn't have done it), or retain the null (admit that chance could have done it, and if it did we couldn't tell the difference between what we found and what chance could do)

Neyman and Pearson introduced one more idea to this mix, the idea of an alternative hypothesis. The alternative hypothesis is the idea that if there is a true effect, then the data sampled into each condition of the experiment must have come from two different distributions. Remember, when there is no effect we assume all of the data came from the same distribution (which by definition can't produce true differences in the long run, because all of the numbers are coming from the same distribution). The graphs of effect-sizes from before show examples of these alternative distributions, with samples for condition A coming from one distribution, and samples from condition B coming from a shifted distribution with a different mean.

So, under the Neyman-Pearson tradition, when a researcher finds a significant effect they do more than one thing. First, they reject the null-hypothesis of no differences, and they accept the alternative hypothesis

that there was differences. This seems like a sensible thing to do. And, because the researcher is actually interested in the properties of the real effect, they might be interested in learning more about the actual alternative hypothesis, that is they might want to know if their data come from two different distributions that were separated by some amount...in other words, they would want to know the size of the effect that they were measuring.

12.2.2 Back to power

We have now discussed enough ideas to formalize the concept of statistical power. For this concept to exist we need to do a couple things.

1. Agree to set an alpha criterion. When the p-value for our test-statistic is below this value we will call our finding statistically significant, and agree to reject the null hypothesis and accept the “alternative” hypothesis (sidenote, usually it isn’t very clear which specific alternative hypothesis was accepted)
2. In advance of conducting the study, figure out what kinds of effect-sizes our design is capable of detecting with particular probabilities.

The power of a study is determined by the relationship between

1. The sample-size of the study
2. The effect-size of the manipulation
3. The alpha value set by the researcher.

To see this in practice let’s do a simulation. We will do a t-test on a between-groups design 10 subjects in each group. Group A will be a control group with scores sampled from a normal distribution with mean of 10, and standard deviation of 5. Group B will be a treatment group, we will say the treatment has an effect-size of Cohen’s $d = .5$, that’s a standard deviation shift of .5, so the scores will come from a normal distribution with mean =12.5 and standard deviation of 5. Remember 1 standard deviation here is 5, so half of a standard deviation is 2.5.

The following R script runs this simulated experiment 1000 times. We set the alpha criterion to .05, this means we will reject the null whenever the p -value is less than .05. With this specific design, how many times out of 1000 do we reject the null, and accept the alternative hypothesis?

```
## [1] 170
```

The answer is that we reject the null, and accept the alternative 170 times out of 1000. In other words our experiment successfully accepts the alternative hypothesis 17 percent of the time, this is known as the power of the study. Power is the probability that a design will successfully detect an effect of a specific size.

Importantly, power is completely abstract idea that is completely determined by many assumptions including N, effect-size, and alpha. As a result, it is best not to think of power as a single number, but instead as a family of numbers.

For example, power is different when we change N. If we increase N, our samples will more precisely estimate the true distributions that they came from. Increasing N reduces sampling error, and shrinks the range of differences that can be produced by chance. Let’s increase our N in this simulation from 10 to 20 in each group and see what happens.

```
## [1] 330
```

Now the number of significant experiments is 330 out of 1000, or a power of 33 percent. That’s roughly doubled from before. We have made the design more sensitive to the effect by increasing N.

We can change the power of the design by changing the alpha-value, which tells us how much evidence we need to reject the null. For example, if we set the alpha criterion to 0.01, then we will be more conservative, only rejecting the null when chance can produce the observed difference 1% of the time. In our example, this will have the effect of reducing power. Let’s keep N at 20, but reduce the alpha to 0.01 and see what happens:

```
## [1] 177
```

Now only 177 out of 1000 experiments are significant, that's 17.7 power.

Finally, the power of the design depends on the actual size of the effect caused by the manipulation. In our example, we hypothesized that the effect caused a shift of .5 standard deviations. What if the effect causes a bigger shift? Say, a shift of 2 standard deviations. Let's keep N= 20, and alpha < .01, but change the effect-size to two standard deviations. When the effect in the real-world is bigger, it should be easier to measure, so our power will increase.

```
## [1] 1000
```

Neat, if the effect-size is actually huge (2 standard deviation shift), then we have power 100 percent to detect the true effect.

12.2.3 Power curves

We mentioned that it is best to think of power as a family of numbers, rather than as a single number. To elaborate on this consider the power curve below. This is the power curve for a specific design: a between groups experiments with two levels, that uses an independent samples t-test to test whether an observed difference is due to chance. Critically, N is set to 10 in each group, and alpha is set to .05

Power (as a proportion, not a percentage) is plotted on the y-axis, and effect-size (Cohen's d) in standard deviation units is plotted on the x-axis.

Power curve for N=10,

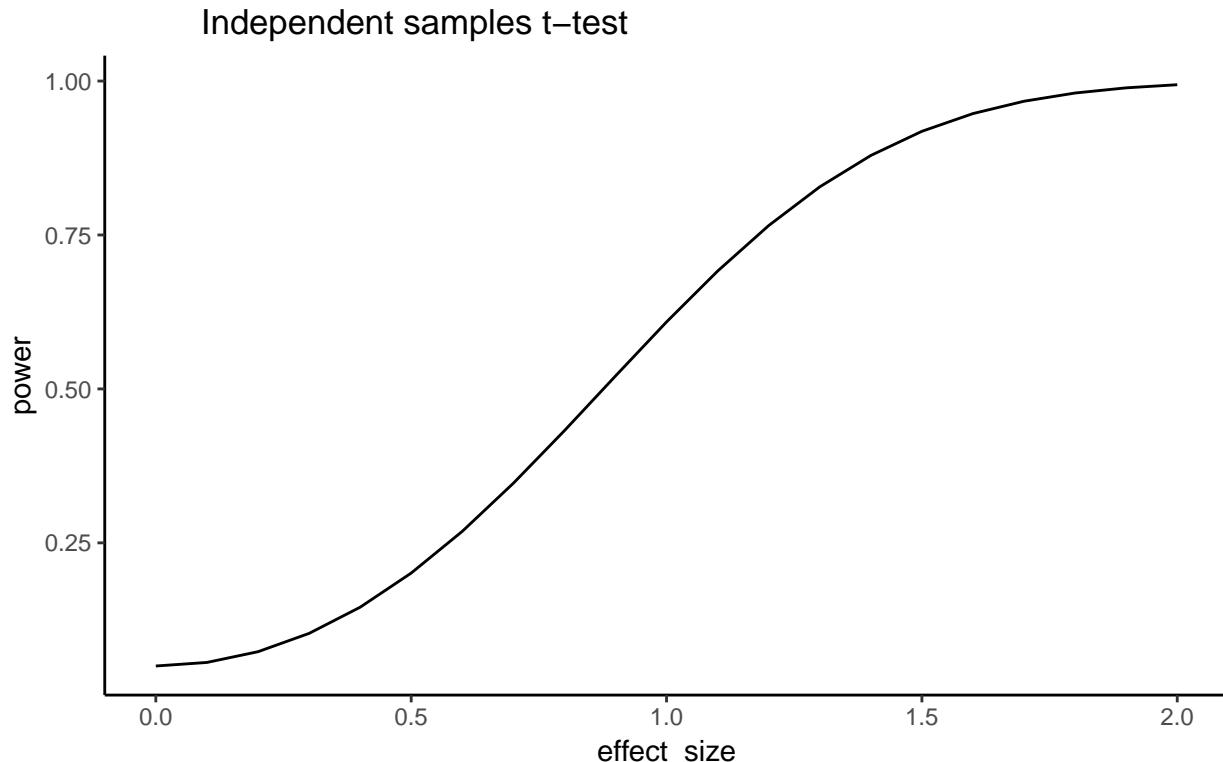


Figure 12.3: This figure shows power as a function of effect-size (Cohen's d) for a between-subjects independent samples t-test, with N=10, and alpha criterion 0.05.

A power curve like this one is very helpful to understand the sensitivity of a particular design. For example, we can see that a between subjects design with $N=10$ in both groups, will detect an effect of $d=.5$ (half a standard deviation shift) about 20% of the time, will detect an effect of $d=.8$ about 50% of the time, and will detect an effect of $d=2$ about 100% of the time. All of the percentages reflect the power of the design, which is the percentage of times the design would be expected to find a $p < 0.05$.

Let's imagine that based on prior research, the effect you are interested in measuring is fairly small, $d=0.2$. If you want to run an experiment that will detect an effect of this size a large percentage of the time, how many subjects do you need to have in each group? We know from the above graph that with $N=10$, power is very low to detect an effect of $d=0.2$. Let's make another graph, but vary the number of subjects rather than the size of the effect.

Power curve for $d=0.2$,

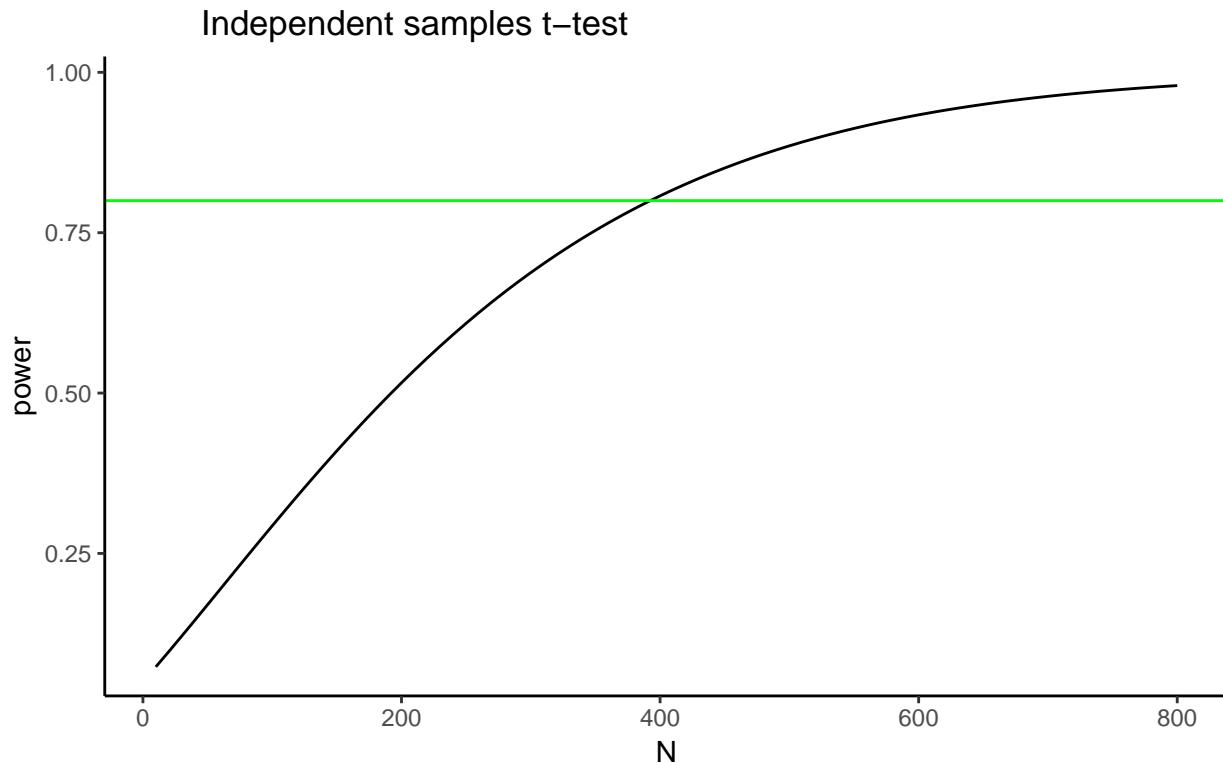


Figure 12.4: This figure shows power as a function of N for a between-subjects independent samples t-test, with $d=0.2$, and alpha criterion 0.05.

The figure plots power to detect an effect of $d=0.2$, as a function of N . The green line shows where power = .8, or 80%. It looks like we would need about 380 subjects in each group to measure an effect of $d=0.2$, with power = .8. This means that 80% of our experiments would successfully show $p < 0.05$. Often times power of 80% is recommended as a reasonable level of power, however even when your design has power = 80%, your experiment will still fail to find an effect (associated with that level of power) 20% of the time!

12.3 Planning your design

Our discussion of effect size and power highlight the importance of the understanding the statistical limitations of an experimental design. In particular, we have seen the relationship between:

1. Sample-size

2. Effect-size
3. Alpha criterion
4. Power

As a general rule of thumb, small N designs can only reliably detect very large effects, whereas large N designs can reliably detect much smaller effects. As a researcher, it is your responsibility to plan your design accordingly so that it is capable of reliably detecting the kinds of effects it is intended to measure.

12.4 Some considerations

12.4.1 Low powered studies

Consider the following case. A researcher runs a study to detect an effect of interest. There is good reason, from prior research, to believe the effect-size is $d=0.5$. The researcher uses a design that has 30% power to detect the effect. They run the experiment and find a significant p-value, ($p<.05$). They conclude their manipulation worked, because it was unlikely that their result could have been caused by chance. How would you interpret the results of a study like this? Would you agree with the researchers that the manipulation likely caused the difference? Would you be skeptical of the result?

The situation above requires thinking about two kinds of probabilities. On the one hand we know that the result observed by the researchers does not occur often by chance (p is less than 0.05). At the same time, we know that the design was underpowered, it only detects results of the expected size 30% of the time. We are faced with wondering what kind of luck was driving the difference. The researchers could have gotten unlucky, and the difference really could be due to chance. In this case, they would be making a type I error (saying the result is real when it isn't). If the result was not due to chance, then they would also be lucky, as their design only detects this effect 30% of the time.

Perhaps another way to look at this situation is in terms of the replicability of the result. Replicability refers to whether or not the findings of the study would be the same if the experiment was repeated. Because we know that power is low here (only 30%), we would expect that most replications of this experiment would not find a significant effect. Instead, the experiment would be expected to replicate only 30% of the time.

12.4.2 Large N and small effects

Perhaps you have noticed that there is an intriguing relationship between N (sample-size) and power and effect-size. As N increases, so does power to detect an effect of a particular size. Additionally, as N increases, a design is capable of detecting smaller and smaller effects with greater and greater power. For example, if N was large enough, we would have high power to detect very small effects, say $d= 0.01$, or even $d=0.001$. Let's think about what this means.

Imagine a drug company told you that they ran an experiment with 1 billion people to test whether their drug causes a significant change in headache pain. Let's say they found a significant effect (with power =100%), but the effect was very small, it turns out the drug reduces headache pain by less than 1%, let's say 0.01%. For our imaginary study we will also assume that this effect is very real, and not caused by chance.

Clearly the design had enough power to detect the effect, and the effect was there, so the design did detect the effect. However, the issue is that there is little practical value to this effect. Nobody is going to buy a drug to reduce their headache pain by 0.01%, even if it was "scientifically proven" to work. This example brings up two issues. First, increasing N to very large levels will allow designs to detect almost any effect (even very tiny ones) with very high power. Second, sometimes effects are meaningless when they are very small, especially in applied research such as drug studies.

These two issues can lead to interesting suggestions. For example, someone might claim that large N studies aren't very useful, because they can always detect really tiny effects that are practically meaningless. On

the other hand, large N studies will also detect larger effects too, and they will give a better estimate of the “true” effect in the population (because we know that larger samples do a better job of estimating population parameters). Additionally, although really small effects are often not interesting in the context of applied research, they can be very important in theoretical research. For example, one theory might predict that manipulating X should have no effect, but another theory might predict that X does have an effect, even if it is a small one. So, detecting a small effect can have theoretical implication that can help rule out false theories. Generally speaking, researchers asking both theoretical and applied questions should think about and establish guidelines for “meaningful” effect-sizes so that they can run designs of appropriate size to detect effects of “meaningful size”.

12.4.3 Small N and Large effects

All other things being equal would you trust the results from a study with small N or large N? This isn’t a trick question, but sometimes people tie themselves into a knot trying to answer it. We already know that large sample-sizes provide better estimates of the distributions the samples come from. As a result, we can safely conclude that we should trust the data from large N studies more than small N studies.

At the same time, you might try to convince yourself otherwise. For example, you know that large N studies can detect very small effects that are practically and possibly even theoretically meaningless. You also know that that small N studies are only capable of reliably detecting very large effects. So, you might reason that a small N study is better than a large N study because if a small N study detects an effect, that effect must be big and meaningful; whereas, a large N study could easily detect an effect that is tiny and meaningless.

This line of thinking needs some improvement. First, just because a large N study can detect small effects, doesn’t mean that it only detects small effects. If the effect is large, a large N study will easily detect it. Large N studies have the power to detect a much wider range of effects, from small to large. Second, just because a small N study detected an effect, does not mean that the effect is real, or that the effect is large. For example, small N studies have more variability, so the estimate of the effect size will have more error. Also, there is 5% (or alpha rate) chance that the effect was spurious. Interestingly, there is a pernicious relationship between effect-size and type I error rate

12.4.4 Type I errors are convincing when N is small

So what is this pernicious relationship between Type I errors and effect-size? Mainly, this relationship is pernicious for small N studies. For example, the following figure illustrates the results of 1000s of simulated experiments, all assuming the null distribution. In other words, for all of these simulations there is no true effect, as the numbers are all sampled from an identical distribution (normal distribution with mean =0, and standard deviation =1). The true effect-size is 0 in all cases.

We know that under the null, researchers will find p values that are less than 5% about 5% of the time, remember that is the definition. So, if a researcher happened to be in this situation (where there manipulation did absolutely nothing), they would make a type I error 5% of the time, or if they conducted 100 experiments, they would expect to find a significant result for 5 of them.

The following graph reports the findings from only the type I errors, where the simulated study did produce $p < 0.05$. For each type I error, we calculated the exact p-value, as well as the effect-size (cohen’s D) (mean difference divided by standard deviation). We already know that the true effect-size is zero, however take a look at this graph, and pay close attention to the smaller sample-sizes.

For example, look at the red dots, when sample size is 10. Here we see that the effect-sizes are quite large. When p is near 0.05 the effect-size is around .8, and it goes up and up as when p gets smaller and smaller. What does this mean? It means that when you get unlucky with a small N design, and your manipulation does not work, but you by chance find a “significant” effect, the effect-size measurement will show you a “big effect”. This is the pernicious aspect. When you make a type I error for small N, your data will make you

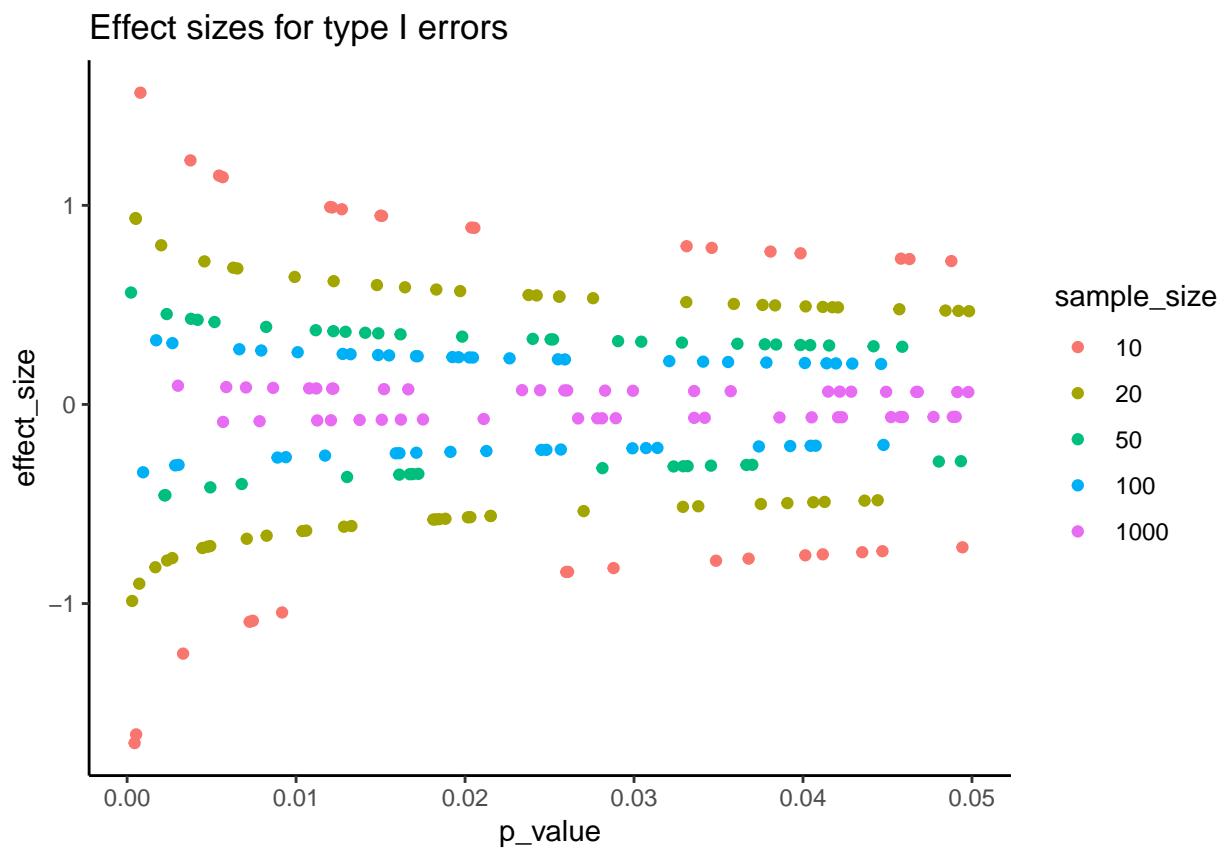


Figure 12.5: Effect size as a function of p-values for type 1 Errors under the null, for a paired samples t-test.

think there is no way it could be a type I error because the effect is just so big!. Notice that when N is very large, like 1000, the measure of effect-size approaches 0 (which is the true effect-size in the simulation).

Bibliography

- Adair, G. (1984). The hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, 69:334–345.
- Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, 27:17–21.
- Behmer, L. P. and Crump, M. J. (2017). Spatial knowledge during skilled action sequencing: Hierarchical versus nonhierarchical representations. *Attention, Perception, & Psychophysics*, 79(8):2435–2448.
- Bickel, P. J., Hammel, E. A., and O’Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187:398–404.
- Campbell, D. T. and Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin, Boston, MA.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum, 2nd edition.
- Evans, J. S. B. T., Barston, J. L., and Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, 11:295–306.
- Fisher, R. A. (1922). On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222:309–368.
- Hothsall, D. (2004). *History of Psychology*. McGraw-Hill.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, 2(8):697–701.
- James, E. L., Bonsall, M. B., Hoppitt, L., Tunbridge, E. M., Geddes, J. R., Milton, A. L., and Holmes, E. A. (2015). Computer game play reduces intrusive memories of experimental trauma via reconsolidation-update mechanisms. *Psychological science*, 26(8):1201–1215.
- Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80:237–251.
- Keynes, J. M. (1923). *A Tract on Monetary Reform*. Macmillan and Company, London.
- Kühberger, A., Fritz, A., and Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *Public Library of Science One*, 9:1–8.
- Matejka, J. and Fitzmaurice, G. (2017). Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1290–1294. ACM.
- Maul, A. (2017). Rethinking Traditional Methods of Survey Validation. *Measurement: Interdisciplinary Research and Perspectives*, 15(2):51–69.
- Meehl, P. H. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34:103–115.
- Mehr, S. A., Song, L. A., and Spelke, E. S. (2016). For 5-month-old infants, melodies are social. *Psychological Science*, 27(4):486–501.

- Pfungst, O. (1911). *Clever Hans (The horse of Mr. von Osten): A contribution to experimental animal and human psychology.* Henry Holt.
- Rosenbaum, D., Mama, Y., and Algom, D. (2017). Stand by your stroop: Standing up enhances selective attention and cognitive control. *Psychological science*, 28(12):1864–1867.
- Rosenthal, R. (1966). *Experimenter effects in behavioral research.* Appleton.
- Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century.* Macmillan.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103:677–680.
- Student, A. (1908). The probable error of a mean. *Biometrika*, 6:1–2.