LEARNING STATISTICS **WITH JASP**

DANIELLE J. NAVARRO
DAVID R. FOXCROFT
THOMAS J. FAULKENBERRY
CLAUDE J. BAJADA

A Tutorial for
Medical Students and
Other Beginners

Learning Statistics with JASP: A Tutorial for Medical Students and Other Beginners

(Version 0.8)

Danielle Navarro
University of New South Wales
d.navarro@unsw.edu.au

David Foxcroft
Oxford Brookes University
david.foxcroft@brookes.ac.uk

Thomas J. Faulkenberry
Tarleton State University
faulkenberry@tarleton.edu

Claude J. Bajada
L' Universita' ta' Malta
claude.bajada@um.edu.mt

October 18, 2023

Overview

Learning Statistics with JASP covers the contents of an introductory statistics class, as typically taught to undergraduate students. The book discusses how to get started in JASP as well as giving an introduction to data manipulation. From a statistical perspective, the book discusses descriptive statistics and graphing first, followed by chapters on probability theory, sampling and estimation, and null hypothesis testing. After introducing the theory, the book covers the analysis of contingency tables, correlation, *t*-tests, regression, ANOVA and factor analysis. Bayesian statistics is covered at the end of the book.

Citation

Navarro, D.J., Foxcroft, D.R., Faulkenberry, T.J. & Bajada, C.J. (2023). Learning Statistics with JASP: A Tutorial for Medical Students and Other Beginners. (Version 0.8).

This book is a revision of the book by Navarro, Foxcroft and Faulkenberry. Each author of the previous editions have their own dedications and I urge you to go back to their versions should you wish to read them.

For this version, I dedicate it to Beni, my son.

Table of Contents

Preface to Version 0.8

Work in Progress

Claude J. Bajada
October 18, 2023

Preface to Version $1/\sqrt{2}$

I am happy to introduce "Learning Statistics with JASP", an adaptation of the excellent "Learning statistics with jamovi" and "Learning Statistics with R". This version builds on the wonderful previous work of Dani Navarro and David Foxcroft, without whose previous efforts a book of this quality would not be possible. I had a simple aim when I began working on this adaption: I wanted to use the Navarro and Foxcroft text in my own statistics courses, but for reasons I won't get into right now, I use JASP instead of jamovi. Both are wonderful tools, but I have a not-so-slight tendency to prefer JASP, possibly because I was using JASP before jamovi split off as a separate project. Nonetheless, I am happy to help bring this book into the world for JASP users.

I am grateful to the Center for Instructional Innovation at Tarleton State University, who gave me a grant to pursue the writing of this open educational resource (OER) in Summer 2019. I am looking forward to providing my future students (and students everywhere) with a quality statistics text that is (and forever shall be) 100% free!

I invite readers everywhere to find ways to make this text better (including identiying the ever-present typos). Please send me an email if you'd like to contribute (or feel free to go to my Github page and just fork it yourself. Go crazy!

Thomas J. Faulkenberry
July 12, 2019

Preface to Version 0.70

This update from version 0.65 introduces some new analyses. In the ANOVA chapters we have added sections on repeated measures ANOVA and analysis of covariance (ANCOVA). In a new chapter we have introduced Factor Analysis and related techniques. Hopefully the style of this new material is consistent with the rest of the book, though eagle-eyed readers might spot a bit more of an emphasis on conceptual and practical explanations, and a bit less algebra. I'm not sure this is a good thing, and

might add the algebra in a bit later. But it reflects both my approach to understanding and teaching statistics, and also some feedback I have received from students on a course I teach. In line with this, I have also been through the rest of the book and tried to separate out some of the algebra by putting it into a box or frame. It's not that this stuff is not important or useful, but for some students they may wish to skip over it and therefore the boxing of these parts should help some readers.

With this version I am very grateful to comments and feedback received from my students and colleagues, notably Wakefield Morys-Carter, and also to numerous people all over the world who have sent in small suggestions and corrections - much appreciated, and keep them coming! One pretty neat new feature is that the example data files for the book can now be loaded into jamovi as an add-on module - thanks to Jonathon Love for helping with that.

David Foxcroft
February 1st, 2019

Preface to Version 0.65

In this adaptation of the excellent 'Learning statistics with R', by Danielle Navarro, we have replaced the statistical software used for the analyses and examples with jamovi. Although R is a powerful statistical programming language, it is not the first choice for every instructor and student at the beginning of their statistical learning. Some instructors and students tend to prefer the point-and-click style of software, and that's where jamovi comes in. jamovi is software that aims to simplify two aspects of using R. It offers a point-and-click graphical user interface (GUI), and it also provides functions that combine the capabilities of many others, bringing a more SPSS- or SAS-like method of programming to R. Importantly, jamovi will always be free and open - that's one of its core values - because jamovi is made by the scientific community, for the scientific community.

With this version I am very grateful for the help of others who have read through drafts and provided excellent suggestions and corrections, particularly Dr David Emery and Kirsty Walter.

David Foxcroft
July 1st, 2018

Preface to Version 0.6

The book hasn't changed much since 2015 when I released Version 0.5 – it's probably fair to say that I've changed more than it has. I moved from Adelaide to Sydney in 2016 and my teaching profile at

UNSW is different to what it was at Adelaide, and I haven't really had a chance to work on it since arriving here! It's a little strange looking back at this actually. A few quick comments...

- Weirdly, the book *consistently* misgenders me, but I suppose I have only myself to blame for that one :-) There's now a brief footnote on page 12 that mentions this issue; in real life I've been working through a gender affirmation process for the last two years and mostly go by she/her pronouns. I am, however, just as lazy as I ever was so I haven't bothered updating the text in the book.
- For Version 0.6 I haven't changed much I've made a few minor changes when people have pointed out typos or other errors. In particular it's worth noting the issue associated with the etaSquared function in the **lsr** package (which isn't really being maintained any more) in Section 14.4. The function works fine for the simple examples in the book, but there are definitely bugs in there that I haven't found time to check! So please take care with that one.
- The biggest change really is the licensing! I've released it under a Creative Commons licence (CC BY-SA 4.0, specifically), and placed all the source files to the associated GitHub repository, if anyone wants to adapt it.

Maybe someone would like to write a version that makes use of the **tidyverse**... I hear that's become rather important to R these days :-)

Best,
Danielle Navarro

Preface to Version 0.5

Another year, another update. This time around, the update has focused almost entirely on the theory sections of the book. Chapters 9, 10 and 11 have been rewritten, hopefully for the better. Along the same lines, Chapter 17 is entirely new, and focuses on Bayesian statistics. I think the changes have improved the book a great deal. I've always felt uncomfortable about the fact that all the inferential statistics in the book are presented from an orthodox perspective, even though I almost always present Bayesian data analyses in my own work. Now that I've managed to squeeze Bayesian methods into the book somewhere, I'm starting to feel better about the book as a whole. I wanted to get a few other things done in this update, but as usual I'm running into teaching deadlines, so the update has to go out the way it is!

Dan Navarro

February 16, 2015

Preface to Version 0.4

A year has gone by since I wrote the last preface. The book has changed in a few important ways: Chapters 3 and 4 do a better job of documenting some of the time saving features of Rstudio, Chapters 12 and 13 now make use of new functions in the lsr package for running chi-square tests and t tests, and the discussion of correlations has been adapted to refer to the new functions in the lsr package. The soft copy of 0.4 now has better internal referencing (i.e., actual hyperlinks between sections), though that was introduced in 0.3.1. There's a few tweaks here and there, and many typo corrections (thank you to everyone who pointed out typos!), but overall 0.4 isn't massively different from 0.3.

I wish I'd had more time over the last 12 months to add more content. The absence of any discussion of repeated measures ANOVA and mixed models more generally really does annoy me. My excuse for this lack of progress is that my second child was born at the start of 2013, and so I spent most of last year just trying to keep my head above water. As a consequence, unpaid side projects like this book got sidelined in favour of things that actually pay my salary! Things are a little calmer now, so with any luck version 0.5 will be a bigger step forward.

One thing that has surprised me is the number of downloads the book gets. I finally got some basic tracking information from the website a couple of months ago, and (after excluding obvious robots) the book has been averaging about 90 downloads per day. That's encouraging: there's at least a few people who find the book useful!

Dan Navarro
February 4, 2014

Preface to Version 0.3

There's a part of me that really doesn't want to publish this book. It's not finished.

And when I say that, I mean it. The referencing is spotty at best, the chapter summaries are just lists of section titles, there's no index, there are no exercises for the reader, the organisation is suboptimal, and the coverage of topics is just not comprehensive enough for my liking. Additionally, there are sections with content that I'm not happy with, figures that really need to be redrawn, and I've had almost no time to hunt down inconsistencies, typos, or errors. In other words, *this book is not finished*. If I didn't have a looming teaching deadline and a baby due in a few weeks, I really wouldn't be making this available at all.

What this means is that if you are an academic looking for teaching materials, a Ph.D. student

looking to learn R, or just a member of the general public interested in statistics, I would advise you to be cautious. What you're looking at is a first draft, and it may not serve your purposes. If we were living in the days when publishing was expensive and the internet wasn't around, I would never consider releasing a book in this form. The thought of someone shelling out $80 for this (which is what a commercial publisher told me it would retail for when they offered to distribute it) makes me feel more than a little uncomfortable. However, it's the 21st century, so I can post the pdf on my website for free, and I can distribute hard copies via a print-on-demand service for less than half what a textbook publisher would charge. And so my guilt is assuaged, and I'm willing to share! With that in mind, you can obtain free soft copies and cheap hard copies online, from the following webpages:

Soft copy: http://www.compcogscisydney.com/learning-statistics-with-r.html
Hard copy: www.lulu.com/content/13570633

Even so, the warning still stands: what you are looking at is Version 0.3 of a work in progress. If and when it hits Version 1.0, I would be willing to stand behind the work and say, yes, this is a textbook that I would encourage other people to use. At that point, I'll probably start shamelessly flogging the thing on the internet and generally acting like a tool. But until that day comes, I'd like it to be made clear that I'm really ambivalent about the work as it stands.

All of the above being said, there is one group of people that I can enthusiastically endorse this book to: the psychology students taking our undergraduate research methods classes (DRIP and DRIP:A) in 2013. For you, this book is ideal, because it was written to accompany your stats lectures. If a problem arises due to a shortcoming of these notes, I can and will adapt content on the fly to fix that problem. Effectively, you've got a textbook written specifically for your classes, distributed for free (electronic copy) or at near-cost prices (hard copy). Better yet, the notes have been tested: Version 0.1 of these notes was used in the 2011 class, Version 0.2 was used in the 2012 class, and now you're looking at the new and improved Version 0.3. I'[for a historical summary]m not saying these notes are titanium plated awesomeness on a stick – though if *you* wanted to say so on the student evaluation forms, then you're totally welcome to – because they're not. But I am saying that they've been tried out in previous years and they seem to work okay. Besides, there's a group of us around to troubleshoot if any problems come up, and you can guarantee that at least *one* of your lecturers has read the whole thing cover to cover!

Okay, with all that out of the way, I should say something about what the book aims to be. At its core, it is an introductory statistics textbook pitched primarily at psychology students. As such, it covers the standard topics that you'd expect of such a book: study design, descriptive statistics, the theory of hypothesis testing, t-tests, χ^2 tests, ANOVA and regression. However, there are also several chapters devoted to the R statistical package, including a chapter on data manipulation and another one on scripts and programming. Moreover, when you look at the content presented in the book, you'll notice a lot of topics that are traditionally swept under the carpet when teaching statistics to psychology students. The Bayesian/frequentist divide is openly disussed in the probability chapter, and the disagreement between Neyman and Fisher about hypothesis testing makes an appearance. The difference between probability and density is discussed. A detailed treatment of Type I, II and III sums of squares for unbalanced factorial ANOVA is provided. And if you have a look in the Epilogue,

it should be clear that my intention is to add a lot more advanced content.

My reasons for pursuing this approach are pretty simple: the students can handle it, and they even seem to enjoy it. Over the last few years I've been pleasantly surprised at just how little difficulty I've had in getting undergraduate psych students to learn R. It's certainly not easy for them, and I've found I need to be a little charitable in setting marking standards, but they do eventually get there. Similarly, they don't seem to have a lot of problems tolerating ambiguity and complexity in presentation of statistical ideas, as long as they are assured that the assessment standards will be set in a fashion that is appropriate for them. So if the students can handle it, why *not* teach it? The potential gains are pretty enticing. If they learn R, the students get access to CRAN, which is perhaps the largest and most comprehensive library of statistical tools in existence. And if they learn about probability theory in detail, it's easier for them to switch from orthodox null hypothesis testing to Bayesian methods if they want to. Better yet, they learn data analysis skills that they can take to an employer without being dependent on expensive and proprietary software.

Sadly, this book isn't the silver bullet that makes all this possible. It's a work in progress, and maybe when it is finished it will be a useful tool. One among many, I would think. There are a number of other books that try to provide a basic introduction to statistics using R, and I'm not arrogant enough to believe that mine is better. Still, I rather like the book, and maybe other people will find it useful, incomplete though it is.

Dan Navarro
January 13, 2013

Part I.

Background

1. Why do we learn statistics?

"Thou shalt not answer questionnaires
Or quizzes upon World Affairs,
Nor with compliance
Take any test. Thou shalt not sit
With statisticians nor commit
A social science"

– W.H. Auden[1]

1.1 The Role of Statistics in Medical Education

For many medical students, it comes as a surprise that statistics plays a significant role in their training. It's safe to say that statistics is rarely the most favored subject in the medical curriculum. If you were genuinely passionate about statistics, chances are you'd be enrolled in a dedicated statistics course rather than pursuing a medical degree. Given this backdrop, it seems pertinent to start by addressing some common questions students have about the relevance of statistics in medicine.

One of the primary concerns is understanding what statistics is, its purpose, and why it's so integral to medical research. Scientists, particularly in the medical field, seem to rely heavily on statistical analysis. So much so that the rationale for it often goes unexplained. For many, the belief is almost axiomatic: your findings aren't credible until they've undergone statistical scrutiny. This might lead medical students to wonder:

Why is there a need for statistics? Why not rely on clinical judgment alone?

[1] The quote comes from Auden's 1946 poem *Under Which Lyre: A Reactionary Tract for the Times*, delivered as part of a commencement address at Harvard University. The history of the poem is kind of interesting: http://harvardmagazine.com/2007/11/a-poets-warning.html

Though this might appear to be a simplistic question, it's a critical one to address. Many reasons can be cited, but perhaps the most straightforward one is that human judgment is fallible. We are all prone to biases, temptations, and errors that could significantly influence medical outcomes. Relying solely on "clinical judgment" to evaluate medical evidence would involve depending solely on intuition, anecdotal experiences, or the sheer reasoning power of the human mind. In the medical field, this is considered an unreliable approach for decision-making.

Exploring this issue further, we can question the trustworthiness of relying solely on "clinical judgment." While medical terminologies are framed in a specific language, language itself has its biases. Some concepts are more challenging to articulate, not necessarily because they are incorrect but because they are complex (e.g., the intricacies of cellular biology or the pharmacokinetics of a drug). Furthermore, our intuitive "gut feelings" are not designed to tackle complex medical issues; they are adapted for day-to-day problem-solving in a world that is rapidly evolving. At the core, making sound judgments requires the use of "induction," where one must extrapolate from available evidence to form general conclusions. If you believe you can make such extrapolations without being swayed by biases or inaccuracies, then that's a risky proposition in the medical setting, where lives are often at stake.

While statistics may not be everyone's favorite subject, it serves as an essential tool to counteract the limitations of human judgment and intuition. It brings an additional layer of rigor and objectivity, helping us make more informed and reliable decisions in medical practice.

1.1.1 The Pitfall of Cognitive Bias in Medical Decision-Making

Humans are remarkably intelligent beings, surpassing other species on Earth in cognitive abilities. This intelligence enables us to engage in complex reasoning and problem-solving. However, being highly intelligent doesn't exempt us from cognitive biases that can skew our judgement. One notable example of this is the **belief bias effect** in logical reasoning. In medicine, just as in other fields, the ability to evaluate evidence impartially is crucial. The belief bias effect demonstrates that when assessing the validity of an argument—i.e., whether the conclusion logically follows from the premises—we tend to be influenced by how believable the conclusion appears to us.

Consider the following logically valid argument that aligns with common beliefs:

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

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

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

The logical *structure* of argument #2 is identical to the structure of argument #1, and they're both

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

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

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

And finally, an invalid argument with an unbelievable conclusion:

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

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

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

In a study relevant to this topic (Evans, Barston, and Pollard 1983), it was found that people are fairly good at identifying valid arguments when their beliefs align with the argument's conclusion. But when their beliefs run counter to a valid argument's conclusion, their ability to recognize its validity drops significantly. Furthermore, when people encounter an invalid argument that aligns with their pre-existing beliefs, they often fail to recognize its flaws.

What they found is that when pre-existing biases (i.e., beliefs) were in agreement with the structure of the data, everything went the way you'd hope:

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

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

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

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

While the data may not seem overly concerning, it's noteworthy that the accuracy in overcoming prior biases was about 60%, which is better than the 50% you'd expect by mere chance. Now, consider you're a medical professional tasked with diagnosing patients. If a tool could elevate your diagnostic accuracy from 60% to, let's say, 95%, wouldn't you want to utilize it? Absolutely, you would. Fortunately, such a tool exists: it's not magic, but statistics. This is one major reason why medical researchers value statistical methods. It's all too easy to fall prey to our own biases. If medical professionals could perfectly set aside cognitive biases, relying on intuition to evaluate data might be sufficient. However, that's far from the reality. The stakes in medical decision-making are extremely high, and even a moderate rate of error could have serious consequences. Statistics serves as a safeguard, helping us maintain objectivity and ensuring we rely on empirical evidence rather than subjective opinion. 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%

Given this, they were worried about being sued![2] 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.

[2] Earlier versions of these notes incorrectly suggested that they actually were sued. But 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.

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

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

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

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

Before leaving this topic entirely, I want to point out something else really critical that is often overlooked in a research methods class. Statistics only solves *part* of the problem. Remember that we started all this with the concern that Berkeley's admissions processes might be unfairly biased against female applicants. When we looked at the "aggregated" data, it did seem like the university

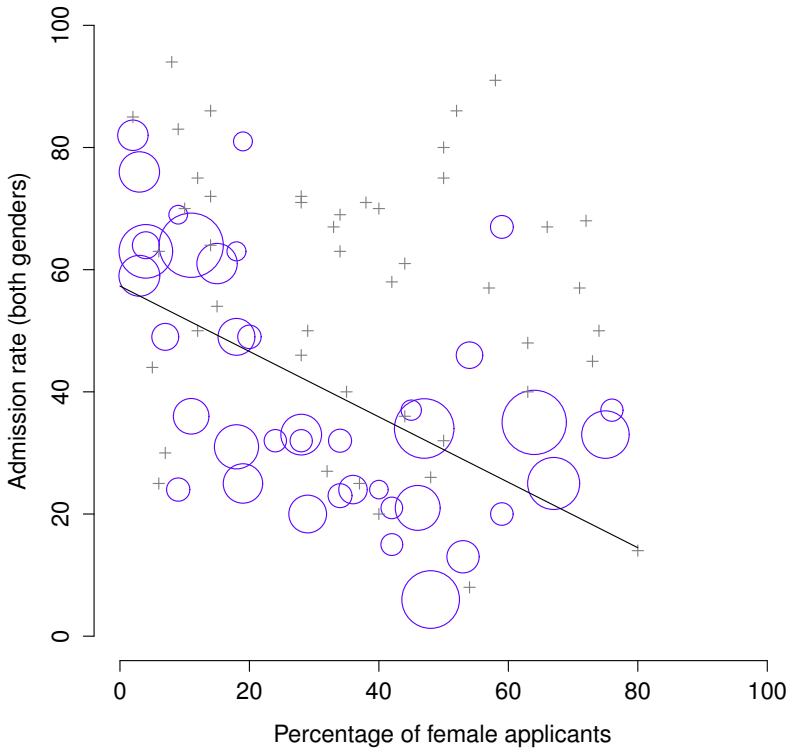

Figure 1.1: The Berkeley 1973 college admissions data. This figure plots the admission rate for the 85 departments that had at least one female applicant, as a function of the percentage of applicants that were female. The plot is a redrawing of Figure 1 from Bickel, Hammel, and O'Connell (1975). Circles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

. . . . . .

was discriminating against women, but when we "disaggregate" and looked at the individual behaviour of all the departments, it turned out that the actual departments were, if anything, slightly biased in favour of women. The gender bias in total admissions was caused by the fact that women tended to self-select for harder departments. From a legal perspective, that would probably put the university in the clear. Postgraduate admissions are determined at the level of the individual department, and there are good reasons to do that. At the level of individual departments the decisions are more or less unbiased (the weak bias in favour of females at that level is small, and not consistent across departments). Since the university can't dictate which departments people choose to apply to, and the decision making takes place at the level of the department it can hardly be held accountable for

any biases that those choices produce.

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

In short there are a lot of critical questions that you can't answer with statistics, but the answers to those questions will have a huge impact on how you analyse and interpret data. And this is the reason why you should always think of statistics as a *tool* to help 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 Medicine

I trust that the preceding discussion has shed light on the general significance of statistics in the realm of science. Yet, you might still be wondering what specific role statistics plays in medicine, especially considering the emphasis placed on this subject in your medical curriculum. Here, I will attempt to clarify some of your queries.

- **Why is statistics so prominent in medicine?**

  Frankly, there are several reasons, each with its own level of importance. The most crucial aspect is that medicine is a statistical science. The "subjects" we examine are *patients*—complex, multidimensional, ever-changing individuals. Unlike the predictable elements of certain natural sciences, human physiology and pathology are far from static. Patients can respond unpredictably to treatments, have differing baseline conditions, and are influenced by a myriad of external factors such as lifestyle and genetics.

In essence, if you are going into the medical field, you will find that statistics is indispensable. Unlike some fields where the dictum might be "if your experiment needs statistics, you should have designed a better experiment," medicine doesn't have this luxury. We are dealing with complex biological systems, not inanimate objects. Hence, understanding statistics is not a choice but a necessity.

- **Can't a specialist handle the statistics?**

  While it's true you don't need to be a statistical expert to practice medicine, a basic level of proficiency in statistics is essential. Here are three reasons why:

  – First, statistics and research design go hand-in-hand. Being good at one will inherently make you better at the other, especially when considering treatment efficacy and medical trials.
  – Second, most medical literature is replete with statistical data and analyses. Being able to understand this is vital for keeping up-to-date with medical advancements.
  – Third, employing a statistician for every piece of medical research is impractical and costly. Due to the shortage of trained statisticians, mastering basic statistical methods is not only practical but also economical.

  Moreover, these reasons are not limited to researchers alone. Even as a practicing clinician, you will benefit from being literate in statistics to interpret the latest findings in medical science.

- **What if I'm not interested in research or clinical practice? Do I still need statistics?**

  This might sound like a rhetorical question, but the importance of statistics transcends vocational concerns. We live in a data-driven era, and statistical literacy is almost a life skill. Understanding statistical concepts will empower you to make informed decisions, whether in clinical practice, research, or even in understanding health trends in the media.

1.4 Statistics in everyday life

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

When I started writing up my lecture notes I took the 20 most recent news articles posted to the ABC news website. Of those 20 articles, it turned out that 8 of them involved a discussion of something that I would call a statistical topic and 6 of those made a mistake. The most common error, if you're curious, was failing to report baseline data (e.g., the article mentions that 5% of people in situation X have some characteristic Y, but doesn't say how common the characteristic is for everyone else!). The

point I'm trying to make here isn't that journalists are bad at statistics (though they almost always are), it's that a basic knowledge of statistics is very helpful for trying to figure out when someone else is either making a mistake or even lying to you. In fact, one of the biggest things that a knowledge of statistics does to you is cause you to get angry at the newspaper or the internet on a far more frequent basis. You can find a good example of this in Section **??**. In later versions of this book I'll try to include more anecdotes along those lines.

1.5 There's more to research methods than statistics

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

But note that "urgent" is different from "important" – they both matter. I really do want to stress that research design is just as important as data analysis, and this book does spend a fair amount of time on it. However, while statistics has a kind of universality, and provides a set of core tools that are useful for most types of medical 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.

2. A brief introduction to research design

> *To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*
> – Sir Ronald Fisher[1]

In this chapter, we're going to start thinking about the basic ideas that go into designing a study, collecting data, checking whether your data collection works, and so on. It won't give you enough information to allow you to design studies of your own, but it will give you a lot of the basic tools that you need to assess the studies done by other people. However, since the focus of this book is much more on data analysis than on data collection, I'm only giving a very brief overview. Note that this chapter is "special" in two ways. Firstly, it's much more medicine-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.

2.1 Introduction to measurement

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

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

2.1.1 Some Considerations on Medical Measurement

Measurement in a medical context is a nuanced idea, but fundamentally, it involves assigning numbers, labels, or other kinds of well-defined descriptors to specific variables or "entities." Thus, the following can be considered examples of medical measurements:

- My **age** is *37 years old*.
- I *do not* have a **history of diabetes**.
- My **genetic predisposition** is *low risk for cardiovascular disease*.
- My **self-reported level of pain** is *4 on a scale of 10*.

In the list above, the **bolded text** represents "the variable to be measured," while the *italicized text* signifies "the actual measurement." We can delve further into this by considering the range of possible measurements that could be obtained for each variable:

- 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 questioned about my **history of diabetes**, the answers could be *Yes, I have diabetes*, *No, I don't have diabetes*, or *I have prediabetes*.
- My **genetic predisposition for cardiovascular disease** could be *low risk, medium risk, high risk*, or it could specify certain genetic markers associated with risk factors.
- My **self-reported level of pain** could be any number on a scale from *0 to 10*. Other scales, such as the Visual Analogue Scale, could also be used, allowing for a broader range of responses.

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. Yet, if you are focused on pediatric medicine, measuring age merely in years could be too imprecise. You may want to quantify age in *years and months*, often notated as "2;11" for a child who is 2 years and 11 months old. When dealing with neonates, age might even be measured in *days since birth* or *hours since birth*. Therefore, the manner in which you specify the possible range of measurements is critical and often contextual, depending on the specific needs and focus of the medical study.

Upon closer examination, even a seemingly straightforward concept like "age" can be complex and context-dependent. Generally, when we mention "age," it is implicitly understood as "the duration since birth." However, this may not always be the most scientifically meaningful measure. Consider a scenario in neonatology: if Baby Alice is born three weeks premature and Baby Bianca is born one

week post-term, would it be accurate to consider them of the "same age" if evaluated "two hours after birth"?

From a social standpoint, birth serves as a common reference point for age, marking the time a person has existed as a separate entity in the world. However, the medical and scientific perspective often necessitates a more nuanced understanding. When considering biological development, the relevant timeframe could span from conception rather than birth, as development begins long before an infant's birth. In such a case, Alice and Bianca would not be considered the same age from a developmental viewpoint.

Therefore, depending on the medical context, you might want to define "age" in multiple ways: the duration since conception and the duration since birth. While this distinction may not matter significantly in adult medicine, it could be crucial when focusing on neonates or pediatrics.

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

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

2.1.2 Operationalisation: defining your measurement

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

- Being precise about what you are trying to measure. For instance, does "age" mean "time since birth" or "time since conception" in the context of your research?
- Determining what method you will use to measure it. Will you use self-report to measure age, ask a parent, or look up an official record? If you're using self-report, how will you phrase the question?

- Defining the set of allowable values that the measurement can take. Note that these values don't always have to be numerical, though they often are. When measuring age the values are numerical, but we still need to think carefully about what numbers are allowed. Do we want age in years, years and months, days, or hours? For other types of measurements (e.g., gender) the values aren't numerical. But, just as before, we need to think about what values are allowed. If we're asking people to self-report their gender, what options to we allow them to choose between? Is it enough to allow only "male" or "female"? Do you need an "other" option? Or should we not give people specific options and instead let them answer in their own words? And if you open up the set of possible values to include all verbal response, how will you interpret their answers?

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

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

- **A theoretical construct**. This is the thing that you're trying to take a measurement of, like "age", "gender" or an "opinion". A theoretical construct can't be directly observed, and often they're actually a bit vague.
- **A measure**. The measure refers to the method or the tool that you use to make your observations. A question in a survey, a behavioural observation or a brain scan could all count as a measure.
- **An operationalisation**. The term "operationalisation" refers to the logical connection between the measure and the theoretical construct, or to the process by which we try to derive a measure from a theoretical construct.
- **A variable**. Finally, a new term. A variable is what we end up with when we apply our measure to something in the world. That is, variables are the actual "data" that we end up with in our data sets.

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

2.2

Scales of measurement

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

2.2.1 Nominal scale

A **nominal scale** variable (often termed a **categorical** variable) is one where the categories have no intrinsic order or hierarchy between them. For such variables, it's not meaningful to say that one category is "greater" or "more valuable" than another, and it's equally nonsensical to compute an average. A typical example in the medical field might be "blood types." Blood types can be A, B, AB, or O, but it would be absurd to say one type is "better" than another or to talk about an "average blood type." The same applies to categories like "disease type" or "medical procedure": whether a person has diabetes or hypertension is a nominal scale variable, as neither condition is inherently "greater" or "less" than the other.

To delve deeper, let's say you're conducting research on patient preferences for different types of pain management techniques. You might have a variable called "preferred pain management," with potential categories such as "medication," "physical therapy," "surgery," and "alternative therapies."

Preferred Pain Management	Number of Patients
(1) Medication	20
(2) Physical Therapy	35
(3) Surgery	15
(4) Alternative Therapies	30

So, can we talk about an "average" preferred pain management technique? Clearly, the notion is nonsensical. You can, however, observe that physical therapy is the most preferred method, while surgery is the least preferred. But beyond such observations, the sequence in which the options are listed is inconsequential.

Preferred Pain Management	Number of Patients
(3) Surgery	15
(1) Medication	20
(4) Alternative Therapies	30
(2) Physical Therapy	35

Switching the order of listing doesn't alter the essence of the data, underscoring the nominal nature

of this variable. Understanding such variables is crucial for medical researchers, particularly when examining patient populations or preferences, as it dictates the kinds of statistical analyses that are appropriate.

2.2.2 Ordinal scale

Ordinal scale variables possess more structure than nominal scale variables in that there is a meaningful way to order the categories, even though no other mathematical operations make sense on them. For instance, within a medical context, the "severity of a symptom" could serve as an ordinal scale variable. You can say that "severe" is worse than "moderate," which in turn is worse than "mild," but you can't quantify the exact difference between these categories. This structure can be mathematically represented as "severe" > "moderate" > "mild" "severe">"moderate">"mild", but it doesn't mean that the difference between "severe" and "moderate" is the same as between "moderate" and "mild."

Here's an example more directly related to a medical student's experience. Suppose you are surveying your colleagues to gauge their level of confidence in their understanding of medical research articles. You offer the following statements for them to choose from:

(1) I completely understand medical research articles
(2) I somewhat understand medical research articles
(3) I rarely understand medical research articles
(4) I don't understand medical research articles at all

These statements naturally follow an order regarding the level of understanding, which can be expressed as $1 > 2 > 3 > 4$. Such an order is crucial when presenting the options, as listing them in a disjointed manner like this would be confusing:

(3) I rarely understand medical research articles
(1) I completely understand medical research articles
(4) I don't understand medical research articles at all
(2) I somewhat understand medical research articles

Now, imagine that 100 students answered this survey with the following distribution:

Response	Number of Patients
(1) I completely understand medical research articles	51
(2) I somewhat understand medical research articles	20
(3) I rarely understand medical research articles	10
(4) I don't understand medical research articles at all	19

When analysing these data it seems quite reasonable to try to group (1), (2) and (3) together, and say that 81 out of 100 people were willing to *at least partially* understand research articles. And it's

also quite reasonable to group (2), (3) and (4) together and say that 49 out of 100 people registered *some level of difficulty* in comprehending medical literature. However, it would be entirely bizarre to try to group (1), (2) and (4) together and say that 90 out of 100 people said. . . what? There's nothing sensible that allows you to group those responses together at all.

That said, notice that while we *can* use the natural ordering of these items to construct sensible groupings, what we *can't* do is average them. For instance, in my simple example here, the "average" response to the question is 1.97. If you can tell me what that means I'd love to know, because it seems like gibberish to me! Despite its mathematical calculability, the "average" does not offer any interpretable insight into the collective understanding of medical research articles by medical students.

2.2.3 Interval scale

In contrast to nominal and ordinal scale variables, **interval scale** and ratio scale variables are variables for which the numerical value is genuinely meaningful. In the case of interval scale variables the *differences* between the numbers are interpretable, but the variable doesn't have a "natural" zero value. A good example of an interval scale variable in the context of medical research is the measurement of body temperature in degrees Celsius. If one patient has a body temperature of 37°C and another has a temperature of 39°C, the 2°C difference between them is meaningful. Furthermore, that 2°C difference is *exactly the same* as the 2°C difference between 35°C and 37°C. In this sense, addition and subtraction are meaningful operations for interval scale variables. [2]

However, it's important to recognize that 0°C does not mean "absence of body temperature." Therefore, it is misleading to perform multiplication and division operations on this scale. You cannot say that a body temperature of 40°C is "twice as hot" as 20°C, as this does not offer a meaningful interpretation.

In a medical research setting, suppose you're examining the effect of an antipyretic medication on postoperative fever. You would record the body temperatures of patients at different time intervals to determine the effectiveness of the medication. You could meaningfully interpret differences in temperature to conclude whether the medication effectively lowers postoperative fever. However, you would avoid stating that the medication makes patients "twice as less feverish," since this interpretation is inconsistent with the properties of interval scale data. Accurate understanding of your data scales is crucial for conducting rigorous medical research and for later translating these findings into clinical practice.

Lets look at an everyday example that should make things obvious. 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,

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

it would be completely daft for me to divide 2008 by 2003 and say that the second student started "1.0024 times later" than the first one. That doesn't make any sense at all.

2.2.4 Ratio scale

The last category of variables we'll explore is the **ratio scale** variable, where zero genuinely signifies zero, and multiplication and division are permissible. An example pertinent to medical research is the dosage of a medication. Often in clinical trials, it's crucial to record the dosage needed to achieve a particular therapeutic effect, as this serves as an indicator of the drug's efficacy. Let's say Patient A needs 10mg of Drug X to alleviate symptoms, whereas Patient B needs 14mg. Just like with an interval scale variable, arithmetic operations such as addition and subtraction are meaningful here. Patient B did indeed require $14 - 10 = 4$ mg more of the drug than Patient A. Furthermore, multiplication and division are also meaningful: Patient B needed $14/10 = 1.4$ times the dosage that Patient A needed. This is permissible because in a ratio scale variable like dosage, "zero mg" genuinely means "no medication at all."

2.2.5 Continuous versus discrete variables

There's a second kind of distinction that you need to be aware of, regarding what types of variables you can run into. This is the distinction between continuous variables and discrete variables. 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 become much more tangible when applied to clinical settings. Let's consider *blood pressure* as a continuous variable. If Patient A has a systolic blood pressure of 120 mmHg and Patient B has a systolic blood pressure of 130 mmHg, then Patient C with a pressure of 125 mmHg would fall in between them. Moreover, it's theoretically possible for Patient D to have a systolic pressure of 125.5 mmHg, a value that falls between Patient C's and Patient A's measurements. Although in clinical practice you might not measure blood pressure with such high precision, the principle holds that you *could*. Because we can always find a new value for blood pressure in between any two other ones we regard blood pressure as a continuous measure.

Discrete variables emerge when this rule is violated. For instance, variables on a nominal scale are always discrete in nature. In medical terminology, think about different types of tissues—epithelial, muscular, or nervous; there is no 'in-between' category that mathematically falls between epithelial and muscular tissue, much like there's no value that falls in between 2 and 3. Similarly, variables on an ordinal scale are always discrete. In medical research, the stages of cancer can be an example. Although 'Stage II' falls between 'Stage I' and 'Stage III', you won't find a stage that logically fits between 'Stage I' and 'Stage II'.

Table 2.1: The relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible.

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

Interval scale and ratio scale variables can go either way. As we saw above, blood pressure (a ratio scale variable) is continuous. Temperature in degrees celsius (an interval scale variable) is also continuous. However, the year you went to school (an interval scale variable) is discrete. There's no year in between 2002 and 2003. The number of questions you get right on a true-or-false test (a ratio scale variable) is also discrete. Since a true-or-false question doesn't allow you to be "partially correct", there's nothing in between 5/10 and 6/10.

Table 2.1 summarises the relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible. I'm trying to hammer this point home, because (a) some textbooks get this wrong, and (b) people very often say things like "discrete variable" when they mean "nominal scale variable".

2.2.6 Some complexities

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

Let's explore a fundamental measurement tool in medical surveys, the **Likert scale**. You've probably answered many such scales in medical settings, or perhaps you've even utilized one in your own research. Imagine this survey question:

> How would you rate your agreement with the statement, "Routine screenings are essential for early detection of diseases"?

The options given to the respondent are as follows:

(1) Strongly disagree
(2) Disagree
(3) Neither agree nor disagree

(4) Agree
(5) Strongly agree

This questionnaire is a classic 5-point Likert scale where participants choose from among several ordered possibilities, often accompanied by verbal descriptors. Though not mandatory, these descriptors aid in the participant's understanding. For instance, this version is equally valid:

(1) Strongly disagree
(2)
(3)
(4)
(5) Strongly agree

So, what kind of variable do these Likert scales represent? They are certainly discrete, ruling out the possibility of a 2.5 response. They are not nominal, as there is a clear order, and not ratio scale since there's no natural zero point.

The ambiguity arises when considering whether they are ordinal or interval scale variables. One could argue that the difference between "Strongly agree" and "Agree" isn't necessarily equal to the difference between "Agree" and "Neither agree nor disagree." This supports classifying the variable as ordinal scale. Yet, in many cases, respondents interpret the scale as approximately equal intervals, leading some researchers to treat these as interval scale. While not strictly interval scale, it's pragmatically close enough to be termed **quasi-interval scale**.

2.3 Assessing the reliability of a measurement

By now, we've considered how to operationalise a medical construct, thereby deriving a measure suitable for clinical or research settings. In the process, we obtain variables, which can manifest in various forms. It's time to address an essential question: is the measurement we've created any good for medical research? We'll do this in terms of two related ideas: *reliability* and *validity*. Put simply, the **reliability** of a measure tells you how *precisely* you are measuring something, whereas the validity of a measure tells you how *accurate* the measure is. In this section I'll talk about reliability; we'll talk about validity in section 2.6.

Reliability is actually a very simple concept. It refers to the repeatability or consistency of your measurement. For instance, measuring a patient's blood pressure using a sphygmomanometer is generally quite reliable. If you take multiple readings in quick succession under consistent conditions, you'll likely get the same result. On the other hand, assessing a patient's pain level by simply asking them might be less reliable, as the answer can vary based on subjective experience or emotional state at that moment. Note that reliability is distinct from the concept of validity, which concerns whether the measurement is actually correct or accurate.

Consider this example: if you weigh a patient while they are holding a heavy bag, the measurement will still be consistent if repeated. In other words, it's reliable. However, this does not mean it accurately represents the patient's true weight, making it an *reliable but invalid* measurement. Conversely, a quick assessment of a patient's mental state by a family member might fluctuate but could still generally be accurate, exemplifying an *unreliable but valid* measure.

It's important to understand that highly unreliable measures often end up being practically invalid because they offer inconsistent information. Thus, many researchers and clinicians argue that reliability is necessary (though not sufficient) to achieve 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 blood pressure) will they produce the same answer?
- **Parallel forms reliability**. This relates to consistency across theoretically-equivalent measurements. If I use a different set of weighing 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 mental state questionnaire result is added up across several questions) do the individual parts tend to give similar answers.

Not all measurements require every form of reliability either. Consider the evaluation process for a particular clinical condition, such as diabetes. The diagnostic approach might involve multiple components like blood sugar tests, patient history, and a physical examination. Each component is designed to measure different aspects of the patient's health and therefore may not be internally consistent with each other, which is acceptable in this context.

For instance, a fasting glucose test is meant to measure blood sugar levels at a specific point in time and is expected to be highly consistent if repeated under the same conditions. Conversely, patient history might include a range of symptoms and lifestyle factors that are qualitatively different but equally important. In this case, the fasting glucose test on its own should have high internal consistency, as it aims to measure the same variable each time it's performed.

The demand for reliability is contingent on what exactly you intend to measure. If multiple components of an evaluation or research study aim to measure different things, then internal consistency across the entire set of measures may not be critical. Understanding this nuance is essential, particularly as you read medical research or design your own studies, allowing you to critically assess the reliability and, consequently, the validity of the measures involved. You should only demand reliability in those situations where you want to be measuring the same thing!

Table 2.2: The terminology used to distinguish between different roles that a variable can play when analysing a data set. Note that this book will tend to avoid the classical terminology in favour of the newer names.

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

2.4 The "role" of variables: predictors and outcomes

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

When we are doing an analysis we have different names for X and Y, since they play different roles in the analysis. The classical names for these roles are **independent variable** (IV) and **dependent variable** (DV). The IV is the variable that you use to do the explaining (i.e., X) and the DV is the variable being explained (i.e., Y). The logic behind these names goes like this: if there really is a relationship between X and Y then we can say that Y depends on X, and if we have designed our study "properly" then X isn't dependent on anything else. However, I personally find those names horrible. They're hard to remember and they're highly misleading because (a) the IV is never actually "independent of everything else", and (b) if there's no relationship then the DV doesn't actually depend on the IV. And in fact, because I'm not the only person who thinks that IV and DV are just awful names, there are a number of alternatives that I find more appealing. The terms that I'll use in this book are **predictors** and **outcomes**. The idea here is that what you're trying to do is use X (the predictors) to make guesses about Y (the outcomes).[3] This is summarised in Table 2.2.

[3] Annoyingly though, there's a lot of different names used out there. I won't list all of them – there would be no point in doing that – other than to note that "response variable" is sometimes used where I've used "outcome". This sort of terminological confusion is very common, I'm afraid.

2.5 Experimental and non-experimental research

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

2.5.1 Experimental research

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

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

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

In the realm of medical research, experimental designs hold a pivotal role. Such experimental methods are often employed in pharmacological studies, clinical trials, and even in the evaluation of medical interventions like surgical techniques. For instance, to evaluate the effectiveness of a new antihypertensive drug, researchers might conduct a **randomized controlled trial** (RCT). In an RCT, participants are randomly assigned to either a treatment group receiving the new drug or a control group receiving a placebo. All other variables, like diet and exercise, would be controlled or balanced to ensure they don't interfere with the results. This way, any observed differences in blood pressure between the two groups can be causally linked to the drug itself. Similarly, to understand the impact of a new surgical technique, researchers might compare outcomes like recovery time or complication rates between a group undergoing the new technique and another receiving the standard procedure. By meticulously controlling the conditions and employing randomisation, researchers strive to eliminate or minimize confounders, thus making the conclusions more robust and reliable. Therefore, understanding the nuances of experimental research is crucial for anyone who aims to engage in medical research or be adept in interpreting medical literature.

2.5.2 Non-experimental research

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

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

The alternative approach, case studies, aims to provide a very detailed description of one or a few instances. In general, you can't use statistics to analyse the results of case studies and it's usually very hard to draw any general conclusions about "people in general" from a few isolated examples. However, case studies are very useful in some situations. Firstly, there are situations where you don't have any alternative. Nonetheless, in the medical field, case studies possess unique merits. For example, when dealing with rare diseases or unique medical conditions, clinicians may have limited options but to rely on case studies. This is often encountered in fields like neurology, where specific types of brain lesions are infrequent, so the only thing you can do is describe those cases that you do have in as much detail and with as much care as you can.

However, there's also some genuine advantages to case studies. Because you don't have as many people to study you have the ability to invest lots of time and effort trying to understand the specific factors at play in each case. This is a very valuable thing to do. As a consequence, case studies can complement the more statistically-oriented approaches that you see in experimental and quasi-experimental designs. We won't talk much about case studies in this book, but they are nevertheless very valuable tools!

2.6 Assessing the validity of a study

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

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

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

2.6.1 Internal validity

Internal Validity refers to the degree to which one can accurately infer causal connections between variables within a research study. It's called "internal" because it refers to the relationships between things "inside" the study. Let's illustrate the concept with a simple example. Imagine you want to investigate if a specific antihypertensive medication effectively lowers blood pressure. You administer the drug to a group of newly diagnosed hypertensive patients and observe a decrease in their blood pressure levels. Concurrently, you notice that a group of hypertensive patients who have been on medication for several years also show reduced blood pressure levels. One might hastily conclude that the medication is effective in lowering blood pressure. However, a critical issue arises; the long-term medication group is inherently different—they have been managed and possibly adherent to lifestyle changes for a longer period. Thus, the question arises: what is the true *cause* of the blood pressure reduction? Is it the medication, the lifestyle changes, or simply changes that occur naturally over time? In this scenario, the internal validity is compromised as the study design fails to isolate the *causal* relationships between the variables adequately. This is particularly pertinent for medical research, where multiple factors often interact, making it crucial to design experiments that accurately determine causality. This ability to delineate causal relationships is vital when interpreting clinical trials and epidemiological studies, which directly influence patient care and medical guidelines.

2.6.2 External validity

External Validity is concerned with the **generalizability** or **applicability** of research findings to broader contexts or populations. Specifically, it addresses how well the results of a study can be extrapolated to "real-world" situations outside of the research environment. In medical research, a study often involves a specific clinical setting, patient population, or medical condition. If the findings are not applicable to broader healthcare contexts or different patient populations, the study suffers from limited external validity.

A classic example within medical research involves clinical trials that primarily include participants from a specific age group, ethnicity, or health status. While the results may provide valuable insights for that particular demographic, they may not necessarily be applicable to the general patient population. This becomes particularly problematic if the study aims to inform treatment guidelines or therapeutic interventions for a diverse group of patients.

Thus, a study that lacks diversity in its participant pool runs the risk of low external validity. If there is something "unique" about the chosen sample that differentiates them from the larger population in a *relevant* manner, concerns about the generalizability of the findings arise. Understanding external validity is crucial when interpreting medical research, as healthcare decisions often depend on the applicability of research findings to diverse patient populations.

It's essential to clarify that a study focusing on a specific group, such as medical students as your research subjects, does not automatically suffer from a lack of external validity. The external validity of a study is compromised when two conditions are met: (a) the participant sample is drawn from a narrow population (e.g., medical students), and (b) this specific population differs from the broader

population *in a manner that is relevant to the medical or health phenomenon under investigation*. This subtlety is crucial and often overlooked. Medical students, like any specialized group, will vary from the general population in multiple aspects, but these differences may not necessarily impede the study's external validity. The key is to identify whether these differences are relevant to the specific medical research question.

To further illustrate this point, consider the following examples:

- You aim to assess "public perceptions of vaccine safety," but your participant pool consists solely of medical students. In this case, external validity is likely to be compromised, given that medical students' views may significantly differ from those of the general public on this matter.
- You want to measure the effectiveness of a visual illusion, and your participants are all medical students. This study is unlikely to have a problem with external validity

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

- Responses to a "medical questionnaire" may not accurately represent patients' behaviors or symptoms in a real-world clinical setting.
- A laboratory experiment on, for example, "drug interactions," may not fully replicate the complexities of polypharmacy often encountered in clinical practice.

2.6.3 Construct validity

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

2.6.4 Face validity

Face Validity refers to the intuitive credibility of a measure or test; essentially, whether it "ap-

pears" to assess what it claims to. For instance, if a diagnostic tool for assessing kidney function is presented, but experts in nephrology argue that the test doesn't seem relevant to kidney function, then the tool lacks face validity. From a rigorous scientific standpoint, face validity is not highly important. What truly matters is whether the measure *actually* accomplishes its intended purpose, rather than just *appearing* to do so. However, face validity has its practical utilities:

- Experienced clinicians or researchers may possess intuitive insights or "hunches" about the ineffectiveness of a particular measure. While these hunches don't serve as empirical evidence, they often warrant attention. Unspoken expertise may inform these instincts, making it prudent to reevaluate the study design if a respected colleague questions its face validity. Still, if no concrete flaws are found, it's generally reasonable to proceed given that face validity is not a critical factor.
- Criticisms about the superficial aspects of research often surface, especially on public forums like social media. These critiques usually target the study's appearance rather than its methodological integrity. Understanding face validity can be helpful for diplomatically urging critics to deepen their arguments.
- In applied medical research, especially when the goal is to influence healthcare policies or guidelines, face validity can become significantly important. Policy makers and non-experts may rely on face validity as a surrogate for actual validity. Despite the scientific rigor of a study, if it lacks face validity, it risks being dismissed in the policy arena. Although this may seem unfair, it reflects the reality that public perception can carry significant weight.

2.6.5 Ecological validity

Ecological Validity refers to how well the conditions and design of a study approximate the real-world medical scenario under investigation. Though related to external validity, it is less overarching. Ecological validity essentially deals with the study's "appearance" of realism, but demands more specific criteria to be met. The underlying rationale is that a study with high ecological validity is more likely to possess external validity, although it's not a guarantee.

For example, consider a study that investigates the impact of a newly developed drug on blood pressure. If the study is conducted in a controlled clinical setting with 24/7 monitoring, excluding any other medications and isolating participants from their usual environment and lifestyle, it may lack ecological validity. In the real world, patients are not isolated; they interact with various environmental factors, take other medications, and are exposed to daily life stressors that could affect blood pressure. Although the controlled setting is beneficial for establishing causal relationships, the absence of these real-world variables could question the study's ecological validity.

The advantage of considering ecological validity is that it is comparatively easier to evaluate than external validity. High ecological validity can often serve as a useful heuristic for expecting good external validity, especially when moving from research to practical medical applications or interventions. However, the lack of ecological validity doesn't automatically imply a lack of external validity; it

merely flags the need for cautious interpretation and potential additional studies that include real-world variables.

2.7 Confounds, artefacts and other threats to validity

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

- **Confounder**: A confounder is an additional, often unmeasured variable[4] that turns out to be related to both the predictors and the outcome. The existence of confounders threatens the internal validity of the study because you can't tell whether the predictor causes the outcome, or if the confounding variable causes it.
- **Artefact**: A result is said to be "artefactual" if it only holds in the special situation that you happened to test in your study. The possibility that your result is an artefact describes a threat to your external validity, because it raises the possibility that you can't generalise or apply your results to the actual population that you care about.

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

However, there's a trade-off to consider when discussing artefacts and confounders, especially in the realm of medical research. Generally speaking, artefactual results are often more of a concern for experimental studies than for non-experimental ones. To understand why, it's important to recognize that studies are often non-experimental precisely because the objective is to investigate phenomena in a more naturalistic setting. When you work in a real-world context, you may lose some degree of experimental control, making your study susceptible to confounders. However, by studying the phenomenon "in the wild" you minimize the likelihood of obtaining artefactual results. In contrast, when you extract a medical phenomenon from its natural environment to study it in a controlled laboratory setting (which is often necessary for more precise measurements), you risk inadvertently investigating something other than your intended focus.

Be warned though. The above is a rough guide only. It's absolutely possible to have confounders in an experiment, and to get artefactual results with non-experimental studies. This can happen for

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

all sorts of reasons, not least of which is experimenter or researcher error. In practice, it's really hard to think everything through ahead of time and even very good researchers make mistakes.

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

History effects refer to the impact that specific, often uncontrolled, events may have during the course of your study that could influence the outcome measure. For example, an event could occur between a pre-test and a post-test, or between the recruitment of one participant and the next. Such effects can also make an older study's conclusions less applicable to the current context.

- You're interested in investigating how medical students perceive the risk and resource allocation during pandemics. You started collecting your data in November 2019. However, recruitment and data collection span over several months, and you are still adding new participants in March 2020. Regrettably, the COVID-19 pandemic escalates during this period, affecting healthcare systems and causing significant mortality. Unsurprisingly, the perceptions of risk and resource allocation among participants in March 2020 are starkly different than those from November 2019. Which set of perceptions reflects the "true" sentiments? Arguably, both do. The COVID-19 pandemic has genuinely, and perhaps enduringly, altered perceptions and attitudes about medical risks and resource allocation. The key here is recognizing that the "history" for the March 2020 cohort is substantially different from that of the November 2019 cohort.
- You're conducting a study to evaluate the effectiveness of a novel antiviral medication. You measure baseline viral loads before administering the drug and plan to measure again post-treatment. However, during the course of the study, a new strain of the virus emerges, significantly affecting the efficacy of existing treatments, including potentially your antiviral medication. This unexpected event significantly confounds your results and needs to be accounted for in your analysis.

2.7.1 Maturation effects

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

- Consider a longitudinal study aimed at understanding the effectiveness of a new antihypertensive drug on blood pressure levels in adults. In such cases, it's important to account for the natural progression of hypertension with age. Failing to consider this maturational effect could lead you to incorrectly attribute changes in blood pressure solely to the intervention, when in fact some changes might be simply due to aging.

- When designing clinical trials that span an extended period (e.g., several hours or even multiple sessions), it's essential to consider that participants may experience fatigue, restlessness, or diminished attention over time. Such physiological maturation effects could potentially confound the study outcomes, obscuring the true impact of the treatment or intervention under investigation.

2.7.2 Repeated testing effects

An important type of history effect is the effect of **repeated testing**. Suppose I want to take two measurements of some 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 Adaptation*: For instance, if you're studying cognitive functions in patients with a neurological condition using a set of standardized tests, the scores at a second time point may appear to improve not because of any intervention, but because the participants have become familiar with the test format from the first session.
- *Acclimatization to the Research Setting*: In studies involving stress or pain levels, participants may report reduced stress or pain during subsequent visits simply because they've become accustomed to the testing environment or procedure, rather than any real reduction in their symptoms.
- *Physiological or Psychological Responses to Initial Testing*: Administering a lengthy or tedious diagnostic test could, in itself, produce fatigue or boredom in participants. These states may influence results at a second measurement point, confounding any changes attributed to an intervention or the natural course of a condition.

2.7.3 Selection bias

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

than males) then you're in a lot of trouble.

2.7.4 Differential attrition

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

The second type of attrition is **heterogeneous attrition**, in which the attrition effect is different for different groups. More often called **differential attrition**, this is a kind of selection bias that is caused by the study itself. For instance, if you are conducting a clinical trial comparing two different types of medication for chronic pain, and it turns out that more people drop out from the group receiving medication A due to side effects, the study's internal validity is compromised.

Let's delve into some more examples:

- *Patient Characteristics Influencing Attrition*: Suppose you are conducting a long-term study on the effects of a new anticoagulant drug. If participants who are more prone to missing medical appointments begin to drop out, your remaining sample might be skewed towards more conscientious patients, affecting the generalizability of your findings.
- *Treatment Effects Causing Differential Attrition*: In a clinical trial comparing a new pain relief method with a standard treatment, if the new method results in discomfort, you may see higher dropout rates in that group. This differential attrition undermines the internal validity, as the individuals who remain are possibly more tolerant to the method, thus biasing the study outcome.

Ethically and professionally, it's important to remind participants that they have the right to withdraw from the study at any point for any reason. However, it's equally important to design your study such that you minimize attrition, and analyze your data in a way that accounts for it, to maintain the integrity of your research.

2.7.5 Non-response bias

Non-response bias is closely related to selection bias and to differential attrition. The simplest version of the problem goes like this. You mail out a survey to 1000 people but only 300 of them reply. The 300 people who replied are almost certainly not a random subsample. People who respond to surveys are systematically different to people who don't. This introduces a problem when trying to generalise from those 300 people who replied to the population at large, since you now have a very non-random sample. The issue of non-response bias is more general than this, though. Among the (say) 300 people that did respond to the survey, you might find that not everyone answers every

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

2.7.6 Regression to the mean

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

Consider this slightly contrived scenario. You are interested in investigating whether a new medication improves the cognitive function of medical students during exam periods. You initially select 20 medical students with the highest baseline cognitive scores and evaluate their cognitive function after administering the medication. While these students continue to perform above average, they no longer hold the highest cognitive scores after the intervention. A tempting conclusion might be that the medication has had an adverse effect on their cognitive function. However, a more plausible explanation is that what you're observing is *regression to the mean*.

What contributes to being an outlier in cognitive function? It's not just natural ability or diligent study habits; an element of luck is also involved. Perhaps the baseline tests played to the specific strengths of these top-performing students. Since luck is not a consistent factor and doesn't transfer from the initial to subsequent measurements, it's likely that these students' scores would naturally decline or "regress" towards the mean upon retesting, regardless of any interventions.

Regression to the mean is surprisingly common. For instance, if two extremely 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 unusually short parents will tend to have short children, but nevertheless those kids will tend to be taller than the parents. It can also be extremely subtle. For instance, there have been studies done that suggested that people learn better from negative feedback than from positive feedback. However, the way that people tried to show this was to give people positive reinforcement whenever they did good, and negative reinforcement when they did bad. And what you see is that after the positive reinforcement people tended to do worse, but after the negative reinforcement they tended to do better. But notice that there's a selection bias here! When people do very well, you're selecting for "high" values, and so you should *expect*, because of regression to the mean, that performance on the next trial should be worse regardless of whether reinforcement is given. Similarly, after a bad trial, people will tend to improve all on their own. The apparent superiority of negative feedback is an artefact caused by regression to the mean (see Kahneman and Tversky 1973,

for discussion).

2.7.7 Experimenter bias

Experimenter bias can come in multiple forms. The basic idea is that the experimenter, despite the best of intentions, can accidentally end up influencing the results of the experiment by subtly communicating the "right answer" or the "desired behaviour" to the participants. Typically, this occurs because the experimenter has special knowledge that the participant does not, for example the right answer to the questions being asked or knowledge of the expected pattern of performance for the condition that the participant is in. The classic example of this happening is the case study of "Clever Hans", which dates back to 1907 (Pfungst 1911; 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 the humans did know how to count and the horse had learned to change its behaviour when people changed theirs.

In clinical research, the gold standard to mitigate investigator bias is a double-blind study design. In this model, neither the medical staff administering the treatment nor the study participants are aware of the intervention being done, thereby eliminating potential biases. However, executing a perfectly double-blind study is challenging. This provides a very good solution to the problem, but it's important to recognise that it's not quite ideal, and (depending on the type of study) may be 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 PhD 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 PhD student) doesn't know anything. Except for the reality that the last part is very unlikely to be true. In order for the Ph.D. student to run the study effectively they need to have been briefed by me, the researcher. And, as it happens, the PhD student also knows me and knows a bit about my general beliefs about people. As a result of all this, it's almost impossible for the experimenter to avoid knowing a little bit about what expectations I have. And even a little bit of knowledge can have an effect. Suppose the experimenter accidentally conveys the fact that the participants are expected to do well in this task. Well, there's a thing called the "Pygmalion effect", where if you expect great things of people they'll tend to rise to the occasion. But if you expect them to fail then they'll do that too. In other words, the expectations become a self-fulfilling prophesy.

2.7.8 Demand effects and reactivity

When talking about experimenter bias, the worry is that the experimenter's knowledge or desires for the experiment are communicated to the participants, and that these can change people's behaviour (Rosenthal 1966). However, even if you manage to stop this from happening, it's almost impossible to stop people from knowing that they're part of a medical study. And the mere fact of knowing

that someone is watching or studying you can have a pretty big effect on behaviour. This is generally referred to as **reactivity** or **demand effects**. The basic idea is captured by the Hawthorne effect: people alter their performance because of the attention that the study focuses on them. The effect takes its name from a study that took place in the "Hawthorne Works" factory outside of Chicago (see Adair 1984). This study, from the 1920s, looked at the effects of factory lighting on worker productivity. But, importantly, change in worker behaviour occurred because the workers *knew* they were being studied, rather than any effect of factory lighting.

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

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

2.7.9 Placebo effects

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

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

2.7.10 Situation, measurement and sub-population effects

In some respects, these terms are a catch-all term for "all other threats to external validity". They refer to the fact that the choice of sub-population from which you draw your participants, the location, timing and manner in which you run your study (including who collects the data) and the tools that

you use to make your measurements might all be influencing the results. Specifically, the worry is that these things might be influencing the results in such a way that the results won't generalise to a wider array of people, places and measures.

2.7.11 Fraud, deception and self-deception

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

There's one final thing I feel I should mention. While reading what the textbooks often have to say about assessing the validity of a study I couldn't help but notice that they seem to make the assumption that the researcher is honest. I find this hilarious. While the vast majority of scientists are honest, in my experience at least, some are not.[5] 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:

- **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 Andrew Wakefield (a doctor who has been accused of fabricating his data connecting the MMR vaccine to autism) and Hwang Woo-suk (who falsified a lot of his data on stem cell research).
- **Hoaxes**. Hoaxes share a lot of similarities with data fabrication, but they differ in the intended purpose. A hoax is often a joke, and many of them are intended to be (eventually) discovered. Often, the point of a hoax is to discredit someone or some field. There's quite a few well known scientific hoaxes that have occurred over the years (e.g., Piltdown man) and some were deliberate attempts to discredit particular fields of research (e.g., the Sokal affair).
- **Data misrepresentation**. While fraud gets most of the headlines, it's much more common in my experience to see data being misrepresented. When I say this I'm not referring to newspapers getting it wrong (which they do, almost always). I'm referring to the fact that often the data don't actually say what the researchers think they say. My guess is that, almost always, this

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

isn't the result of deliberate dishonesty but instead is due to a lack of sophistication in the data analyses. For instance, think back to the example of Simpson's paradox that I discussed in the beginning of this book. It's very common to see people present "aggregated" data of some kind and sometimes, when you dig deeper and find the raw data yourself you find that the aggregated data tell a different story to the disaggregated data. Alternatively, you might find that some aspect of the data is being hidden, because it tells an inconvenient story (e.g., the researcher might choose not to refer to a particular variable). There's a lot of variants on this, many of which are very hard to detect.

- **Study "misdesign"**. Okay, this one is subtle. Basically, the issue here is that a researcher designs a study that has built-in flaws and those flaws are never reported in the paper. The data that are reported are completely real and are correctly analysed, but they are produced by a study that is actually quite wrongly put together. The researcher really wants to find a particular effect and so the study is set up in such a way as to make it "easy" to (artefactually) observe that effect. One sneaky way to do this, in case you're feeling like dabbling in a bit of fraud yourself, is to design an experiment in which it's obvious to the participants what they're "supposed" to be doing, and then let reactivity work its magic for you. If you want you can add all the trappings of double blind experimentation but it won't make a difference since the study materials themselves are subtly telling people what you want them to do. When you write up the results the fraud won't be obvious to the reader. What's obvious to the participant when they're in the experimental context isn't always obvious to the person reading the paper. Of course, the way I've described this makes it sound like it's always fraud. Probably there are cases where this is done deliberately, but in my experience the bigger concern 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 misrepresent the data is by engaging in what's referred to as "data mining" (see Gelman and Loken 2014, for a broader discussion of this as part of the "garden of forking paths" in statistical analysis). As we'll discuss later, if you keep trying to analyse your data in lots of different ways, you'll eventually find something that "looks" like a real effect but isn't. This is referred to as "data mining". It used to be quite rare because data analysis used to take weeks, but now that everyone has very powerful statistical software on their computers it's becoming very common. Data mining per se isn't "wrong", but the more that you do it the bigger the risk you're taking. The thing that is wrong, and I suspect is very common, is *unacknowledged* data mining. That is, the researcher runs every possible analysis known to humanity, finds the one that works, and then pretends that this was the only analysis that they ever conducted. Worse yet, they often "invent" a hypothesis after looking at the data to cover up the data mining. To be clear. It's not wrong to change your beliefs after looking at the data, and to reanalyse your data using your new "post hoc" hypotheses. What is wrong (and I suspect common) is failing to acknowledge that you've done. If you acknowledge that you did it then other researchers are able to take your behaviour into account. If you don't, then they can't. And that makes your behaviour deceptive. Bad!

- **Publication bias & 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.[6] This is an example of a *publication bias*. Since no-one ever published the 19 studies that didn't find an effect, a naive reader would never know that they existed. Worse yet, most researchers "internalise" this bias and end up *self-censoring* their research. Knowing that negative results aren't going to be accepted for publication, they never even try to report them. As a friend of mine says "for every experiment that you get published, you also have 10 failures". And she's right. The catch is, while some (maybe most) of those studies are failures for boring reasons (e.g. you stuffed something up) others might be genuine "null" results that you ought to acknowledge when you write up the "good" experiment. And telling which is which is often hard to do. A good place to start is a paper by Ioannidis (2005) with the depressing title "Why most published research findings are false".

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.

2.8 Summary

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

- *Introduction to measurement* (Section 2.1). What does it mean to operationalise a theoretical construct? What does it mean to have variables and take measurements?
- *Scales of measurement and types of variables* (Section 2.2). 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* (Section 2.3). 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

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

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* (Section 2.4). 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* (Section 2.5). 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* (Section 2.6). 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 et al. (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.

LEARNING STATISTICS

WITH JASP