

# 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' Università ta' Malta  
`claud.bajada@um.edu.mt`

November 6, 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 published under a Creative Commons BY-SA license (CC BY-SA) version 4.0.

This means that this book can be reused, remixed, retained, revised and redistributed (including commercially) as long as appropriate credit is given to the authors. If you remix, or modify the original version of this open textbook, you must redistribute all versions of this open textbook under the same license - CC BY-SA.

<https://creativecommons.org/licenses/by-sa/4.0/>

*To Beni, my son.*

*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.*

# Table of Contents

<b>Preface</b>	<b>ix</b>
<b>I Background</b>	<b>1</b>
<b>1 Why do we learn statistics?</b>	<b>3</b>
1.1 The Role of Statistics in Medical Education . . . . .	3
1.2 The cautionary tale of Simpson's paradox . . . . .	6
1.3 Statistics in Medicine . . . . .	10
1.4 Statistics in everyday life . . . . .	11
1.5 There's more to research methods than statistics . . . . .	11
<b>2 A brief introduction to research design</b>	<b>13</b>
2.1 Introduction to measurement . . . . .	13
2.2 Scales of measurement . . . . .	17
2.3 Assessing the reliability of a measurement . . . . .	23
2.4 The "role" of variables: predictors and outcomes . . . . .	24
2.5 Experimental and non-experimental research . . . . .	25
2.6 Assessing the validity of a study . . . . .	28
2.7 Confounds, artefacts and other threats to validity . . . . .	32
2.8 Summary . . . . .	42
<b>II Describing and displaying data with JASP</b>	<b>45</b>
<b>3 Getting started with JASP</b>	<b>47</b>
3.1 Installing JASP . . . . .	48
3.2 Analyses . . . . .	50
3.3 Loading data in JASP . . . . .	50
3.4 The spreadsheet . . . . .	52
3.5 Changing data from one measurement scale to another . . . . .	54
3.6 Quitting JASP . . . . .	54
3.7 Summary . . . . .	55
<b>4 Descriptive statistics</b>	<b>57</b>
4.1 Measures of central tendency . . . . .	57
4.2 Measures of variability . . . . .	67
4.3 Skew and kurtosis . . . . .	76
4.4 Descriptive statistics separately for each group . . . . .	79

## TABLE OF CONTENTS

4.5	Standard scores . . . . .	79
4.6	Summary . . . . .	83
<b>III</b>	<b>Statistical theory</b>	<b>85</b>
<b>5</b>	<b>Estimating unknown quantities from a sample</b>	<b>93</b>
5.1	Samples, populations and sampling . . . . .	93
5.2	The law of large numbers . . . . .	101
5.3	Sampling distributions and the central limit theorem . . . . .	103
5.4	Estimating population parameters . . . . .	111
5.5	Estimating a confidence interval . . . . .	118
5.6	Summary . . . . .	123
<b>6</b>	<b>Hypothesis testing</b>	<b>125</b>
6.1	A menagerie of hypotheses . . . . .	126
6.2	Two types of errors . . . . .	130
6.3	Test statistics and sampling distributions . . . . .	131
6.4	Making decisions . . . . .	133
6.5	The $p$ value of a test . . . . .	137
6.6	Reporting the results of a hypothesis test . . . . .	140
6.7	Running the hypothesis test in practice . . . . .	142
6.8	Effect size, sample size and power . . . . .	143
6.9	Some issues to consider . . . . .	150
6.10	Summary . . . . .	153

## 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 adaptation: 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 identifying 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.compcogscisysdney.com/learning-statistics-with-r.html>

Hard copy: [www.lulu.com/content/13570633](http://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,  $\chi^2$  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 discussed 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<sup>1</sup>

### 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:

---

<sup>1</sup>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>



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

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"	<b>46% say "valid"</b>
argument is invalid	<b>92% say "valid"</b>	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!<sup>2</sup> 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.

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):

Department	Males		Females	
	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

<sup>2</sup>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.

to be suggesting is that the female applicants tended to apply to “harder” departments. And in fact, if we look at Figure 1.1 we see that this trend is systematic, and quite striking. This effect is known as **Simpson’s paradox**. It’s not common, but it does happen in real life, and most people are very surprised by it when they first encounter it, and many people refuse to even believe that it’s real. It is very real. And while there are lots of very subtle statistical lessons buried in there, I want to use it to make a much more important point: doing research is hard, and there are *lots* of subtle, counter-intuitive traps lying in wait for the unwary. That’s reason #2 why scientists love statistics, and why we teach research methods. Because science is hard, and the truth is sometimes cunningly hidden in the nooks and crannies of complicated data.

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

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

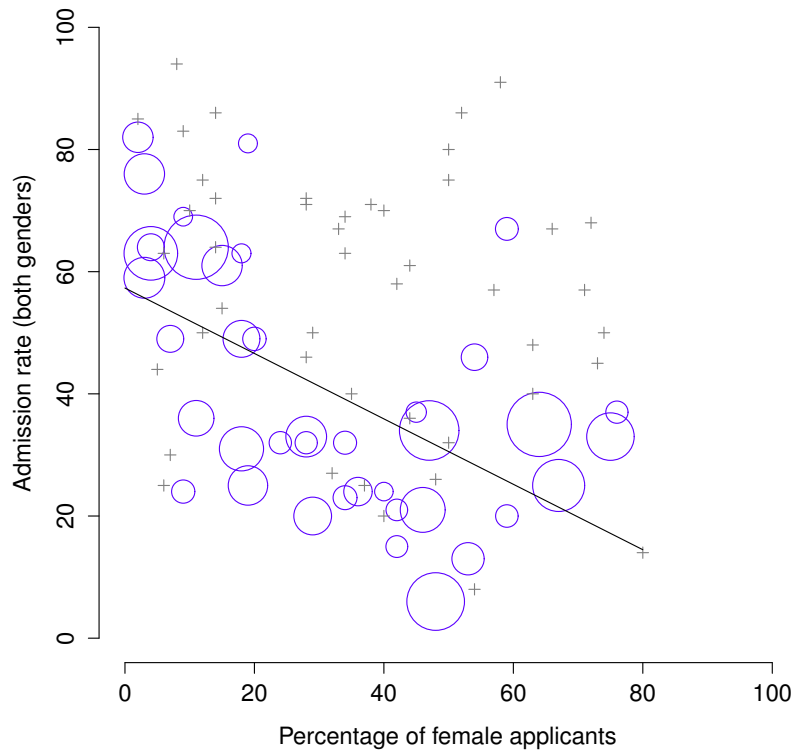


Figure 1.1: The Berkeley 1973 college admissions data. This figure plots the admission rate for the 85 departments that had at least one female applicant, as a function of the percentage of applicants that were female. The plot is a redrawing of Figure 1 from Bickel, Hammel, and O’Connell (1975). Circles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

.....

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.

## 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 important in medicine?**

Frankly, there are several reasons, each with its own level of importance. The most crucial aspect is that many aspects of medicine are statistical. The "subjects" of our research are complex, multidimensional, ever-changing biological creatures. Unlike the predictable elements of certain natural sciences, human physiology, epidemiology, pathology, and other medical sciences are far from static. Patients (or experimental animals) 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 4.1.5. 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<sup>1</sup>

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"?

---

<sup>1</sup>Presidential Address to the First Indian Statistical Congress, 1938. Source: [http://en.wikiquote.org/wiki/Ronald\\_Fisher](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 do we allow them to choose between? Is it enough to allow only "male" or "female"? Do you need an "other" option? Or should we not give people specific options and instead let them answer in their own words? And if you open up the set of possible values to include all verbal response, how will you interpret their answers?

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

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

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

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

## 2.2

---

### Scales of measurement

As the previous section indicates, the outcome of a 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. <sup>2</sup>

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

---

<sup>2</sup>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.



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, 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

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	✓	✓

.....

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'.

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**.

### 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!

## 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

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
.....		

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).<sup>3</sup> This is summarised in Table 2.2.

## 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

<sup>3</sup>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.

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 “appears” 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<sup>4</sup> 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

---

<sup>4</sup>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.

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 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 anti-hypertensive 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.<sup>5</sup> 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)

---

<sup>5</sup>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.

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 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.<sup>6</sup> 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

---

<sup>6</sup>Clearly, the real effect is that only insane people would even try to read *Finnegans Wake*.

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 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.





Part II.

## **Describing and displaying data with JASP**



### 3. Getting started with JASP

---

*Robots are nice to work with.*  
—Roger Zelazny<sup>1</sup>

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

However, before going into any of the specifics, it's worth talking a little about why you might want to use JASP at all. Given that you're reading this you've probably got your own reasons. However, if those reasons are "because that's what my stats class uses", it might be worth explaining a little why your professor has chosen to use JASP for the class. Of course, who really knows why *other* people choose JASP, so really, I will be talking about why I use it.

- It's sort of obvious but worth saying anyway: doing statistics on a computer is faster, easier and more powerful than doing statistics by hand. Computers excel at mindless repetitive tasks, and a lot of statistical calculations are both mindless and repetitive. For most people the only reason to ever do statistical calculations with pencil and paper is for learning purposes (even professionals do this when learning new concepts). In my class I do occasionally suggest doing some calculations that way, but the only real value to it is pedagogical. It does help you to get a "feel" for statistics to do some calculations yourself, so it's worth doing it once. But only once!
- Doing statistics in a conventional spreadsheet (e.g., Microsoft Excel) is generally a bad idea in the long run. Although many people likely feel more familiar with them, spreadsheets are very limited in terms of what analyses they allow you do. If you get into the habit of trying to do your real life data analysis using spreadsheets then you've dug yourself into a very deep hole.

---

<sup>1</sup>Source: *Dismal Light* (1968).

- Avoiding proprietary software is a very good idea. There are a lot of commercial packages out there that you can buy, some of which I like and some of which I don't. They're usually very glossy in their appearance and generally very powerful (much more powerful than spreadsheets). However, they're also very expensive. Usually, the company sells "student versions" (crippled versions of the real thing) very cheaply, and then they they sell full powered "educational versions" at a price that makes me wince. They will also sell commercial licences with a staggeringly high price tag. The business model here is to suck you in during your student days and then leave you dependent on their tools when you go out into the real world. It's hard to blame them for trying, but personally I'm not in favor of shelling out thousands of dollars if I can avoid it. And you can avoid it. If you make use of packages like JASP that are open source and free you never get trapped having to pay exorbitant licensing fees.

Those are the main reasons I use JASP. It's not without its flaws, though. It's relatively new<sup>2</sup> so there is not a huge set of textbooks and other resources to support it, and it has a few annoying quirks that we're all pretty much stuck with, but on the whole I think the strengths outweigh the weakness; more so than any other option I've encountered so far.

### 3.1

---

## Installing JASP

Okay, enough with the sales pitch. Let's get started. Just as with any piece of software, JASP needs to be installed on a computer. Fortunately, JASP is freely distributed online and you can download it from the JASP homepage, which is:

<https://jasp-stats.org/>

At the top of the page, you'll click on the heading "Download". Then, you'll see separate links for Windows users, Mac users, and Linux users. If you follow the relevant link you'll see that the online instructions are pretty self-explanatory. As of this writing, the current version of JASP is 0.18.1, but they usually issue updates every few months, so you'll probably have a newer version.<sup>3</sup>

### 3.1.1 Starting up JASP

JASP runs regardless of what operating system you're using Windows, Mac, Linux, ChromeOS

---

<sup>2</sup>As of writing this in May 2019.

<sup>3</sup>Although JASP is updated frequently it doesn't usually make much of a difference for the sort of work we'll do in this book. In fact, during the writing of the book I upgraded several times and it didn't make much difference at all to what is in this book.

... it even has an option to run online (through rollApp). One way or another, it's time to open JASP and get started. When first starting JASP you will be presented with a user interface which looks something like Figure 3.1.



Figure 3.1: JASP looks like this when you start it.

If you have experience with other statistical software packages, you might be a bit dismayed to see that there is no place to begin typing your data. This is a deliberate decision on the part of the JASP developers; their philosophy is that users should be allowed to use the editor they are most comfortable with <sup>4</sup>. Thus, the preferred method for getting data into JASP is to load a CSV file (.csv), which is a text-based data format that can be created by (and opened in) any spreadsheet program. More details about this will be given shortly.

<sup>4</sup>See <https://jasp-stats.org/2018/05/15/data-editing-in-jasp/> for a discussion of this very issue.

## 3.2

---

### Analyses

Analyses can be selected from several buttons along the top. Selecting an analysis will present an 'options panel' for that particular analysis, allowing you to assign different variables to different parts of the analysis, and select different options. At the same time, the results for the analysis will appear in the right 'Results panel' and will update in real-time as you make changes to the options.

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

If you decide you no longer need a particular analysis, you can remove it with the results context menu. Clicking on the header of a specific results header (or clicking on the ▼ symbol) will bring up a menu and by selecting 'Remove Analysis', the analysis can be removed. But more on this later. First, let's get some data into JASP.

## 3.3

---

### Loading data in JASP

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

- *.jasp files* are those with a `.jasp` file extension. This is the standard kind of file that JASP uses to store data, and variables and analyses.
- *Comma separated value (CSV) files* are those with a `.csv` file extension. These are just regular old text files and they can be opened with many different software programs. It's quite typical for people to store data in csv files, precisely because they're so simple.

#### 3.3.1 Importing data from CSV files

One quite commonly used data format is the humble "comma separated value" file, also called a CSV file, and usually bearing the file extension `.csv`. CSV files are just plain old-fashioned text files and what they store is basically just a table of data. This is illustrated in Figure 3.2, which shows a file called `booksales.csv` that I've created. As you can see, each

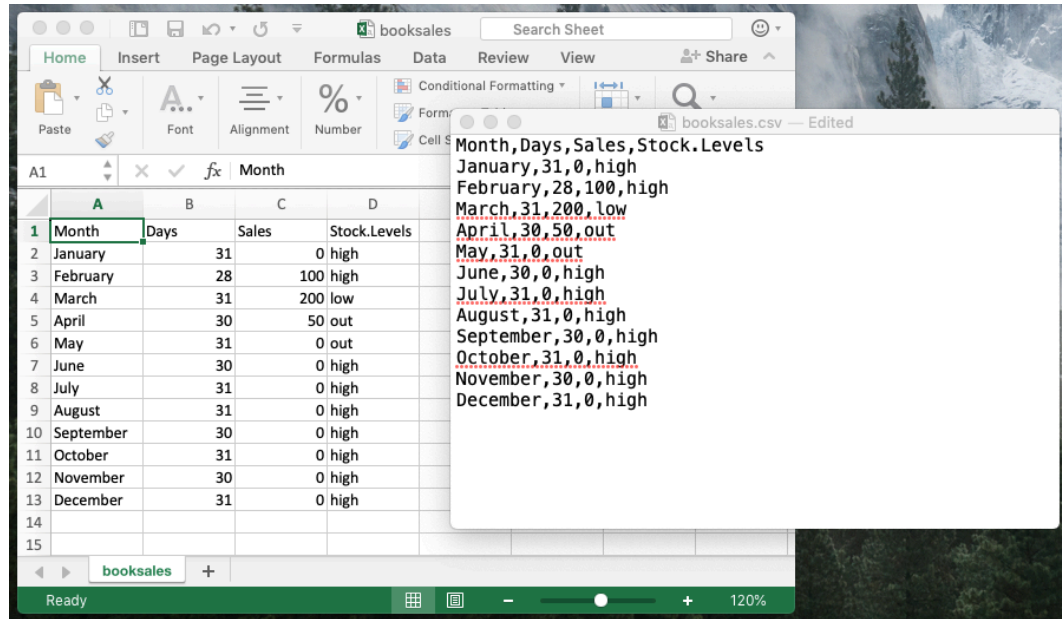


Figure 3.2: The `booksales.csv` data file. On the left I've opened the file using a spreadsheet program, which shows that the file is basically a table. On the right the same file is open in a standard text editor (the TextEdit program on a Mac), which shows how the file is formatted. The entries in the table are separated by commas.

.....

row represents the book sales data for one month. The first row doesn't contain actual data though, it has the names of the variables.

Once you have a CSV file (either that you created or someone has given you), you open the file in JASP by clicking the File tab at the top left hand corner, select 'Open', and then choosing from the options presented. Most commonly, you will select 'Computer' and then 'Browse', which will then open a file browser specific to your operating system. If you're on a Mac, it'll look like the usual Finder window that you use to choose a file; on Windows it looks like an Explorer window. An example of what it looks like on a Mac is shown in Figure 3.3. I'm assuming that you're familiar with your own computer, so you should have no problem finding



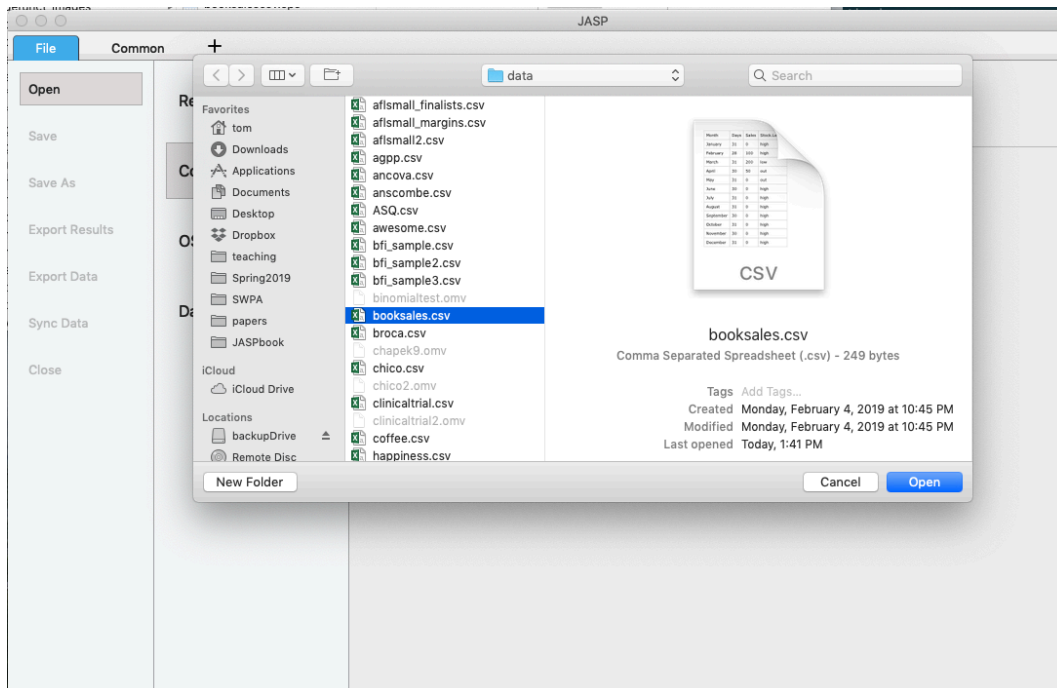


Figure 3.3: A dialog box on a Mac asking you to select the CSV file JASP should try to import. Mac users will recognise this immediately – it’s the usual way in which a Mac asks you to find a file. Windows users won’t see this, but instead will see the usual explorer window that Windows always gives you when it wants you to select a file.

the csv file that you want to import! Find the one you want, then click on the “Open” button.

### 3.4

#### The spreadsheet

Once loaded into JASP, data is represented in a spreadsheet with each column representing a ‘variable’ and each row representing a ‘case’ or ‘participant’.

### 3.4.1 Variables

The most commonly used variables in JASP are ‘Data Variables’, which contain data loaded from a CSV file. Data variables can be one of three measurement levels, which are designated by the symbol in the header of the variable’s column.

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

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

*Scale* variables are variables which exist on a continuous scale. Examples might be height or weight. This is also referred to as ‘Interval’ or ‘Ratio scale’.

Note that when opening a data file JASP will try and guess the variable type from the data in each column. In both cases this automatic approach may not be correct, and it may be necessary to manually specify the variable type with the variable editor.

### 3.4.2 Computed variables

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

Computed variables can be added to the data set with the ‘+’ button in the header row of the data spreadsheet. This will produce a dialog box where you can specify the formula using either R code or a drag-and-drop interface. At this point, I simply want you to know that the capability exists, but describing how to do it is a little beyond our scope right now. More later!

### 3.4.3 Copy and Paste

As a final note, we will mention that JASP produces nice American Psychological Association (APA) formatted tables and attractive plots. It is often useful to be able to copy and paste these, perhaps into a Word document, or into an email to a colleague. To copy results, click on the header of the object of interest and from the menu select exactly what you want to copy. Selecting “copy” copies the content to the clipboard and this can be pasted into other programs in the usual way. You can practice this later on when we do some analyses. Also, if you use the L<sup>A</sup>T<sub>E</sub>X document preparation system, you can select “Copy special” and “LaTeX code”; doing so will place the L<sup>A</sup>T<sub>E</sub>X syntax into your clipboard.

### Changing data from one measurement scale to another

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

In 3.4 we saw how to specify different variable levels, and if you want to change a variable's measurement level then you can do this in the JASP data view for that variable. Just click the check box for the measurement level you want - continuous, ordinal, or nominal.

### Quitting JASP

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

It is good practice to save any changes to the data set as a *new* data set. That way you can always go back to the original data. To save any changes in JASP, select 'Export Data' from the 'File' tab, click 'Browse' and navigate to the directory location in which you want to save the file, and create a new file name for the changed data set.

Alternatively, you can save *both* the changed data and any analyses you have undertaken by saving as a `.jasp` file. To do this, from the 'File' tab select 'Save as', click 'Browse' to navigate to the directory location in which you want to save the file, and type in a file name for this `.jasp` file. Remember to save the file in a location where you can find it again later. I usually create a new folder for specific data sets and analyses.

**Summary**

Every book that tries to teach a new statistical software program to novices has to cover roughly the same topics, and in roughly the same order. Ours is no exception, and so in the grand tradition of doing it just the same way everyone else did it, this chapter covered the following topics:

- Section 3.1. We downloaded and installed JASP, and started it up.
- Section 3.2. We very briefly oriented to the part of JASP where analyses are done and results appear, but then deferred this until later in the book.
- Section 3.3. We saw how to load data files (formatted as `.csv` files) in JASP.
- Section 3.4. We spent more time looking at the spreadsheet part of JASP, and considered different variable types, and briefly mentioned how to compute new variables.
- Section 3.5. And saw that sometimes we need to coerce data from one type to another.
- Section 3.6. Finally, we looked at good practice in terms of saving your data set and analyses when you have finished and are about to quit JASP.

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



## 4. Descriptive statistics

---

Any time that you get a new data set to look at one of the first tasks that you have to do is find ways of summarising the data in a compact, easily understood fashion. This is what **descriptive statistics** (as opposed to inferential statistics) is all about. In fact, to many people the term “statistics” is synonymous with descriptive statistics. It is this topic that we’ll consider in this chapter, but before going into any details, let’s take a moment to get a sense of why we need descriptive statistics. To do this, let’s open the `af1small_margins` file and see what variables are stored in the file.

In fact, there is just one variable here, `af1.margins`. We’ll focus a bit on this variable in this chapter, so I’d better tell you what it is. The `af1.margins` variable contains the reduction in systolic blood pressure reduction (in mmHg) observed in a 178 patients after being administered antihypertensive medication over a period of time.

This output doesn’t make it easy to get a sense of what the data are actually saying. Just “looking at the data” isn’t a terribly effective way of understanding data. In order to get some idea about what the data are actually saying we need to calculate some descriptive statistics (this chapter) and draw some nice pictures (Chapter ??). Since the descriptive statistics are the easier of the two topics I’ll start with those, but nevertheless I’ll show you a histogram of the `af1.margins` data since it should help you get a sense of what the data we’re trying to describe actually look like, see Figure 4.2. We’ll talk a lot more about how to draw histograms in Section ???. For now, it’s enough to look at the histogram and note that it provides a fairly interpretable representation of the `af1.margins` data.

### 4.1

---

#### Measures of central tendency

Drawing pictures of the data, as I did in Figure 4.2, is an excellent way to convey the “gist” of what the data is trying to tell you. It’s often extremely useful to try to condense the data into a few simple “summary” statistics. In most situations, the first thing that you’ll want to calculate is a measure of **central tendency**. That is, you’d like to know something about

	afl.margins
1	56
2	31
3	56
4	8
5	32
6	14
7	36
8	56
9	19
10	1
11	3
12	104
13	43
14	44

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

where the “average” or “middle” of your data lies. The three most commonly used measures are the mean, median and mode. I’ll explain each of these in turn, and then discuss when each of them is useful.

#### 4.1.1 The mean

The **mean** of a set of observations is just a normal, old-fashioned average. Add all of the values up, and then divide by the total number of values. The first five systolic blood pressure reductions were 56, 31, 56, 8 and 32, so the mean of these observations is just:

$$\frac{56 + 31 + 56 + 8 + 32}{5} = \frac{183}{5} = 36.60$$

Of course, this definition of the mean isn’t news to anyone. Averages (i.e., means) are used so often in everyday life that this is pretty familiar stuff. However, since the concept of a mean

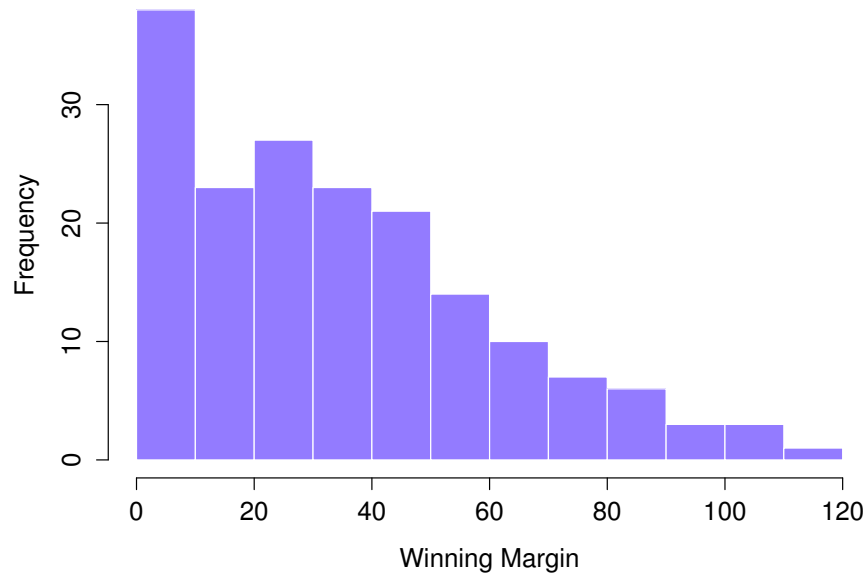


Figure 4.2: A histogram of the AFL 2010 winning margin data (the `af1.margins` variable). As you might expect, the larger the winning margin the less frequently you tend to see it.

.....

is something that everyone already understands, I'll use this as an excuse to start introducing some of the mathematical notation that statisticians use to describe this calculation, and talk about how the calculations would be done in JASP.

The first piece of notation to introduce is  $N$ , which we'll use to refer to the number of observations that we're averaging (in this case  $N = 5$ ). Next, we need to attach a label to the observations themselves. It's traditional to use  $X$  for this, and to use subscripts to indicate which observation we're actually talking about. That is, we'll use  $X_1$  to refer to the first observation,  $X_2$  to refer to the second observation, and so on all the way up to  $X_N$  for the last one. Or, to say the same thing in a slightly more abstract way, we use  $X_i$  to refer to the  $i$ -th observation. Just to make sure we're clear on the notation, the following table lists the 5 observations in the `af1.margins` variable, along with the mathematical symbol used to refer to it and the actual value that the observation corresponds to:



the observation	its symbol	the observed value
Blood pressure decrease, patient 1	$X_1$	56 mmHg
Blood pressure decrease, patient 2	$X_2$	31 mmHg
Blood pressure decrease, patient 3	$X_3$	56 mmHg
Blood pressure decrease, patient 4	$X_4$	8 mmHg
Blood pressure decrease, patient 5	$X_5$	32 mmHg

Okay, now let's try to write a formula for the mean. By tradition, we use  $\bar{X}$  as the notation for the mean. So the calculation for the mean could be expressed using the following formula:

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

This formula is entirely correct but it's terribly long, so we make use of the **summation symbol**  $\Sigma$  to shorten it.<sup>a</sup> If I want to add up the first five observations I could write out the sum the long way,  $X_1 + X_2 + X_3 + X_4 + X_5$  or I could use the summation symbol to shorten it to this:

$$\sum_{i=1}^5 X_i$$

Taken literally, this could be read as “the sum, taken over all  $i$  values from 1 to 5, of the value  $X_i$ ”. But basically what it means is “add up the first five observations”. In any case, we can use this notation to write out the formula for the mean, which looks like this:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

In all honesty, I can't imagine that all this mathematical notation helps clarify the concept of the mean at all. In fact, it's really just a fancy way of writing out the same thing I said in words: add all the values up and then divide by the total number of items. However, that's not really the reason I went into all that detail. My goal was to try to make sure that everyone reading this book is clear on the notation that we'll be using throughout the book:  $\bar{X}$  for the mean,  $\Sigma$  for the idea of summation,  $X_i$  for the  $i$ th observation, and  $N$  for the total number of observations. We're going to be re-using these symbols a fair bit so it's important that you understand them well enough to be able to “read” the equations, and to be able to see that it's just saying “add up lots of things and then divide by another thing”.

<sup>a</sup>The choice to use  $\Sigma$  to denote summation isn't arbitrary. It's the Greek upper case letter sigma, which is the analogue of the letter S in that alphabet. Similarly, there's an equivalent symbol used to denote the multiplication of lots of numbers, because multiplications are also called “products” we use the  $\Pi$  symbol for this (the Greek upper case pi, which is the analogue of the letter P).

#### 4.1.2 Calculating the mean in JASP

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

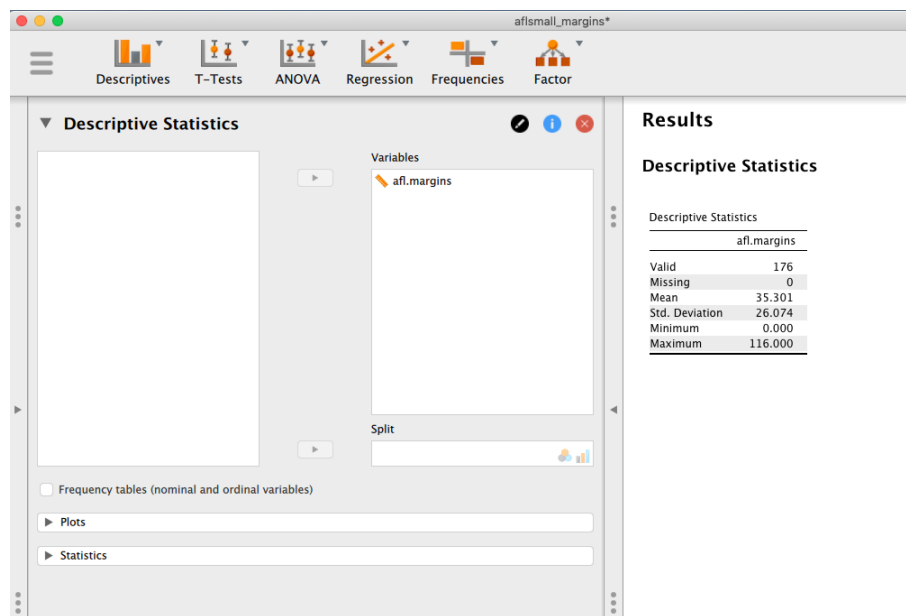


Figure 4.3: Default descriptives for the systolic blood pressure reduction data (the `afl.margins` variable).

As you can see in Figure 4.3, the mean value for the `afl.margins` variable is 35.301. Other information presented includes the total number of observations ( $N=176$ ), the number of missing values (none), and the Median, Minimum and Maximum values for the variable.

#### 4.1.3 The median

The second measure of central tendency that people use a lot is the **median**, and it's even easier to describe than the mean. The median of a set of observations is just the middle value.

As before let's imagine we were interested only in the first 5 AFL winning margins: 56, 31, 56, 8 and 32. To figure out the median we sort these numbers into ascending order:

8, 31, **32**, 56, 56

From inspection, it's obvious that the median value of these 5 observations is 32 since that's the middle one in the sorted list (I've put it in bold to make it even more obvious). Easy stuff. But what should we do if we are interested in the first 6 games rather than the first 5? Since the sixth game in the season had a winning margin of 14 points, our sorted list is now

8, 14, **31**, **32**, 56, 56

and there are *two* middle numbers, 31 and 32. The median is defined as the average of those two numbers, which is of course 31.5. As before, it's very tedious to do this by hand when you've got lots of numbers. In real life, of course, no-one actually calculates the median by sorting the data and then looking for the middle value. In real life we use a computer to do the heavy lifting for us. JASP will give us the median if we ask for it; we simply need to click on the 'Statistics' dropdown menu and select 'Median' from the 'Central Tendency' menu. The results will automatically update to include this median, which JASP reports as 30.500 for the `afl.margins` variable.

#### 4.1.4 Mean or median? What's the difference?

Knowing how to calculate means and medians is only a part of the story. You also need to understand what each one is saying about the data, and what that implies for when you should use each one. This is illustrated in Figure 4.4. The mean is kind of like the "centre of gravity" of the data set, whereas the median is the "middle value" in the data. What this implies, as far as which one you should use, depends a little on what type of data you've got and what you're trying to achieve. As a rough guide:

- If your data are nominal scale you probably shouldn't be using either the mean or the median. Both the mean and the median rely on the idea that the numbers assigned to values are meaningful. If the numbering scheme is arbitrary then it's probably best to use the mode (Section 4.1.6) instead.
- If your data are ordinal scale you're more likely to want to use the median than the mean. The median only makes use of the order information in your data (i.e., which numbers are bigger) but doesn't depend on the precise numbers involved. That's exactly the situation that applies when your data are ordinal scale. The mean, on the other hand, makes use of the precise numeric values assigned to the observations, so it's not really appropriate for ordinal data.
- For interval and ratio scale data either one is generally acceptable. Which one you pick depends a bit on what you're trying to achieve. The mean has the advantage that it uses all the information in the data (which is useful when you don't have a lot of data). But it's very sensitive to extreme, outlying values.

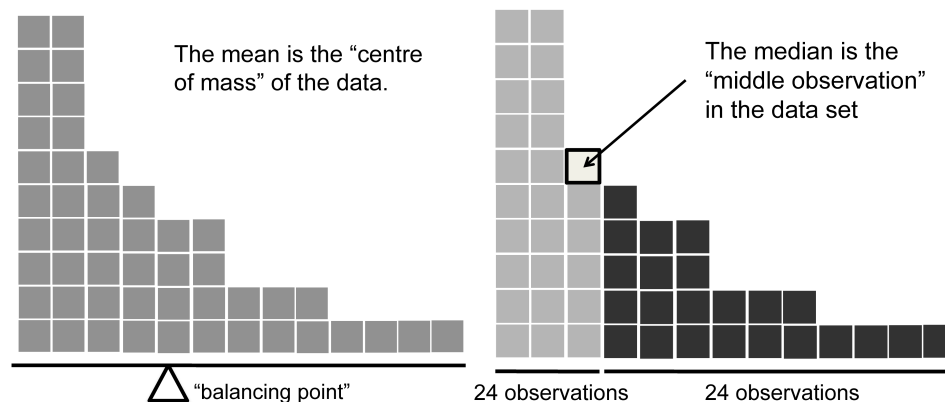


Figure 4.4: An illustration of the difference between how the mean and the median should be interpreted. The mean is basically the “centre of gravity” of the data set. If you imagine that the histogram of the data is a solid object, then the point on which you could balance it (as if on a see-saw) is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

Let’s expand on that last part a little. One consequence is that there are systematic differences between the mean and the median when the histogram is asymmetric (skewed; see Section 4.3). This is illustrated in Figure 4.4. Notice that the median (right hand side) is located closer to the “body” of the histogram, whereas the mean (left hand side) gets dragged towards the “tail” (where the extreme values are). To give a concrete example, consider a General Practitioner who has three patients with blood glucose levels of 5.0, 5.5, and 6.0 mmol/L. Their average glucose level would be 5.5 mmol/L, and the median would be also 5.5 mmol/L. Now, suppose a fourth patient joins with a glucose level of 20 mmol/L. The average glucose level skyrockets to about 9.1 mmol/L, whereas the median merely increases to 5.75 mmol/L. If you’re evaluating the overall blood glucose of the group, the mean might offer valuable insights. However, for a more typical representation of patient responses, the median would serve as a more robust measure.

#### 4.1.5 A real life example

To try to get a sense of why you need to pay attention to the differences between the mean and the median let’s consider a real life example. This will be one of the rare non-medical examples since I believe it make the point excellently. This example is from an excellent article on the ABC news published on 24 September, 2010:

Senior Commonwealth Bank executives have travelled the world in the past couple of weeks with a presentation showing how Australian house prices, and the key price to income ratios, compare favourably with similar countries. “Housing affordability has actually been

going sideways for the last five to six years," said Craig James, the chief economist of the bank's trading arm, CommSec.

This probably comes as a huge surprise to anyone with a mortgage, or who wants a mortgage, or pays rent, or isn't completely oblivious to what's been going on in the Australian housing market over the last several years. Back to the article:

CBA has waged its war against what it believes are housing doomsayers with graphs, numbers and international comparisons. In its presentation, the bank rejects arguments that Australia's housing is relatively expensive compared to incomes. It says Australia's house price to household income ratio of 5.6 in the major cities, and 4.3 nationwide, is comparable to many other developed nations. It says San Francisco and New York have ratios of 7, Auckland's is 6.7, and Vancouver comes in at 9.3.

More excellent news! Except, the article goes on to make the observation that:

Many analysts say that has led the bank to use misleading figures and comparisons. If you go to page four of CBA's presentation and read the source information at the bottom of the graph and table, you would notice there is an additional source on the international comparison – Demographia. However, if the Commonwealth Bank had also used Demographia's analysis of Australia's house price to income ratio, it would have come up with a figure closer to 9 rather than 5.6 or 4.3

That's, um, a rather serious discrepancy. One group of people say 9, another says 4-5. Should we just split the difference and say the truth lies somewhere in between? Absolutely not! This is a situation where there is a right answer and a wrong answer. Demographia is correct, and the Commonwealth Bank is wrong. As the article points out:

[An] obvious problem with the Commonwealth Bank's domestic price to income figures is they compare average incomes with median house prices (unlike the Demographia figures that compare median incomes to median prices). The median is the mid-point, effectively cutting out the highs and lows, and that means the average is generally higher when it comes to incomes and asset prices, because it includes the earnings of Australia's wealthiest people. To put it another way: the Commonwealth Bank's figures count Ralph Norris' multi-million dollar pay packet on the income side, but not his (no doubt) very expensive house in the property price figures, thus understating the house price to income ratio for middle-income Australians.

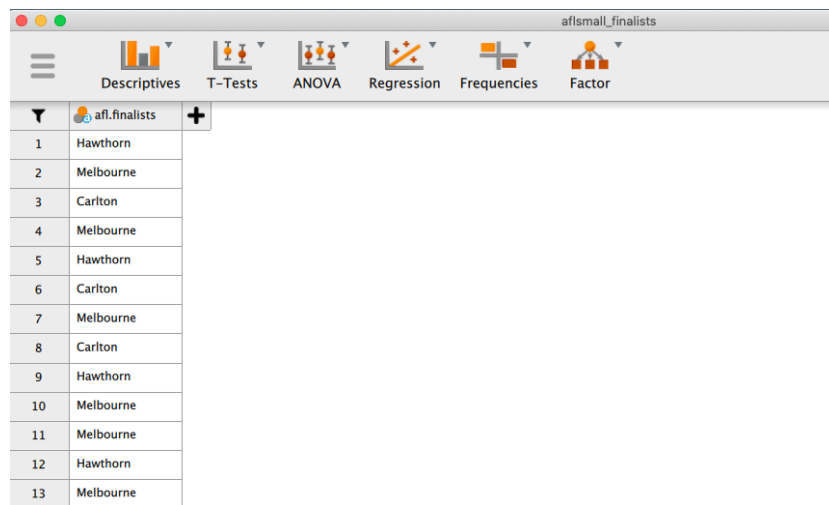
Couldn't have put it better myself. The way that Demographia calculated the ratio is the right thing to do. The way that the Bank did it is incorrect. As for why an extremely quantitatively sophisticated organisation such as a major bank made such an elementary mistake, well... I can't say for sure since I have no special insight into their thinking. But the article itself does happen to mention the following facts, which may or may not be relevant:

[As] Australia's largest home lender, the Commonwealth Bank has one of the biggest vested interests in house prices rising. It effectively owns a massive swathe of Australian housing as security for its home loans as well as many small business loans.

My, my.

#### 4.1.6 Mode

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



	afl.finalists
1	Hawthorn
2	Melbourne
3	Carlton
4	Melbourne
5	Hawthorn
6	Carlton
7	Melbourne
8	Carlton
9	Hawthorn
10	Melbourne
11	Melbourne
12	Hawthorn
13	Melbourne

Figure 4.5: A screenshot of JASP showing the variables stored in the `aflsmall_finalists.csv` file

.....

What we *could* do is read through all 400 entries and count the number of occasions on which each team name appears in our list of finalists, thereby producing a **frequency table**. However, that would be mindless and boring: exactly the sort of task that computers are great at. So let's use JASP to do this for us. Under 'Descriptives' - 'Descriptive Statistics', select the `afl.finalists` variable and move it to the 'Variables' box, then click the small check box labelled 'Frequency tables'. You should get something like Figure 4.6.

Now that we have our frequency table we can just look at it and see that, over the 24 years for which we have data, Geelong has played in more finals than any other team. Thus, the mode of the `afl.finalists` data is "Geelong". We can see that Geelong (39 finals) played in more finals than any other team during the 1987-2010 period. It's also worth noting that in the 'Descriptives' Table no results are calculated for Mean, Median, Minimum or Maximum.

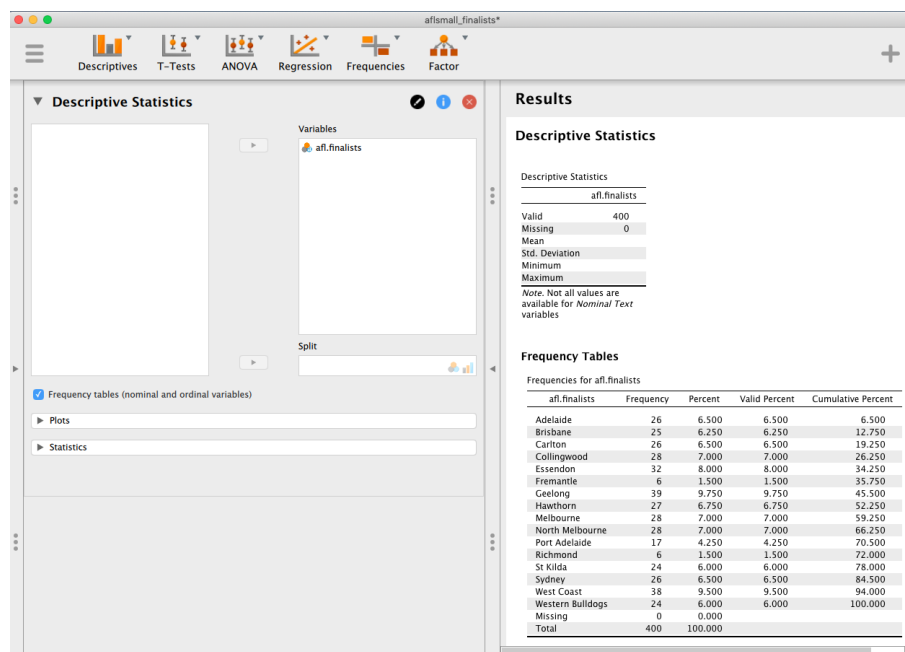


Figure 4.6: A screenshot of JASP showing the frequency table for the afl.finalists variable

This is because the `afl.finalists` variable is a nominal text variable so it makes no sense to calculate these values.

One last point to make regarding the mode. Whilst the mode is most often calculated when you have nominal data, because means and medians are useless for those sorts of variables, there are some situations in which you really do want to know the mode of an ordinal, interval or ratio scale variable. For instance, let's go back to our `afl.margins` variable. This variable is clearly ratio scale (if it's not clear to you, it may help to re-read Section 2.2), and so in most situations the mean or the median is the measure of central tendency that you want. But consider this scenario: a friend of yours is offering a bet and they pick a football game at random. Without knowing who is playing you have to guess the *exact* winning margin. If you guess correctly you win \$50. If you don't you lose \$1. There are no consolation prizes for "almost" getting the right answer. You have to guess exactly the right margin. For this bet, the mean and the median are completely useless to you. It is the mode that you should bet on. To calculate the mode for the `afl.margins` variable in JASP, go back to that data set and on the 'Descriptives' - 'Descriptive Statistics' screen you will see you can expand the section marked 'Statistics'. Click on the checkbox marked 'Mode' and you will see the modal value presented in the 'Descriptive Statistics' Table, as in Figure 4.7. So the 2010 data suggest you should bet on a 3 point margin.

## 4.2

---

### Measures of variability

The statistics that we've discussed so far all relate to *central tendency*. That is, they all talk about which values are "in the middle" or "popular" in the data. However, central tendency is not the only type of summary statistic that we want to calculate. The second thing that we really want is a measure of the **variability** of the data. That is, how "spread out" are the data? How "far" away from the mean or median do the observed values tend to be? For now, let's assume that the data are interval or ratio scale, and we'll continue to use the `afl.margins` data. We'll use this data to discuss several different measures of spread, each with different strengths and weaknesses.

#### 4.2.1 Range

The **range** of a variable is very simple. It's the biggest value minus the smallest value. For the AFL winning margins data the maximum value is 116 and the minimum value is 0. Although the range is the simplest way to quantify the notion of "variability", it's one of the worst. Recall from our discussion of the mean that we want our summary measure to be robust. If



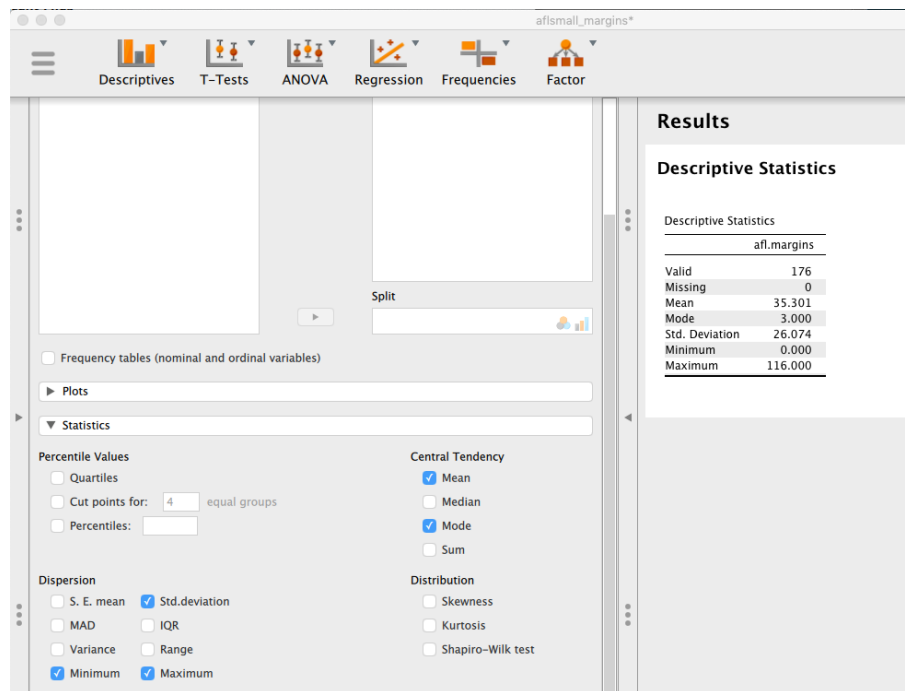


Figure 4.7: A screenshot of JASP showing the modal value for the `afl.margins` variable

the data set has one or two extremely bad values in it we'd like our statistics to not be unduly influenced by these cases. For example, in a variable containing very extreme outliers

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

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

#### 4.2.2 Interquartile range

The **interquartile range** (IQR) is like the range, but instead of the difference between the biggest and smallest value the difference between the 25th percentile and the 75th percentile is taken. If you don't already know what a **percentile** is, the 10th percentile of a data set is the smallest number  $x$  such that 10% of the data is less than  $x$ . In fact, we've already come across the idea. The median of a data set is its 50th percentile! In JASP you can easily specify the 25th, 50th and 75th percentiles by clicking the checkbox 'Quartiles' in the 'Descriptives' - 'Descriptive Statistics' - 'Statistics' screen.

## Descriptive Statistics

Descriptive Statistics	
afl.margins	
Valid	176
Missing	0
Mean	35.301
Mode	3.000
Std. Deviation	26.074
Minimum	0.000
Maximum	116.000
25th percentile	12.250
50th percentile	30.500
75th percentile	51.500

Figure 4.8: A screenshot of JASP showing the Quartiles for the `afl.margins` variable

And not surprisingly, in Figure 4.8 the 50th percentile is the same as the median value. And, by noting that  $50.50 - 12.75 = 37.75$ , we can see that the interquartile range for the 2010 AFL winning margins data is 37.75. While it's obvious how to interpret the range it's a little less obvious how to interpret the IQR. The simplest way to think about it is like this: the interquartile range is the range spanned by the “middle half” of the data. That is, one quarter of the data falls below the 25th percentile and one quarter of the data is above the 75th percentile, leaving the “middle half” of the data lying in between the two. And the IQR is the range covered by that middle half.

### 4.2.3 Mean absolute deviation

The two measures we've looked at so far, the range and the interquartile range, both rely on the idea that we can measure the spread of the data by looking at the percentiles of the data. However, this isn't the only way to think about the problem. A different approach is to select a meaningful reference point (usually the mean or the median) and then report the “typical” deviations from that reference point. What do we mean by “typical” deviation? Usually, this is the mean or median value of these deviations. In practice, this leads to two different measures: the “mean absolute deviation” (from the mean) and the “median absolute deviation” (from the median). From what I've read, the measure based on the median seems to be used in statistics and does seem to be the better of the two. But to be honest I don't think I've seen it used much in academic articles. The measure based on the mean does

occasionally show up though. In this section I'll talk about the first one, and I'll come back to talk about the second one later.

Since the previous paragraph might sound a little abstract, let's go through the **mean absolute deviation** from the mean a little more slowly. One useful thing about this measure is that the name actually tells you exactly how to calculate it. Let's think about our AFL winning margins data, and once again we'll start by pretending that there are only 5 games in total, with winning margins of 56, 31, 56, 8 and 32. Since our calculations rely on an examination of the deviation from some reference point (in this case the mean), the first thing we need to calculate is the mean,  $\bar{X}$ . For these five observations, our mean is  $\bar{X} = 36.6$ . The next step is to convert each of our observations  $X_i$  into a deviation score. We do this by calculating the difference between the observation  $X_i$  and the mean  $\bar{X}$ . That is, the deviation score is defined to be  $X_i - \bar{X}$ . For the first observation in our sample, this is equal to  $56 - 36.6 = 19.4$ . Okay, that's simple enough. The next step in the process is to convert these deviations to absolute deviations, and we do this by converting any negative values to positive ones. Mathematically, we would denote the absolute value of  $-3$  as  $|-3|$ , and so we say that  $|-3| = 3$ . We use the absolute value here because we don't really care whether the value is higher than the mean or lower than the mean, we're just interested in how *close* it is to the mean. To help make this process as obvious as possible, the table below shows these calculations for all five observations:

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

Now that we have calculated the absolute deviation score for every observation in the data set, all that we have to do to calculate the mean of these scores. Let's do that:

$$\frac{19.4 + 5.6 + 19.4 + 28.6 + 4.6}{5} = 15.52$$

And we're done. The mean absolute deviation for these five scores is 15.52.

However, whilst our calculations for this little example are at an end, we do have a couple of things left to talk about. First, we should really try to write down a proper mathematical formula. But in order to do this I need some mathematical notation to refer to the mean absolute deviation. Irritatingly, “mean absolute deviation” and “median absolute deviation” have the same acronym (MAD), which leads to a certain amount of ambiguity so I’d better come up with something different for the mean absolute deviation. Sigh. What I’ll do is use AAD instead, short for *average* absolute deviation. Now that we have some unambiguous notation, here’s the formula that describes what we just calculated:

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

#### 4.2.4 Variance

Although the average absolute deviation measure has its uses, it’s not the best measure of variability to use. From a purely mathematical perspective there are some solid reasons to prefer squared deviations rather than absolute deviations. If we do that we obtain a measure called the **variance**, which has a lot of really nice statistical properties that I’m going to ignore,<sup>1</sup> and one massive psychological flaw that I’m going to make a big deal out of in a moment. The variance of a data set  $X$  is sometimes written as  $\text{Var}(X)$ , but it’s more commonly denoted  $s^2$  (the reason for this will become clearer shortly).

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

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

As you can see, it’s basically the same formula that we used to calculate the average absolute deviation, except that instead of using “absolute deviations” we use “squared deviations”. It is for this reason that the variance is sometimes referred to as the “mean square deviation”.

Now that we’ve got the basic idea, let’s have a look at a concrete example. Once again, let’s use the first five AFL games as our data. If we follow the same approach that we took last time, we end up with the following table:

---

<sup>1</sup>Well, I will very briefly mention the one that I think is coolest, for a very particular definition of “cool”, that is. Variances are *additive*. Here’s what that means. Suppose I have two variables  $X$  and  $Y$ , whose variances are  $\text{Var}(X)$  and  $\text{Var}(Y)$  respectively. Now imagine I want to define a new variable  $Z$  that is the sum of the two,  $Z = X + Y$ . As it turns out, the variance of  $Z$  is equal to  $\text{Var}(X) + \text{Var}(Y)$ . This is a *very* useful property, but it’s not true of the other measures that I talk about in this section.

English: maths:	which game $i$	value $X_i$	deviation from mean $X_i - \bar{X}$	squared deviation $(X_i - \bar{X})^2$
	1	56	19.4	376.36
	2	31	-5.6	31.36
	3	56	19.4	376.36
	4	8	-28.6	817.96
	5	32	-4.6	21.16

That last column contains all of our squared deviations, so all we have to do is average them. If we do that by hand, i.e. using a calculator, we end up with a variance of 324.64. Exciting, isn't it? For the moment, let's ignore the burning question that you're all probably thinking (i.e., what the heck does a variance of 324.64 actually mean?) and instead talk a bit more about how to do the calculations in JASP, because this will reveal something very weird.

First, you'll need to load a new data file that contains *only* the first 5 rows. Do to this, load the file `aflsmall_margins_first5.csv`. Next, click 'Descriptives' – 'Descriptive Statistics', and then under the 'Statistics' menu, click the 'Variance' check box (you'll find it in the 'Dispersion' group). Do you get the same values for variance as the one we calculated by hand (324.64)? No, wait, you get a completely *different* answer (405.800)! That's just weird. Is JASP broken? Is this a typo? What is going on?

As it happens, the answer is no. It's not a typo, and JASP is not making a mistake. In fact, it's very simple to explain what JASP is doing here, but slightly trickier to explain *why* JASP is doing it. So let's start with the "what". What JASP is doing is evaluating a slightly different formula to the one I showed you above. Instead of averaging the squared deviations, which requires you to divide by the number of data points  $N$ , JASP has chosen to divide by  $N - 1$ .

In other words, the formula that JASP is using is this one:

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

So that's the *what*. The real question is *why* JASP is dividing by  $N - 1$  and not by  $N$ . After all, the variance is supposed to be the *mean* squared deviation, right? So shouldn't we be dividing by  $N$ , the actual number of observations in the sample? Well, yes, we should. However, as we'll discuss in Chapter 5, there's a subtle distinction between "describing a sample" and "making guesses about the population from which the sample came". Up to this point, it's been a distinction without a difference. Regardless of whether you're describing a sample or drawing inferences about the population, the mean is calculated exactly the same way. Not so for the variance, or the standard deviation, or for many other measures besides. What I outlined to you initially (i.e., take the actual average, and thus divide by  $N$ ) assumes that you literally intend to calculate the variance of the sample. Most of the time, however,

you're not terribly interested in the sample *in and of itself*. Rather, the sample exists to tell you something about the world. If so, you're actually starting to move away from calculating a "sample statistic" and towards the idea of estimating a "population parameter". However, I'm getting ahead of myself. For now, let's just take it on faith that JASP knows what it's doing, and we'll revisit the question later on when we talk about estimation in Chapter 5.

Okay, one last thing. This section so far has read a bit like a mystery novel. I've shown you how to calculate the variance, described the weird " $N - 1$ " thing that JASP does and hinted at the reason why it's there, but I haven't mentioned the single most important thing. How do you *interpret* the variance? Descriptive statistics are supposed to describe things, after all, and right now the variance is really just a gibberish number. Unfortunately, the reason why I haven't given you the human-friendly interpretation of the variance is that there really isn't one. This is the most serious problem with the variance. Although it has some elegant mathematical properties that suggest that it really is a fundamental quantity for expressing variation, it's completely useless if you want to communicate with an actual human. Variances are completely uninterpretable in terms of the original variable! All the numbers have been squared and they don't mean anything anymore. This is a huge issue. For instance, according to the table I presented earlier, the margin in game 1 was "376.36 points-squared higher than the average margin". This is *exactly* as stupid as it sounds, and so when we calculate a variance of 324.64 we're in the same situation. I've watched a lot of football games, and at no time has anyone ever referred to "points squared". It's *not* a real unit of measurement, and since the variance is expressed in terms of this gibberish unit, it is totally meaningless to a human.

#### 4.2.5 Standard deviation

Okay, suppose that you like the idea of using the variance because of those nice mathematical properties that I haven't talked about, but since you're a human and not a robot you'd like to have a measure that is expressed in the same units as the data itself (i.e., points, not points-squared). What should you do? The solution to the problem is obvious! Take the square root of the variance, known as the **standard deviation**, also called the "root mean squared deviation", or RMSD. This solves our problem fairly neatly. Whilst nobody has a clue what "a variance of 324.68 points-squared" really means, it's much easier to understand "a standard deviation of 18.01 points" since it's expressed in the original units. It is traditional to refer to the standard deviation of a sample of data as  $s$ , though "sd" and "std dev." are also used at times.

Because the standard deviation is equal to the square root of the variance, you probably won't be surprised to see that the formula is:

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

and in JASP there is a check box for 'Std. deviation' in the same section as the check box for 'Variance'. As you can see in Figure 4.8, the value JASP gives for the standard deviation of `afl.margins` is 26.074. Note that because standard deviation is used so often, it is checked by default, so you will rarely have to actually select it yourself!

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

For reasons that will make sense when we return to this topic in Chapter 5 I'll refer to this new quantity as  $\hat{\sigma}$  (read as: "sigma hat"), and the formula for this is:

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

Interpreting standard deviations is slightly more complex. Because the standard deviation is derived from the variance, and the variance is a quantity that has little to no meaning that makes sense to us humans, the standard deviation doesn't have a simple interpretation. As a consequence, most of us just rely on a simple rule of thumb. In general, you should expect 68% of the data to fall within 1 standard deviation of the mean, 95% of the data to fall within 2 standard deviation of the mean, and 99.7% of the data to fall within 3 standard deviations of the mean. This rule tends to work pretty well most of the time, but it's not exact. It's actually calculated based on an *assumption* that the histogram is symmetric and "bell shaped".<sup>2</sup> As you can tell from looking at the AFL winning margins histogram in Figure 4.2, this isn't exactly true of our data! Even so, the rule is approximately correct. As it turns out, 65.3% of the AFL margins data fall within one standard deviation of the mean. This is shown visually in Figure 4.9.

#### 4.2.6 Which measure to use?

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

---

<sup>2</sup>Strictly, the assumption is that the data are *normally* distributed, which is an important concept that we'll discuss more in Chapter ?? and will turn up over and over again later in the book.

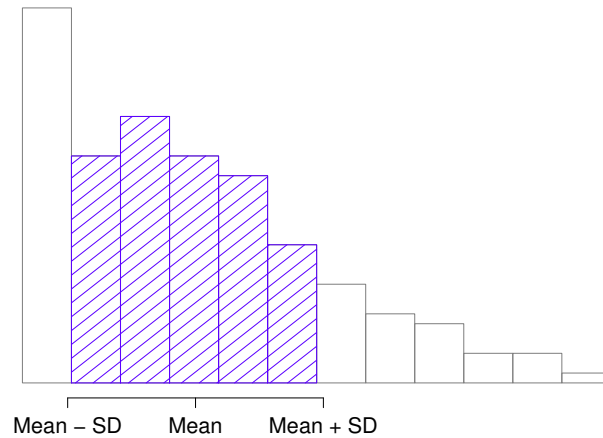


Figure 4.9: An illustration of the standard deviation from the AFL winning margins data. The shaded bars in the histogram show how much of the data fall within one standard deviation of the mean. In this case, 65.3% of the data set lies within this range, which is pretty consistent with the “approximately 68% rule” discussed in the main text.

.....

variance and standard deviation; and hinted at their strengths and weaknesses. Here’s a quick summary:

- *Range*. Gives you the full spread of the data. It’s very vulnerable to outliers and as a consequence it isn’t often used unless you have good reasons to care about the extremes in the data.
- *Interquartile range*. Tells you where the “middle half” of the data sits. It’s pretty robust and complements the median nicely. This is used a lot.
- *Mean absolute deviation*. Tells you how far “on average” the observations are from the mean. It’s very interpretable but has a few minor issues (not discussed here) that make it less attractive to statisticians than the standard deviation. Used sometimes, but not often.
- *Variance*. Tells you the average squared deviation from the mean. It’s mathematically elegant and is probably the “right” way to describe variation around the mean, but it’s completely uninterpretable because it doesn’t use the same units as the data. Almost never used except as a mathematical tool, but it’s buried “under the hood” of a very large number of statistical tools.



- *Standard deviation.* This is the square root of the variance. It's fairly elegant mathematically and it's expressed in the same units as the data so it can be interpreted pretty well. In situations where the mean is the measure of central tendency, this is the default. This is by far the most popular measure of variation.

In short, the IQR and the standard deviation are easily the two most common measures used to report the variability of the data. But there are situations in which the others are used. I've described all of them in this book because there's a fair chance you'll run into most of these somewhere.

### 4.3

---

#### Skew and kurtosis

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

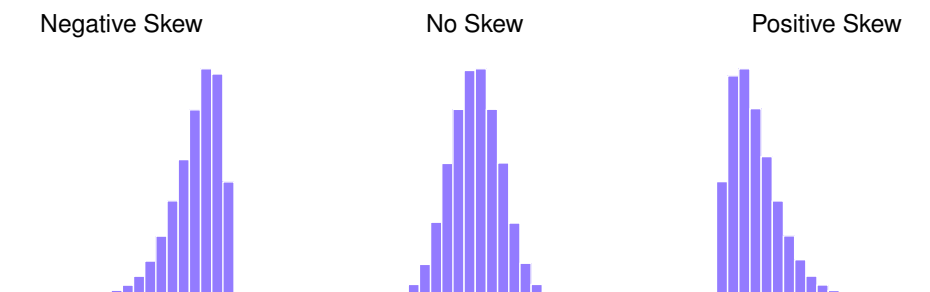


Figure 4.10: An illustration of skewness. On the left we have a negatively skewed data set (skewness =  $-.93$ ), in the middle we have a data set with no skew (well, hardly any: skewness =  $-.006$ ), and on the right we have a positively skewed data set (skewness =  $.93$ ).

.....

Since it's the more interesting of the two, let's start by talking about the **skewness**. Skewness is basically a measure of asymmetry and the easiest way to explain it is by drawing some pictures. As Figure 4.10 illustrates, if the data tend to have a lot of extreme small

values (i.e., the lower tail is “longer” than the upper tail) and not so many extremely large values (left panel) then we say that the data are *negatively skewed*. On the other hand, if there are more extremely large values than extremely small ones (right panel) we say that the data are *positively skewed*. That’s the qualitative idea behind skewness. If there are relatively more values that are far greater than the mean, the distribution is positively skewed or right skewed, with a tail stretching to the right. Negative or left skew is the opposite. A symmetric distribution has a skewness of 0. The skewness value for a positively skewed distribution is positive, and a negative value for a negatively skewed distribution.

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

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

where  $N$  is the number of observations,  $\bar{X}$  is the sample mean, and  $\hat{\sigma}$  is the standard deviation (the “divide by  $N - 1$ ” version, that is).

Perhaps more helpfully, you can use JASP to calculate skewness: it’s a check box in the ‘Statistics’ options under ‘Descriptives’ - ‘Descriptive Statistics’. For the `af1.margins` variable, the skewness figure is 0.780. If you divide the skewness estimate by the Std. error for skewness you have an indication of how skewed the data is. Especially in small samples ( $N < 50$ ), one rule of thumb suggests that a value of 2 or less can mean that the data is not very skewed, and a value of over 2 that there is sufficient skew in the data to possibly limit its use in some statistical analyses. Note that this is a rule of thumb – there is no clear agreement on this interpretation. That said, such analysis does indicate that the AFL winning margins data is somewhat skewed ( $0.780 / 0.183 = 4.262$ , which is certainly greater than 2).

The final measure that is sometimes referred to, though very rarely in practice, is the **kurtosis** of a data set. This is illustrated in Figure 4.11. This term was always one that confused me as a student. If you do a brief search in the literature, it seems as though there is disagreement over what the term “kurtosis” refers to. Some argue that kurtosis tells you something about the “pointiness” of a distribution. After all kurtosis is derived from a Greek “Kurtos” that means curvature. In fact (and if you do not believe me I invite you to read this rather impassioned article by Westfall (2014)) kurtosis actually tells us something about the “tails” or “outliers” of a distribution. The normal distribution has a kurtosis value of zero. We call this mesokurtic. Platykurtic distributions have thinner tails (or less outliers) than the normal distribution, while leptokurtic distributions have fatter tails. It is true that often platykurtic distributions have a “flat peak” and leptokurtic distributions have a “pointy peak”, however, it is possible to contrive plenty of distributions that do not exhibit this property. So it is final. If anyone asks you about kurtosis, you will think about tails!

The values of kurtosis are summarised in the table below:

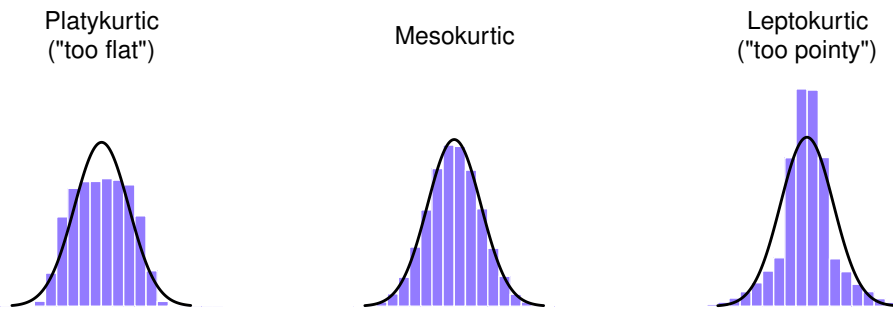


Figure 4.11: An illustration of kurtosis. On the left, we have a “platykurtic” data set (kurtosis =  $-.95$ ). Notice the thin tail. In the middle we have a “mesokurtic” data set (kurtosis is almost exactly 0). Finally, on the right, we have a “leptokurtic” data set (kurtosis =  $2.12$ ). Notice the fat tail. Note that the black line shows the normal distribution for reference.

.....

technical name	tail	kurtosis value
platykurtic	thin	negative
mesokurtic	about right	zero
leptokurtic	heavy	positive

The equation for kurtosis is pretty similar in spirit to the formulas we’ve seen already for the variance and the skewness. Except that where the variance involved squared deviations and the skewness involved cubed deviations, the kurtosis involves raising the deviations to the fourth power:<sup>a</sup>

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

<sup>a</sup>The “ $-3$ ” part is something that statisticians tack on to ensure that the normal curve has kurtosis zero. It looks a bit silly, just sticking a “ $-3$ ” at the end of the formula, but there are good mathematical reasons for doing this.

For completeness, the keen eyed amongst you would have noticed a similarity in the equations of the mean, variance, skewness, and kurtosis. In fact these terms are all related to a mathematical concept of “moments of a distribution”. There are indeed higher moments of a distribution that are needed to describe certain distributions but this is beyond the scope of this book.

More to the point, JASP has a check box for kurtosis just below the check box for skewness, and this gives a value for kurtosis of 0.101 with a standard error of 0.364. Using the same idea as before where we divide the kurtosis by the standard error of the kurtosis, we note that this value is much less than 2 ( $0.101/0.364 = 0.277$ ). This means that the AFL winning margins data are mesokurtic enough.

#### 4.4

---

### **Descriptive statistics separately for each group**

It is very commonly the case that you find yourself needing to look at descriptive statistics broken down by some grouping variable. This is pretty easy to do in JASP. For instance, let's say I want to look at the descriptive statistics for some `clin.trial` data, broken down separately by `therapy` type. This is a new data set, one that you've never seen before. The data is stored in the `clinicaltrial.csv` file and we'll use it a lot in Chapter ?? (you can find a complete description of the data at the start of that chapter). Let's load it and see what we've got:

Evidently there were three drugs: a placebo, something called "anxifree" and something called "joyzepam", and there were 6 people administered each drug. There were 9 people treated using cognitive behavioural therapy (CBT) and 9 people who received no psychological treatment. And we can see from looking at the 'Descriptives' of the `mood.gain` variable that most people did show a mood gain (mean = 0.88), though without knowing what the scale is here it's hard to say much more than that. Still, that's not too bad. Overall I feel that I learned something from that.

We can also go ahead and look at some other descriptive statistics, and this time separately for each type of therapy. In JASP, check Std. deviation, Skewness, and Kurtosis in the 'Statistics' options. At the same time, transfer the `therapy` variable into the 'Split' box, and you should get something like Figure 4.13

#### 4.5

---

### **Standard scores**

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

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

Figure 4.12: A screenshot of JASP showing the variables stored in the `clinicaltrial.csv` file

the standard deviation is 5. In contrast, when I take the questionnaire I answer 35 out of 50 questions in a grumpy way. So, how grumpy am I? One way to think about it would be to say that I have grumpiness of 35/50, so you might say that I'm 70% grumpy. But that's a bit weird, when you think about it. If my friend had phrased her questions a bit differently people might have answered them in a different way, so the overall distribution of answers could easily move up or down depending on the precise way in which the questions were asked. So, I'm only 70% grumpy *with respect to this set of survey questions*. Even if it's a very good questionnaire this isn't very a informative statement.

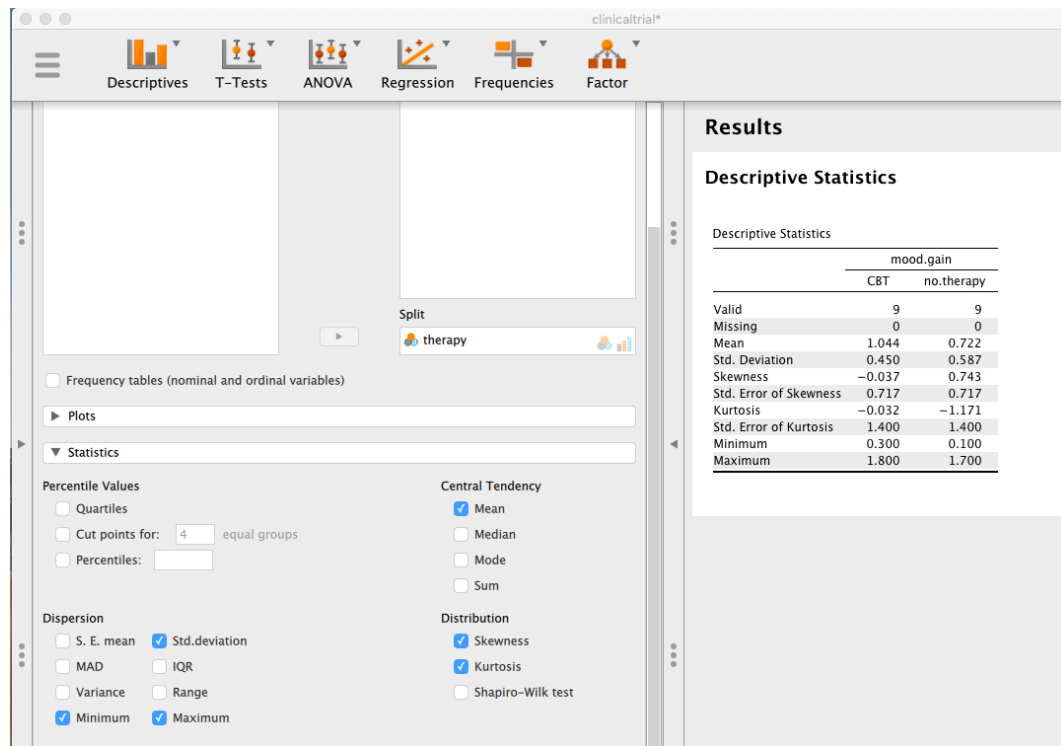


Figure 4.13: A screenshot of JASP showing Descriptives split by therapy type

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

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

maths" the standard score is calculated like this:

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

In actual maths, the equation for the z-score is

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

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

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

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

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

What about the standard scores? Well, this is a little different. If we calculate the standard scores we get  $z = (35 - 17)/5 = 3.6$  for grumpiness and  $z = (2 - 13)/4 = -2.75$  for extraversion. These two numbers *can* be compared to each other.<sup>3</sup> I'm much less extraverted than most people ( $z = -2.75$ ) and much grumpier than most people ( $z = 3.6$ ). But the extent of my unusualness is much more extreme for grumpiness, since 3.6 is a bigger number than 2.75. Because each standardised score is a statement about where an observation falls *relative to its own population*, it is possible to compare standardised scores across completely different variables.

---

<sup>3</sup>Though some caution is usually warranted. It's not always the case that one standard deviation on variable A corresponds to the same "kind" of thing as one standard deviation on variable B. Use common sense when trying to determine whether or not the z scores of two variables can be meaningfully compared.

## Summary

Calculating some basic descriptive statistics is one of the very first things you do when analysing real data, and descriptive statistics are much simpler to understand than inferential statistics, so like every other statistics textbook I've started with descriptives. In this chapter, we talked about the following topics:

- *Measures of central tendency.* Broadly speaking, central tendency measures tell you where the data are. There's three measures that are typically reported in the literature: the mean, median and mode. (Section 4.1)
- *Measures of variability.* In contrast, measures of variability tell you about how "spread out" the data are. The key measures are: range, standard deviation, and interquartile range. (Section 4.2)
- *Measures of skewness and kurtosis.* We also looked at assymetry in a variable's distribution (skew) and pointness (kurtosis). (Section 4.3)
- *Getting group summaries of variables in JASP.* Since this book focuses on doing data analysis in JASP, we spent a bit of time talking about how descriptive statistics are computed for different subgroups. (Section 4.4)
- *Standard scores.* The z-score is a slightly unusual beast. It's not quite a descriptive statistic, and not quite an inference. We talked about it in Section 4.5. Make sure you understand that section. It'll come up again later.

In the next Chapter we'll move on to a discussion of how to draw pictures! Everyone loves a pretty picture, right? But before we do, I want to end on an important point. A traditional first course in statistics spends only a small proportion of the class on descriptive statistics, maybe one or two lectures at most. The vast majority of the lecturer's time is spent on inferential statistics because that's where all the hard stuff is. That makes sense, but it hides the practical everyday importance of choosing good descriptives. With that in mind. . .

### 4.6.1 Epilogue: Good descriptive statistics are descriptive!

*The death of one man is a tragedy.*

*The death of millions is a statistic.*

– Josef Stalin, Potsdam 1945



950,000 – 1,200,000

– Estimate of Soviet repression deaths,  
1937-1938 (Ellman 2002)

Stalin's infamous quote about the statistical character of the deaths of millions is worth giving some thought. The clear intent of his statement is that the death of an individual touches us personally and its force cannot be denied, but that the deaths of a multitude are incomprehensible and as a consequence are mere statistics and more easily ignored. I'd argue that Stalin was half right. A statistic is an abstraction, a description of events beyond our personal experience, and so hard to visualise. Few if any of us can imagine what the deaths of millions is "really" like, but we can imagine one death and this gives the lone death its feeling of immediate tragedy, a feeling that is missing from Ellman's cold statistical description.

Yet it is not so simple. Without numbers, without counts, without a description of what happened, we have *no chance* of understanding what really happened, no opportunity even to try to summon the missing feeling. And in truth, as I write this sitting in comfort on a Saturday morning half a world and a whole lifetime away from the Gulags, when I put the Ellman estimate next to the Stalin quote a dull dread settles in my stomach and a chill settles over me. The Stalinist repression is something truly beyond my experience, but with a combination of statistical data and those recorded personal histories that have come down to us, it is not entirely beyond my comprehension. Because what Ellman's numbers tell us is this: over a two year period Stalinist repression wiped out the equivalent of every man, woman and child currently alive in the city where I live. Each one of those deaths had it's own story, was it's own tragedy, and only some of those are known to us now. Even so, with a few carefully chosen statistics, the scale of the atrocity starts to come into focus.

Thus it is no small thing to say that the first task of the statistician and the scientist is to summarise the data, to find some collection of numbers that can convey to an audience a sense of what has happened. This is the job of descriptive statistics, but it's not a job that can be told solely using the numbers. You are a data analyst, and not a statistical software package. Part of your job is to take these *statistics* and turn them into a *description*. When you analyse data it is not sufficient to list off a collection of numbers. Always remember that what you're really trying to do is communicate with a human audience. The numbers are important, but they need to be put together into a meaningful story that your audience can interpret. That means you need to think about framing. You need to think about context. And you need to think about the individual events that your statistics are summarising.

Part III.

## **Statistical theory**



---

## Prelude to Part IV

Part IV of the book is by far the most theoretical, focusing as it does on the theory of statistical inference. Over the next three chapters my goal is to give you an introduction to probability theory (Chapter ??), sampling and estimation (Chapter 5) and statistical hypothesis testing (Chapter 6). Before we get started though, I want to say something about the big picture. Statistical inference is primarily about *learning from data*. The goal is no longer merely to describe our data but to use the data to draw conclusions about the world. To motivate the discussion I want to spend a bit of time talking about a philosophical puzzle known as the *riddle of induction*, because it speaks to an issue that will pop up over and over again throughout the book: statistical inference relies on *assumptions*. This sounds like a bad thing. In everyday life people say things like “you should never make assumptions”, and psychology classes often talk about assumptions and biases as bad things that we should try to avoid. From bitter personal experience I have learned never to say such things around philosophers!

### On the limits of logical reasoning

*The whole art of war consists in getting at what is on the other side of the hill,  
or, in other words, in learning what we do not know from what we do.*

– Arthur Wellesley, 1st Duke of Wellington

I am told that quote above came about as a consequence of a carriage ride across the countryside.<sup>4</sup> He and his companion, J. W. Croker, were playing a guessing game, each trying to predict what would be on the other side of each hill. In every case it turned out that Wellesley was right and Croker was wrong. Many years later when Wellesley was asked about the game he explained that “the whole art of war consists in getting at what is on the other side of the hill”. Indeed, war is not special in this respect. All of life is a guessing game of one form or another, and getting by on a day to day basis requires us to make good guesses. So let’s play a guessing game of our own.

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

WWW

---

<sup>4</sup>Source: <http://www.bartleby.com/344/400.html>.

Our conversation goes like this:

you: Three in a row doesn't mean much. I suppose Wellesley might be better at this than Croker, but it might just be luck. Still, I'm a bit of a gambler. I'll bet on Wellesley.

me: I agree that three in a row isn't informative and I see no reason to prefer Wellesley's guesses over Croker's. I can't justify betting at this stage. Sorry. No bet for me.

Your gamble paid off: three more hills go by and Wellesley wins all three. Going into the next round of our game the score is 1-0 in favour of you and our data set looks like this:

WWW WWW

I've organised the data into blocks of three so that you can see which batch corresponds to the observations that we had available at each step in our little side game. After seeing this new batch, our conversation continues:

you: Six wins in a row for Duke Wellesley. This is starting to feel a bit suspicious. I'm still not certain, but I reckon that he's going to win the next one too.

me: I guess I don't see that. Sure, I agree that Wellesley has won six in a row, but I don't see any logical reason why that means he'll win the seventh one. No bet.

you: Do you really think so? Fair enough, but my bet worked out last time and I'm okay with my choice.

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

WWW WWW WWW

And our conversation goes like this:

you: Okay, this is pretty obvious. Wellesley is way better at this game. We both agree he's going to win the next hill, right?

me: Is there really any logical evidence for that? Before we started this game, there were lots of possibilities for the first 10 outcomes, and I had no idea which one to expect. WWW WWW WWW W was one possibility,

but so was WCC CWC WWC C and WWW WWW WWW C or even CCC CCC CCC C. Because I had no idea what would happen so I'd have said they were all equally likely. I assume you would have too, right? I mean, that's what it *means* to say you have "no idea", isn't it?

you: I suppose so.

me: Well then, the observations we've made logically rule out all possibilities except two: WWW WWW WWW C or WWW WWW WWW W. Both of these are perfectly consistent with the evidence we've encountered so far, aren't they?

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

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

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

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

WWW WWW WWW WWW

and the conversation continues:

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

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

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

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

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

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

### **Learning without making assumptions is a myth**

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

The riddle of induction is most associated with the philosophical work of David Hume and more recently Nelson Goodman, but you can find examples of the problem popping up in fields as diverse as literature (Lewis Carroll) and machine learning (the "no free lunch" theorem). There really is something weird about trying to "learn what we do not know from what we do know". The critical point is that assumptions and biases are unavoidable if you want to learn anything about the world. There is no escape from this, and it is just as true for statistical inference as it is for human reasoning. In the dialogue I was taking aim at your perfectly sensible inferences as a human being, but the common sense reasoning that you relied on is no different to what a statistician would have done. Your "common sense" half of the dialog relied on an implicit *assumption* that there exists some difference in skill between Wellesley

and Croker, and what you were doing was trying to work out what that difference in skill level would be. My “logical analysis” rejects that assumption entirely. All I was willing to accept is that there are sequences of wins and losses and that I did not know which sequences would be observed. Throughout the dialogue I kept insisting that all logically possible data sets were equally plausible at the start of the Wellesely-Croker game, and the only way in which I ever revised my beliefs was to eliminate those possibilities that were factually inconsistent with the observations.

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

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





## 5. Estimating unknown quantities from a sample

---

At the start of the last chapter I highlighted the critical distinction between *descriptive statistics* and *inferential statistics*. As discussed in Chapter 4, the role of descriptive statistics is to concisely summarise what we *do* know. In contrast, the purpose of inferential statistics is to “learn what we do not know from what we do”. Now that we have a foundation in probability theory we are in a good position to think about the problem of statistical inference. What kinds of things would we like to learn about? And how do we learn them? These are the questions that lie at the heart of inferential statistics, and they are traditionally divided into two “big ideas”: estimation and hypothesis testing. The goal in this chapter is to introduce the first of these big ideas, estimation theory, but I’m going to witter on about sampling theory first because estimation theory doesn’t make sense until you understand sampling. As a consequence, this chapter divides naturally into two parts Sections 5.1 through 5.3 are focused on sampling theory, and Sections 5.4 and 5.5 make use of sampling theory to discuss how statisticians think about estimation.

### 5.1

---

#### **Samples, populations and sampling**

In the prelude to Part III I discussed the riddle of induction and highlighted the fact that *all* learning requires you to make assumptions. Accepting that this is true, our first task to come up with some fairly general assumptions about data that make sense. This is where **sampling theory** comes in. If probability theory is the foundations upon which all statistical theory builds, sampling theory is the frame around which you can build the rest of the house. Sampling theory plays a huge role in specifying the assumptions upon which your statistical inferences rely. And in order to talk about “making inferences” the way statisticians think about it we need to be a bit more explicit about what it is that we’re drawing inferences *from* (the sample) and what it is that we’re drawing inferences *about* (the population).

In almost every situation of interest what we have available to us as researchers is a **sample** of data. This could be a subset of patients enrolled in a clinical trial, blood samples from a

specific population, or medical records from a certain hospital. Given the limitations of time, resources, and ethical considerations, it's impossible to include every individual or data point relevant to our research question. For instance, it's unfeasible to test a new drug on every patient in the world who has a particular condition. In our earlier discussion of descriptive statistics (Chapter 4) this sample was the only thing we were interested in. Our only goal was to find ways of describing, summarising and graphing that sample. This is about to change.

### 5.1.1 Defining a population

A sample is a concrete thing. You can open up a data file and there's the data from your sample. A **population**, on the other hand, is a more abstract idea. It refers to the set of all possible people, or all possible observations, that you want to draw conclusions about and is generally *much* bigger than the sample. In an ideal world the researcher would begin the study with a clear idea of what the population of interest is, since the process of designing a study and testing hypotheses with the data does depend on the population about which you want to make statements.

Defining the target population can also be challenging. Consider a study where you investigate the effectiveness of a new antihypertensive drug using a sample of 100 patients from a specific hospital. Your ultimate aim, as a researcher in medicine, may be to generalize your findings to a broader context. In this situation, which of the following would qualify as "the population":

- All hypertension patients at the hospital where the study was conducted?
- All hypertension patients across all hospitals in the country?
- Adults currently living in the country?
- Adults worldwide who have hypertension?
- All adults, regardless of their hypertension status?
- All human beings, irrespective of their age, geography, or health condition?
- Any mammal susceptible to hypertension?
- Any organism with a cardiovascular system?

Each of these defines a real group of mind-possessing entities, all of which might be of interest to me as a medical scientist, and it's not at all clear which one ought to be the true population of interest. As another example, consider the Wellesley-Croker game that we discussed in the prelude. The sample here is a specific sequence of 12 wins and 0 losses for Wellesley. What is the population?

- All outcomes until Wellesley and Croker arrived at their destination?
- All outcomes if Wellesley and Croker had played the game for the rest of their lives?
- All outcomes if Wellesley and Croker lived forever and played the game until the world ran out of hills?

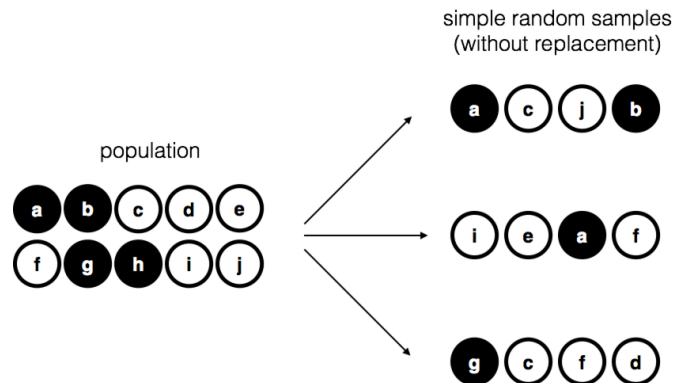


Figure 5.1: Simple random sampling without replacement from a finite population

- All outcomes if we created an infinite set of parallel universes and the Welleseley/Croker pair made guesses about the same 12 hills in each universe?

Again, it's not obvious what the population is.

#### 5.1.2 Simple random samples

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

To keep things simple, let's imagine that we have a bag containing 10 chips. Each chip has a unique letter printed on it so we can distinguish between the 10 chips. The chips come in two colours, black and white. This set of chips is the population of interest and it is depicted graphically on the left of Figure 5.1. As you can see from looking at the picture there are 4 black chips and 6 white chips, but of course in real life we wouldn't know that unless we looked in the bag. Now imagine you run the following "experiment": you shake up the bag, close your eyes, and pull out 4 chips without putting any of them back into the bag. First out comes the *a* chip (black), then the *c* chip (white), then *j* (white) and then finally *b* (black). If you wanted you could then put all the chips back in the bag and repeat the experiment, as depicted on the right hand side of Figure 5.1. Each time you get different results but the procedure is identical in each case. The fact that the same procedure can lead to different results each

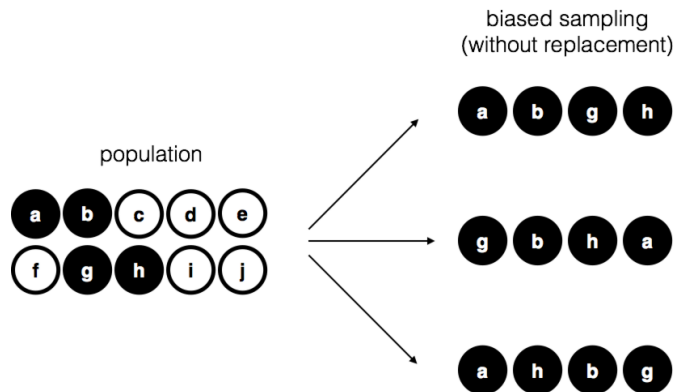


Figure 5.2: Biased sampling without replacement from a finite population

time we refer to as a *random* process.<sup>1</sup> However, because we shook the bag before pulling any chips out, it seems reasonable to think that every chip has the same chance of being selected. A procedure in which every member of the population has the same chance of being selected is called a **simple random sample**. The fact that we did *not* put the chips back in the bag after pulling them out means that you can't observe the same thing twice, and in such cases the observations are said to have been sampled **without replacement**.

To help make sure you understand the importance of the sampling procedure, consider an alternative way in which the experiment could have been run. Suppose that my 5-year old son had opened the bag and decided to pull out four black chips without putting any of them back in the bag. This *biased* sampling scheme is depicted in Figure 5.2. Now consider the evidential value of seeing 4 black chips and 0 white chips. Clearly it depends on the sampling scheme, does it not? If you know that the sampling scheme is biased to select only black chips then a sample that consists of only black chips doesn't tell you very much about the population! For this reason statisticians really like it when a data set can be considered a simple random sample, because it makes the data analysis *much* easier.

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

<sup>1</sup>The proper mathematical definition of randomness is extraordinarily technical, and way beyond the scope of this book. We'll be non-technical here and say that a process has an element of randomness to it whenever it is possible to repeat the process and get different answers each time.

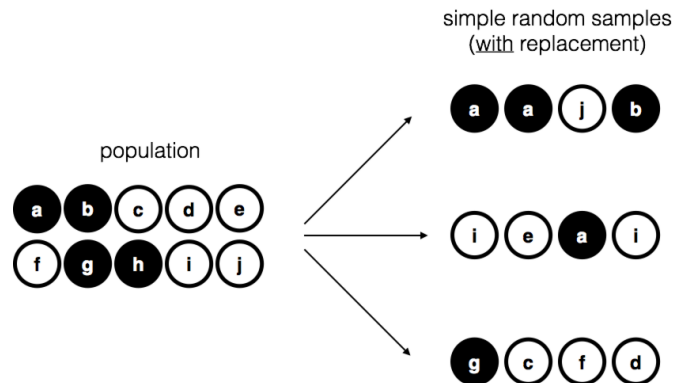


Figure 5.3: Simple random sampling *with* replacement from a finite population

in Figure 5.3.

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

### 5.1.3 Most samples are not simple random samples

As you can see from looking at the list of possible populations that I showed above, it is almost impossible to obtain a simple random sample from most populations of interest. For instance, in a clinical trial, achieving a random sample from all eligible patients within a specific hospital would already be a considerable feat, let alone expanding the scope to include a broader demographic or geographic range. A thorough discussion of other types of sampling schemes is beyond the scope of this book, but to give you a sense of what's out there I'll list a few of the more important ones.

- *Stratified sampling.* Suppose your population is (or can be) divided into several different sub-populations, or *strata*. Perhaps you're running a study at several different sites, for example. Instead of trying to sample randomly from the population as a whole, you instead try to collect a separate random sample from each of the strata. Stratified sampling is sometimes easier to do than simple random sampling, especially when the population is already divided into the distinct strata. It can also be more efficient than simple random sampling, especially when some of the sub-populations are rare. For instance, when studying schizophrenia it would be much better to divide the population

into two<sup>2</sup> strata (schizophrenic and not-schizophrenic) and then sample an equal number of people from each group. If you selected people randomly you would get so few schizophrenic people in the sample that your study would be useless. This specific kind of stratified sampling is referred to as *oversampling* because it makes a deliberate attempt to over-represent rare groups.

- *Snowball sampling* is a technique that is especially useful when sampling from a “hidden” or hard to access population and is especially common in social sciences. For instance, suppose the researchers want to conduct an opinion poll among transgender people. The research team might only have contact details for a few trans folks, so the survey starts by asking them to participate (stage 1). At the end of the survey the participants are asked to provide contact details for other people who might want to participate. In stage 2 those new contacts are surveyed. The process continues until the researchers have sufficient data. The big advantage to snowball sampling is that it gets you data in situations that might otherwise be impossible to get any. On the statistical side, the main disadvantage is that the sample is highly non-random, and non-random in ways that are difficult to address. On the real life side, the disadvantage is that the procedure can be unethical if not handled well, because hidden populations are often hidden for a reason. I chose transgender people as an example here to highlight this issue. If you weren’t careful you might end up outing people who don’t want to be outed (very, very bad form), and even if you don’t make that mistake it can still be intrusive to use people’s social networks to study them. It’s certainly very hard to get people’s informed consent *before* contacting them, yet in many cases the simple act of contacting them and saying “hey we want to study you” can be hurtful. Social networks are complex things, and just because you can use them to get data doesn’t always mean you should.
- *Convenience sampling* is more or less what it sounds like. The samples are chosen in a way that is convenient to the researcher, and not selected at random from the population of interest. Snowball sampling is one type of convenience sampling, but there are many others. A common example in student research are studies that rely on fellow students. These samples are generally non-random in two respects. First, reliance on undergraduate students automatically means that your data are restricted to a single sub-population. Second, the students usually get to pick which studies they participate in, so the sample is a self selected subset of students and not a randomly selected subset. In real life most studies are convenience samples of one form or another. This is sometimes a severe limitation, but not always.

---

<sup>2</sup>Nothing in life is that simple. There’s not an obvious division of people into binary categories like “schizophrenic” and “not schizophrenic”. But this isn’t a clinical psychiatry text so please forgive me a few simplifications here and there.

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

Okay, so real world data collection tends not to involve nice simple random samples. Does that matter? A little thought should make it clear to you that it *can* matter if your data are not a simple random sample. Just think about the difference between Figures 5.1 and 5.2. However, it's not quite as bad as it sounds. Some types of biased samples are entirely unproblematic. For instance, when using a stratified sampling technique you actually *know* what the bias is because you created it deliberately, often to *increase* the effectiveness of your study, and there are statistical techniques that you can use to adjust for the biases you've introduced (not covered in this book!). So in those situations it's not a problem.

More generally though, it's important to remember that random sampling is a means to an end, and not the end in itself. Let's assume you've relied on a convenience sample, and as such you can assume it's biased. A bias in your sampling method is only a problem if it causes you to draw the wrong conclusions. When viewed from that perspective, I'd argue that we don't need the sample to be randomly generated in *every* respect, we only need it to be random with respect to the relevant phenomenon of interest. Suppose you're conducting two different studies focusing on the efficacy of a new diabetes management protocol. In Study 1, you have the ability to randomly sample from all individuals worldwide with diabetes, with one exception: you can only include patients who were diagnosed on a Monday. In Study 2, you can sample randomly from diabetic patients within the country of Malta. If the objective is to generalize the results to all individuals with diabetes globally, then Study 1 would be the better choice. Why? There's little reason to think that being "diagnosed on a Monday" would have a meaningful impact on one's response to a diabetes management protocol. Conversely, being Maltese could introduce certain variables such as access to healthcare, lifestyle factors, genetics, or existing co-morbidities, which could differ from those in the broader global population. These variables could influence the effectiveness of the new diabetes management protocol, potentially leading to biased results that aren't universally applicable. Therefore, carefully considering the specifics of your sample is crucial in medical research for ensuring the broad applicability of your findings.

There are two points hidden in this discussion. First, when designing your own studies, it's important to think about what population you care about and try hard to sample in a way that is appropriate to that population. In practice, you're usually forced to put up with a "sample of convenience" (e.g., students tend to sample other students because that's the least expensive way to collect data), but if so you should at least spend some time thinking about what the dangers of this practice might be. Second, if you're going to criticise someone else's study because they've used a sample of convenience rather than laboriously sampling randomly from the entire human population, at least have the courtesy to offer a specific theory as to *how* this might have distorted the results.

#### 5.1.5 Population parameters and sample statistics

Okay. Setting aside the thorny methodological issues associated with obtaining a random



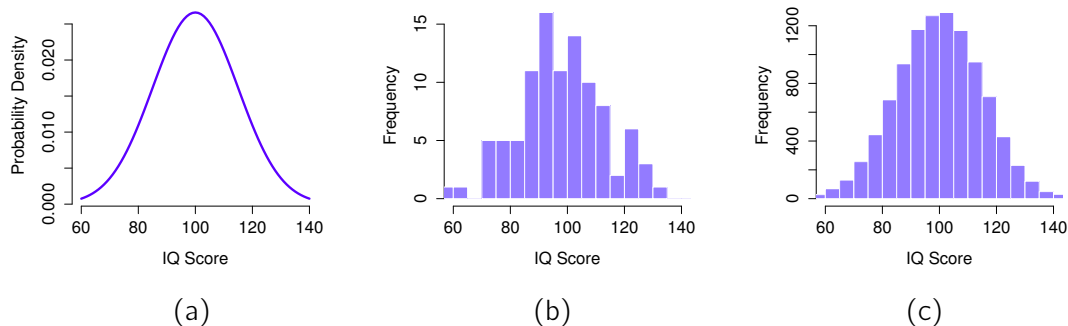


Figure 5.4: The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and panel c we have a sample of 10,000 observations.

.....

sample, let's consider a slightly different issue. Up to this point we have been talking about populations the way a scientist might. To a medical researcher a population might be a group of patients. To an ecologist a population might be a group of bears. In most cases the populations that scientists care about are concrete things that actually exist in the real world. Statisticians, however, are a funny lot. On the one hand, they *are* interested in real world data and real science in the same way that scientists are. On the other hand, they also operate in the realm of pure abstraction in the way that mathematicians do. As a consequence, statistical theory tends to be a bit abstract in how a population is defined. In much the same way that scientists operationalise abstract theoretical ideas in terms of concrete measurements (Section 2.1), statisticians operationalise the concept of a “population” in terms of mathematical objects that they know how to work with. You've already come across these objects in Chapter ???. They're called probability distributions.

The idea is quite simple. Let's say we're talking about IQ scores. To a psychiatrist the population of interest is a group of actual humans who have IQ scores. A statistician “simplifies” this by operationally defining the population as the probability distribution depicted in Figure 5.4a. IQ tests are designed so that the average IQ is 100, the standard deviation of IQ scores is 15, and the distribution of IQ scores is normal. These values are referred to as the **population parameters** because they are characteristics of the entire population. That is, we say that the population mean  $\mu$  is 100 and the population standard deviation  $\sigma$  is 15.

Now suppose I run an experiment. I select 100 people at random and administer an IQ test, giving me a simple random sample from the population. My sample would consist of a collection of numbers like this:

106 101 98 80 74 ... 107 72 100

Each of these IQ scores is sampled from a normal distribution with mean 100 and standard deviation 15. So if I plot a histogram of the sample I get something like the one shown in Figure 5.4b. As you can see, the histogram is *roughly* the right shape but it's a very crude approximation to the true population distribution shown in Figure 5.4a. When I calculate the mean of my sample, I get a number that is fairly close to the population mean 100 but not identical. In this case, it turns out that the people in my sample have a mean IQ of 98.5, and the standard deviation of their IQ scores is 15.9. These **sample statistics** are properties of my data set, and although they are fairly similar to the true population values they are not the same. In general, sample statistics are the things you can calculate from your data set and the population parameters are the things you want to learn about. Later on in this chapter I'll talk about how you can estimate population parameters using your sample statistics (Section 5.4) and how to work out how confident you are in your estimates (Section 5.5) but before we get to that there's a few more ideas in sampling theory that you need to know about.

## 5.2

---

### The law of large numbers

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

The obvious answer is to collect more data. Suppose that we ran a much larger experiment, this time measuring the IQs of 10,000 people. We can simulate the results of this experiment using JASP. The `IQsim.jasp` file is a JASP data file. In this file I have generated 10,000 random numbers sampled from a normal distribution for a population with `mean = 100` and `sd = 15`. By the way, I did this entirely within JASP computing a new variable using the R code `rnorm(10000, 100, 15)`. A histogram and density plot shows that this larger sample is a much better approximation to the true population distribution than the smaller one. This is reflected in the sample statistics. The mean IQ for the larger sample turns out to be 100.107 and the standard deviation is 14.995. These values are now very close to the true population. See Figure 5.5

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

*For even the most stupid of men, by some instinct of nature, by himself and*

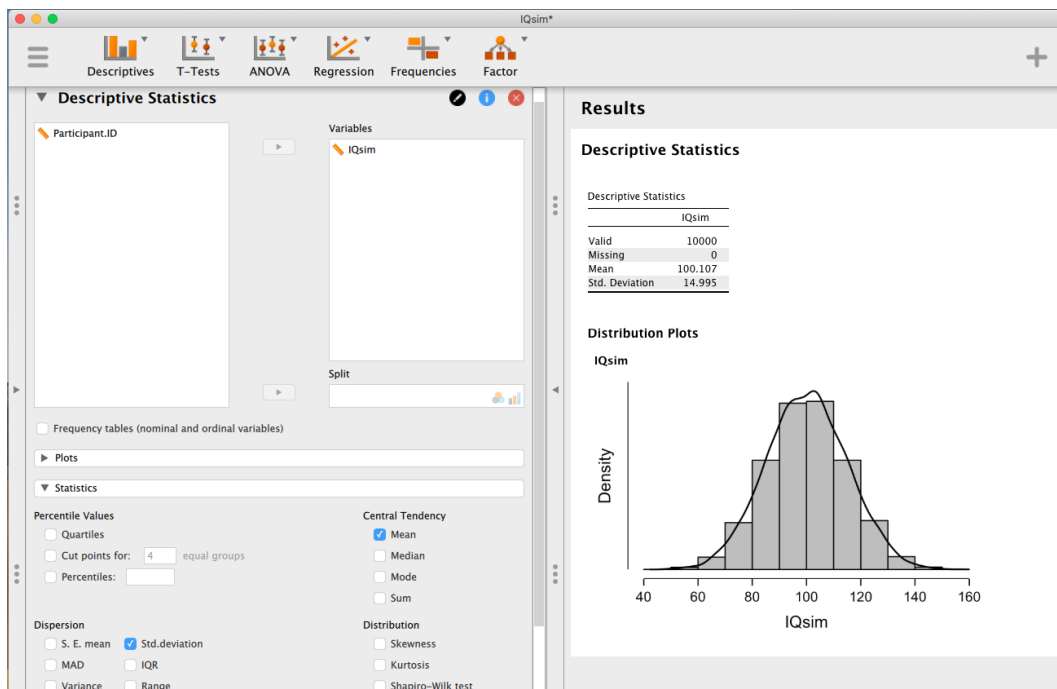


Figure 5.5: A random sample drawn from a normal distribution using JASP

*without any instruction (which is a remarkable thing), is convinced that the more observations have been made, the less danger there is of wandering from one's goal (see Stigler 1986, p65)*

Okay, so the passage comes across as a bit condescending (not to mention sexist), but his main point is correct. It really does feel obvious that more data will give you better answers. The question is, why is this so? Not surprisingly, this intuition that we all share turns out to be correct, and statisticians refer to it as the **law of large numbers**. The law of large numbers is a mathematical law that applies to many different sample statistics but the simplest way to think about it is as a law about averages. The sample mean is the most obvious example of a statistic that relies on averaging (because that's what the mean is... an average), so let's look at that. When applied to the sample mean what the law of large numbers states is that as the sample gets larger, the sample mean tends to get closer to the true population mean. Or, to say it a little bit more precisely, as the sample size "approaches" infinity (written as  $N \rightarrow \infty$ ), the sample mean approaches the population mean ( $\bar{X} \rightarrow \mu$ ).<sup>3</sup>

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

### 5.3

---

#### **Sampling distributions and the central limit theorem**

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

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

---

<sup>3</sup>Technically, the law of large numbers pertains to any sample statistic that can be described as an average of independent quantities. That's certainly true for the sample mean. However, it's also possible to write many other sample statistics as averages of one form or another. The variance of a sample, for instance, can be rewritten as a kind of average and so is subject to the law of large numbers. The minimum value of a sample, however, cannot be written as an average of anything and is therefore not governed by the law of large numbers.

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

### 5.3.1 Sampling distribution of the mean

With this in mind, let's abandon the idea that our studies will have sample sizes of 10,000 and consider instead a very modest experiment indeed. This time around we'll sample  $N = 5$  people and measure their IQ scores. As before, I can simulate this experiment in JASP by modifying the `rnorm` function that was used to generate the `IQsim` data column. If you double-click on the  $f_x$  label beside `IQsim`, JASP will open up the 'Computed Column' dialog, which contains the R code `rnorm(10000, 100, 15)`. Since I only need 5 participant IDs this time, I simply need to change 10000 to 5 and then click 'Compute column' (see Figure 5.6). These are the five numbers that JASP generated for me (yours will be different!). I rounded to the nearest whole number for convenience:

124 74 87 86 109

The mean IQ in this sample turns out to be exactly 96. Not surprisingly, this is much less accurate than the previous experiment. Now imagine that I decided to **replicate** the experiment. That is, I repeat the procedure as closely as possible and I randomly sample 5 new people and measure their IQ. Again, JASP allows me to simulate the results of this procedure, and generates these five numbers:

91 125 104 106 109

This time around, the mean IQ in my sample is 107. If I repeat the experiment 10 times I obtain the results shown in Table 5.1, and as you can see the sample mean varies from one replication to the next.

Now suppose that I decided to keep going in this fashion, replicating this "five IQ scores" experiment over and over again. Every time I replicate the experiment I write down the sample mean. Over time, I'd be amassing a new data set, in which every experiment generates a single data point. The first 10 observations from my data set are the sample means listed in Table 5.1, so my data set starts out like this:

96.0 107.0 101.6 103.8 104.4 ...

What if I continued like this for 10,000 replications, and then drew a histogram. Well that's exactly what I did, and you can see the results in Figure 5.7. As this picture illustrates,

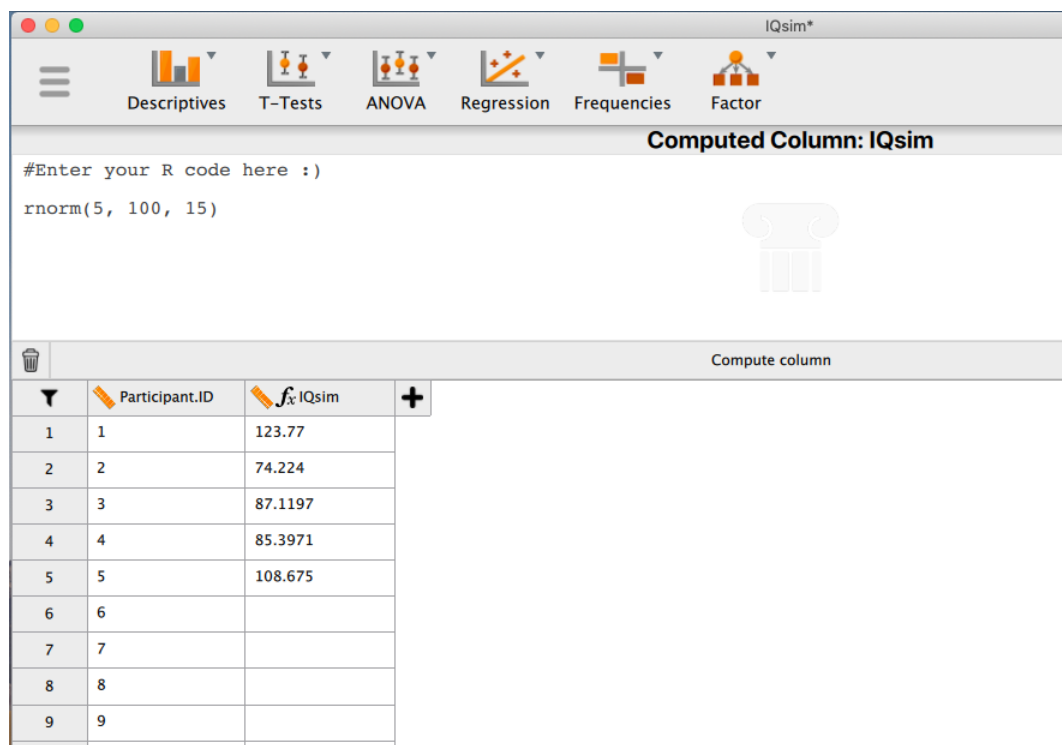


Figure 5.6: Using JASP to draw a random sample of 5 from a normal distribution with  $\mu = 100$  and  $\sigma = 15$ .

Table 5.1: Ten replications of the IQ experiment, each with a sample size of  $N = 5$ .

	Person 1	Person 2	Person 3	Person 4	Person 5	Sample Mean
Replication 1	124	74	87	86	109	96.0
Replication 2	91	125	104	106	109	107.0
Replication 3	111	122	91	98	86	101.6
Replication 4	98	96	119	99	107	103.8
Replication 5	105	113	103	103	98	104.4
Replication 6	81	89	93	85	114	92.4
Replication 7	100	93	108	98	133	106.4
Replication 8	107	100	105	117	85	102.8
Replication 9	86	119	108	73	116	100.4
Replication 10	95	126	112	120	76	105.8

the average of 5 IQ scores is usually between 90 and 110. But more importantly, what it highlights is that if we replicate an experiment over and over again, what we end up with is a *distribution* of sample means! This distribution has a special name in statistics, it's called the **sampling distribution of the mean**.

Sampling distributions are another important theoretical idea in statistics, and they're crucial for understanding the behaviour of small samples. For instance, when I ran the very first "five IQ scores" experiment, the sample mean turned out to be 96. What the sampling distribution in Figure 5.7 tells us, though, is that the "five IQ scores" experiment is not very accurate. If I repeat the experiment, the sampling distribution tells me that I can expect to see a sample mean anywhere between 80 and 120.

### 5.3.2 Sampling distributions exist for any sample statistic!

One thing to keep in mind when thinking about sampling distributions is that *any* sample statistic you might care to calculate has a sampling distribution. For example, suppose that each time I replicated the "five IQ scores" experiment I wrote down the largest IQ score in the experiment. This would give me a data set that started out like this:

124 125 122 119 113 ...

Doing this over and over again would give me a very different sampling distribution, namely the *sampling distribution of the maximum*. The sampling distribution of the maximum of 5 IQ scores is shown in Figure 5.8. Not surprisingly, if you pick 5 people at random and then find the person with the highest IQ score, they're going to have an above average IQ. Most of the time you'll end up with someone whose IQ is measured in the 100 to 140 range.

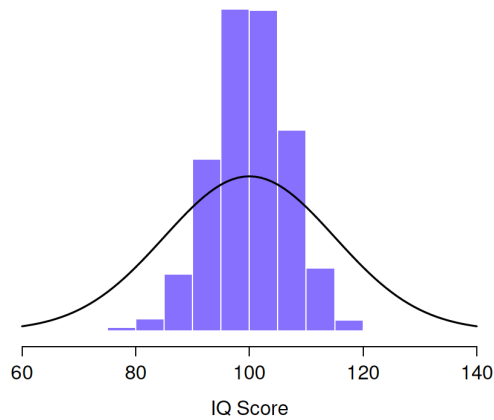


Figure 5.7: The sampling distribution of the mean for the “five IQ scores experiment”. If you sample 5 people at random and calculate their *average* IQ you’ll almost certainly get a number between 80 and 120, even though there are quite a lot of individuals who have IQs above 120 or below 80. For comparison, the black line plots the population distribution of IQ scores.

.....

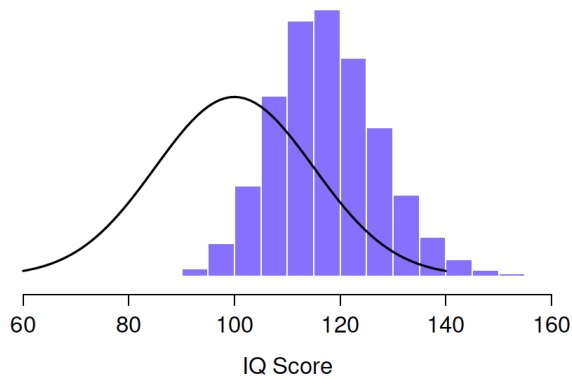


Figure 5.8: The sampling distribution of the *maximum* for the “five IQ scores experiment”. If you sample 5 people at random and select the one with the highest IQ score you’ll probably see someone with an IQ between 100 and 140.

.....



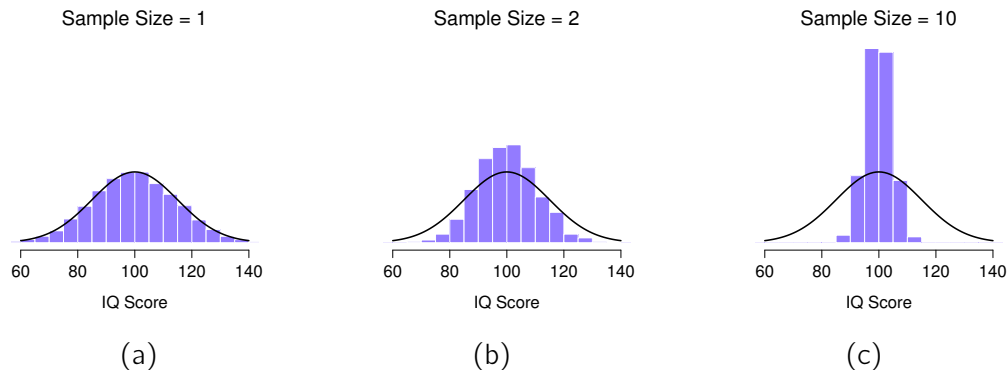


Figure 5.9: An illustration of the how sampling distribution of the mean depends on sample size. In each panel I generated 10,000 samples of IQ data and calculated the mean IQ observed within each of these data sets. The histograms in these plots show the distribution of these means (i.e., the sampling distribution of the mean). Each individual IQ score was drawn from a normal distribution with mean 100 and standard deviation 15, which is shown as the solid black line. In panel a, each data set contained only a single observation, so the mean of each sample is just one person’s IQ score. As a consequence, the sampling distribution of the mean is of course identical to the population distribution of IQ scores. However, when we raise the sample size to 2 the mean of any one sample tends to be closer to the population mean than a one person’s IQ score, and so the histogram (i.e., the sampling distribution) is a bit narrower than the population distribution. By the time we raise the sample size to 10 (panel c), we can see that the distribution of sample means tend to be fairly tightly clustered around the true population mean.

### 5.3.3 The central limit theorem

At this point I hope you have a pretty good sense of what sampling distributions are, and in particular what the sampling distribution of the mean is. In this section I want to talk about how the sampling distribution of the mean changes as a function of sample size. Intuitively, you already know part of the answer. If you only have a few observations, the sample mean is likely to be quite inaccurate. If you replicate a small experiment and recalculate the mean you’ll get a very different answer. In other words, the sampling distribution is quite wide. If you replicate a large experiment and recalculate the sample mean you’ll probably get the same answer you got last time, so the sampling distribution will be very narrow. You can see this visually in Figure 5.9, showing that the bigger the sample size, the narrower the sampling distribution gets. We can quantify this effect by calculating the standard deviation of the sampling distribution, which is referred to as the **standard error**. The standard error of a statistic is often denoted SE, and since we’re usually interested in the standard error of the

sample *mean*, we often use the acronym SEM. As you can see just by looking at the picture, as the sample size  $N$  increases, the SEM decreases.

Okay, so that's one part of the story. However, there's something I've been glossing over so far. All my examples up to this point have been based on the "IQ scores" experiments, and because IQ scores are roughly normally distributed I've assumed that the population distribution is normal. What if it isn't normal? What happens to the sampling distribution of the mean? The remarkable thing is this, no matter what shape your population distribution is, as  $N$  increases the sampling distribution of the mean starts to look more like a normal distribution. To give you a sense of this I ran some simulations. To do this, I started with the "ramped" distribution shown in the histogram in Figure 5.10. As you can see by comparing the triangular shaped histogram to the bell curve plotted by the black line, the population distribution doesn't look very much like a normal distribution at all. Next, I simulated the results of a large number of experiments. In each experiment I took  $N = 2$  samples from this distribution, and then calculated the sample mean. Figure 5.10b plots the histogram of these sample means (i.e., the sampling distribution of the mean for  $N = 2$ ). This time, the histogram produces a  $\cap$ -shaped distribution. It's still not normal, but it's a lot closer to the black line than the population distribution in Figure 5.10a. When I increase the sample size to  $N = 4$ , the sampling distribution of the mean is very close to normal (Figure 5.10c), and by the time we reach a sample size of  $N = 8$  it's almost perfectly normal. In other words, as long as your sample size isn't tiny, the sampling distribution of the mean will be approximately normal no matter what your population distribution looks like!

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

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

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

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

Because we divide the population standard deviation  $\sigma$  by the square root of the sample size  $N$ , the SEM gets smaller as the sample size increases. It also tells us that the shape of the

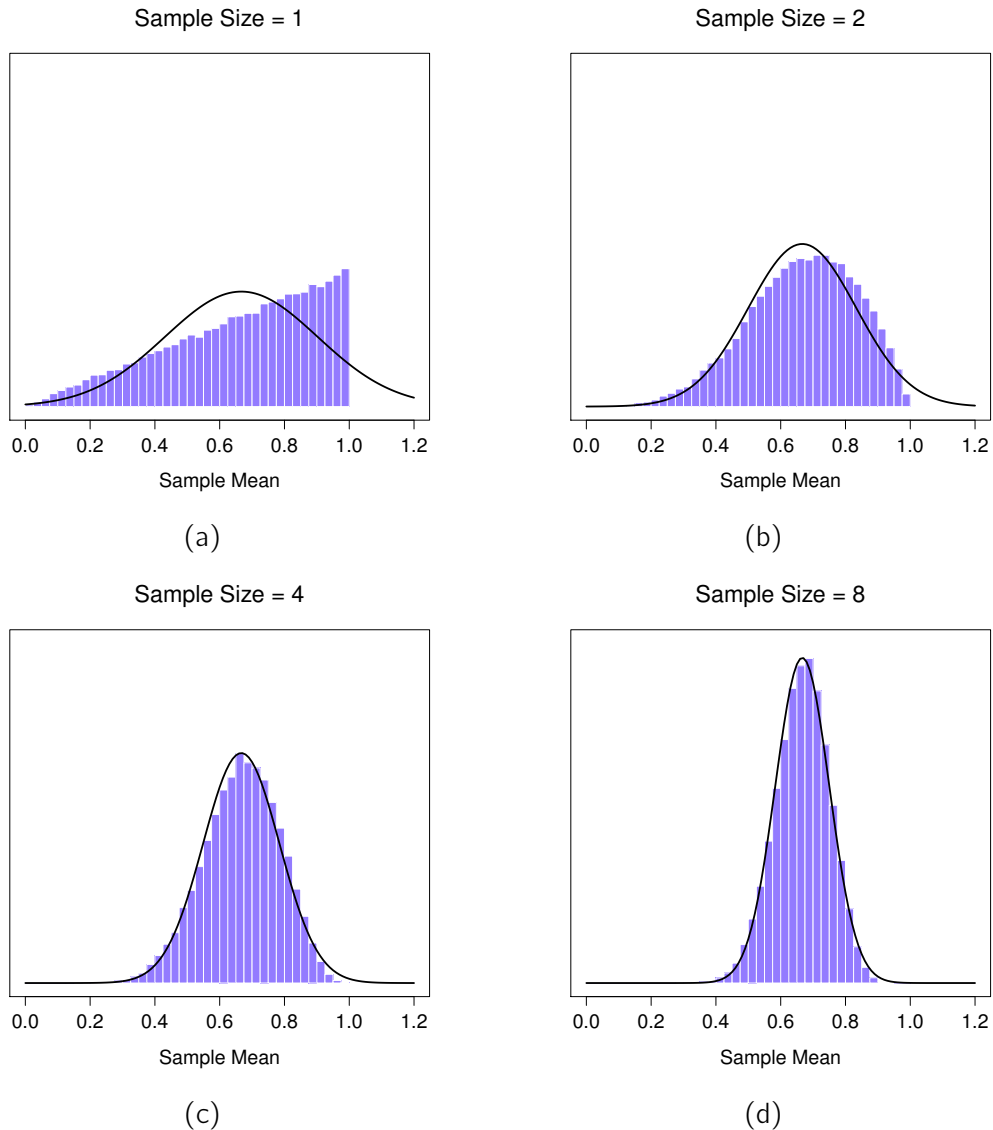


Figure 5.10: A demonstration of the central limit theorem. In panel a, we have a non-normal population distribution, and panels b-d show the sampling distribution of the mean for samples of size 2,4 and 8 for data drawn from the distribution in panel a. As you can see, even though the original population distribution is non-normal the sampling distribution of the mean becomes pretty close to normal by the time you have a sample of even 4 observations.

.....

sampling distribution becomes normal.<sup>4</sup>

This result is useful for all sorts of things. It tells us why large experiments are more reliable than small ones, and because it gives us an explicit formula for the standard error it tells us *how much* more reliable a large experiment is. It tells us why the normal distribution is, well, *normal*. In real experiments, many of the things that we want to measure are actually averages of lots of different quantities (e.g., arguably, “general” intelligence as measured by IQ is an average of a large number of “specific” skills and abilities), and when that happens, the averaged quantity should follow a normal distribution. Because of this mathematical law, the normal distribution pops up over and over again in real data.

## 5.4

---

### Estimating population parameters

In all the IQ examples in the previous sections we actually knew the population parameters ahead of time. As every undergraduate gets taught in their very first lecture on the measurement of intelligence, IQ scores are *defined* to have mean 100 and standard deviation 15. However, this is a bit of a lie. How do we know that IQ scores have a true population mean of 100? Well, we know this because the people who designed the tests have administered them to very large samples, and have then “rigged” the scoring rules so that their sample has mean 100. That’s not a bad thing of course, it’s an important part of designing a psychological measurement. However, it’s important to keep in mind that this theoretical mean of 100 only attaches to the population that the test designers used to design the tests. Good test designers will actually go to some lengths to provide “test norms” that can apply to lots of different populations (e.g., different age groups, nationalities etc).

This is very handy, but of course almost every research project of interest involves looking at a different population of people to those used in the test norms. For instance, you might be interested in the impact of pesticide exposure on cognitive function in Gozo, an island in Malta that is reasonably agricultural. You could decide to compare IQ scores among individuals living in close proximity to farms in Gozo with those from a less agriculturally-intensive area in Malta,

---

<sup>4</sup>As usual, I’m being a bit sloppy here. The central limit theorem is a bit more general than this section implies. Like most introductory stats texts I’ve discussed one situation where the central limit theorem holds: when you’re taking an average across lots of independent events drawn from the same distribution. However, the central limit theorem is much broader than this. There’s a whole class of things called “*U*-statistics” for instance, all of which satisfy the central limit theorem and therefore become normally distributed for large sample sizes. The mean is one such statistic, but it’s not the only one.

like Sliema.<sup>5</sup> Regardless of which town you're thinking about, it doesn't make a lot of sense simply to *assume* that the true population mean IQ is 100. No-one has, to my knowledge, produced sensible norming data that can automatically be applied to different Maltese localities (and nor would it be particularly sensible to do this in reality). We're going to have to **estimate** the population parameters from a sample of data. So how do we do this?

#### 5.4.1 Estimating the population mean

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

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

---

<sup>5</sup>Please note that if you were *actually* interested in this question you would need to be a *lot* more careful than I'm being here. You *can't* just compare IQ scores between Sliema and Gozo and assume any differences are solely due to pesticide exposure. Even if we assume that the only major variable is pesticide levels (which is unlikely), it's important to remember that people generally believe pesticide exposure can influence cognitive abilities. As mentioned in Chapter 2, this could lead to different demand effects between the two sample groups. In other words, you might end up with an illusory group difference in your data, caused by the fact that people *think* that there is a real difference. I find it pretty implausible to think that the locals wouldn't be well aware of what you were trying to do if a bunch of researchers turned up in Gozo with lab coats and IQ tests, and even less plausible to think that a lot of people would be pretty resentful of you for doing it. Those people won't be as co-operative in the tests. Others might be *more* motivated to do well because they don't want their community to look bad. These motivational biases might not be as pronounced in Sliema, where pesticide exposure is not as much of a concern. Conducting rigorous medical research is complex.

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

#### 5.4.2 Estimating the population standard deviation

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

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

20

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

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

Let's extend this example a little. Suppose I now make a second observation. My data set

now has  $N = 2$  observations of the cromulence of blood, and the complete sample now looks like this:

20, 22

This time around, our sample is *just* large enough for us to be able to observe some variability: two observations is the bare minimum number needed for any variability to be observed! For our new data set, the sample mean is  $\bar{X} = 21$ , and the sample standard deviation is  $s = 1$ . What intuitions do we have about the population? Again, as far as the population mean goes, the best guess we can possibly make is the sample mean. If forced to guess we'd probably guess that the population mean cromulence is 21. What about the standard deviation? This is a little more complicated. The sample standard deviation is only based on two observations, and if you're at all like me you probably have the intuition that, with only two observations we haven't given the population "enough of a chance" to reveal its true variability to us. It's not just that we suspect that the estimate is *wrong*, after all with only two observations we expect it to be wrong to some degree. The worry is that the error is *systematic*. Specifically, we suspect that the sample standard deviation is likely to be smaller than the population standard deviation.

This intuition feels right, but it would be nice to demonstrate this somehow. There are in fact mathematical proofs that confirm this intuition, but unless you have the right mathematical background they don't help very much. Instead, what I'll do is simulate the results of some experiments. With that in mind, let's return to our IQ studies. Suppose the true population mean IQ is 100 and the standard deviation is 15. First I'll conduct an experiment in which I measure  $N = 2$  IQ scores and I'll calculate the sample standard deviation. If I do this over and over again, and plot a histogram of these sample standard deviations, what I have is the *sampling distribution of the standard deviation*. I've plotted this distribution in Figure 5.11. Even though the true population standard deviation is 15 the average of the *sample* standard deviations is only 8.5. Notice that this is a very different result to what we found in Figure 5.9b when we plotted the sampling distribution of the mean, where the population mean is 100 and the average of the sample means is also 100.

Now let's extend the simulation. Instead of restricting ourselves to the situation where  $N = 2$ , let's repeat the exercise for sample sizes from 1 to 10. If we plot the average sample mean and average sample standard deviation as a function of sample size, you get the results shown in Figure 5.12. On the left hand side (panel a) I've plotted the average sample mean and on the right hand side (panel b) I've plotted the average standard deviation. The two plots are quite different: *on average*, the average sample mean is equal to the population mean. It is an **unbiased estimator**, which is essentially the reason why your best estimate for the population mean is the sample mean.<sup>6</sup> The plot on the right is quite different: on average, the sample standard deviation  $s$  is *smaller* than the population standard deviation  $\sigma$ . It is a

---

<sup>6</sup>I should note that I'm hiding something here. Unbiasedness is a desirable characteristic for an estimator, but there are other things that matter besides bias. However, it's beyond the scope of this book to discuss this in any detail. I just want to draw your attention to the fact that there's some hidden complexity here.

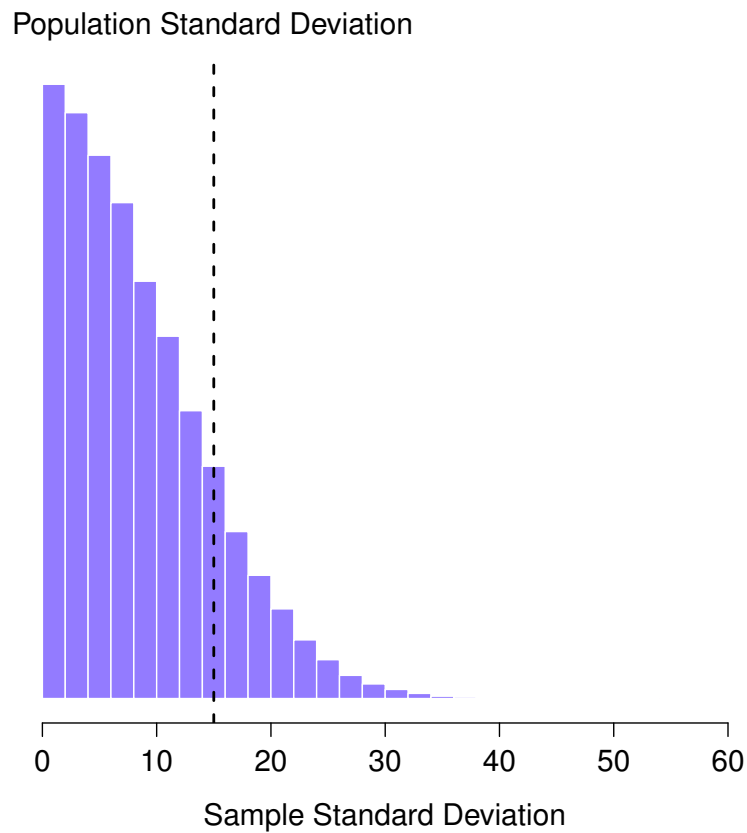


Figure 5.11: The sampling distribution of the sample standard deviation for a “two IQ scores” experiment. The true population standard deviation is 15 (dashed line), but as you can see from the histogram the vast majority of experiments will produce a much smaller sample standard deviation than this. On average, this experiment would produce a sample standard deviation of only 8.5, well below the true value! In other words, the sample standard deviation is a *biased* estimate of the population standard deviation.

.....



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

The fix to this systematic bias turns out to be very simple. Here’s how it works. Before tackling the standard deviation let’s look at the variance. If you recall from Section 4.2, the sample variance is defined to be the average of the squared deviations from the sample mean. That is:

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

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

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

This is an unbiased estimator of the population variance  $\sigma$ . Moreover, this finally answers the question we raised in Section 4.2. Why did JASP give us slightly different answers for variance? It’s because JASP calculates  $\hat{\sigma}^2$  not  $s^2$ , that’s why. A similar story applies for the standard deviation. If we divide by  $N - 1$  rather than  $N$  our estimate of the population standard deviation becomes:

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

and when we use JASP’s built in standard deviation function, what it’s doing is calculating  $\hat{\sigma}$ , not  $s$ .<sup>a</sup>

---

<sup>a</sup>Okay, I’m hiding something else here. In a bizarre and counter-intuitive twist, since  $\hat{\sigma}^2$  is an unbiased estimator of  $\sigma^2$ , you’d assume that taking the square root would be fine and  $\hat{\sigma}$  would be an unbiased estimator of  $\sigma$ . Right? Weirdly, it’s not. There’s actually a subtle, tiny bias in  $\hat{\sigma}$ . This is just bizarre:  $\hat{\sigma}^2$  is an unbiased estimate of the population variance  $\sigma^2$ , but when you take the square root, it turns out that  $\hat{\sigma}$  is a biased estimator of the population standard deviation  $\sigma$ . Weird, weird, weird, right? So, why is  $\hat{\sigma}$  biased? The technical answer is “because non-linear transformations (e.g., the square root) don’t commute with expectation”, but that just sounds like gibberish to everyone who hasn’t taken a course in mathematical statistics. Fortunately, it doesn’t matter for practical purposes. The bias is small, and in real life everyone uses  $\hat{\sigma}$  and it works just fine. Sometimes mathematics is just annoying.

One final point. In practice, a lot of people tend to refer to  $\hat{\sigma}$  (i.e., the formula where we divide by  $N - 1$ ) as the *sample* standard deviation. Technically, this is incorrect. The *sample* standard deviation should be equal to  $s$  (i.e., the formula where we divide by  $N$ ). These aren’t the same thing, either conceptually or numerically. One is a property of the sample, the other

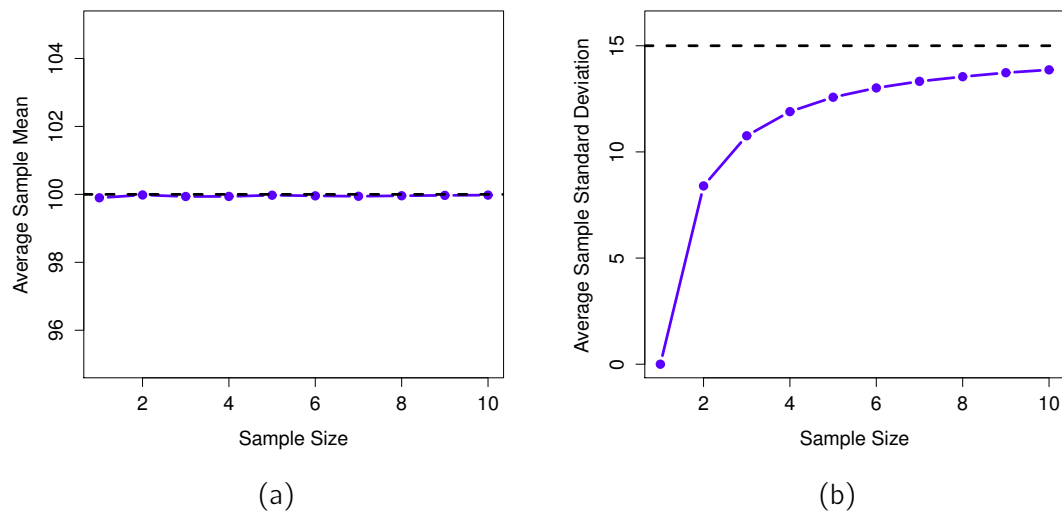


Figure 5.12: An illustration of the fact that the sample mean is an unbiased estimator of the population mean (panel a), but the sample standard deviation is a biased estimator of the population standard deviation (panel b). For the figure I generated 10,000 simulated data sets with 1 observation each, 10,000 more with 2 observations, and so on up to a sample size of 10. Each data set consisted of fake IQ data, that is the data were normally distributed with a true population mean of 100 and standard deviation 15. *On average*, the sample means turn out to be 100, regardless of sample size (panel a). However, the sample standard deviations turn out to be systematically too small (panel b), especially for small sample sizes.

.....

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

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

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

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

## 5.5

### Estimating a confidence interval

*Statistics means never having to say you're certain*

– Unknown origin<sup>7</sup>

Up to this point in this chapter, I've outlined the basics of sampling theory which statisticians rely on to make guesses about population parameters on the basis of a sample of data. As this discussion illustrates, one of the reasons we need all this sampling theory is that every

<sup>7</sup>This quote appears on a great many t-shirts and websites, and even gets a mention in a few academic papers (e.g., <http://www.amstat.org/publications/jse/v10n3/friedman.html>, but I've never found the original source.

data set leaves us with a some of uncertainty, so our estimates are never going to be perfectly accurate. The thing that has been missing from this discussion is an attempt to *quantify* the amount of uncertainty that attaches to our estimate. It's not enough to be able guess that, say, the mean IQ of undergraduate psychology students is 115 (yes, I just made that number up). We also want to be able to say something that expresses the degree of certainty that we have in our guess. For example, it would be nice to be able to say that there is a 95% chance that the true mean lies between 109 and 121. The name for this is a **confidence interval** for the mean.

Armed with an understanding of sampling distributions, constructing a confidence interval for the mean is actually pretty easy. Here's how it works. Suppose the true population mean is  $\mu$  and the standard deviation is  $\sigma$ . I've just finished running my study that has  $N$  participants, and the mean IQ among those participants is  $\bar{X}$ . We know from our discussion of the central limit theorem (Section 5.3.3) that the sampling distribution of the mean is approximately normal. We also know from our discussion of the normal distribution Section ?? that there is a 95% chance that a normally-distributed quantity will fall within about two standard deviations of the true mean.

To be more precise, the more correct answer is that there is a 95% chance that a normally-distributed quantity will fall within 1.96 standard deviations of the true mean. Next, recall that the standard deviation of the sampling distribution is referred to as the standard error, and the standard error of the mean is written as SEM. When we put all these pieces together, we learn that there is a 95% probability that the sample mean  $\bar{X}$  that we have actually observed lies within 1.96 standard errors of the population mean.

Mathematically, we write this as:

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

where the SEM is equal to  $\sigma/\sqrt{N}$  and we can be 95% confident that this is true. However, that's not answering the question that we're actually interested in. The equation above tells us what we should expect about the sample mean given that we know what the population parameters are. What we *want* is to have this work the other way around. We want to know what we should believe about the population parameters, given that we have observed a particular sample. However, it's not too difficult to do this. Using a little high school algebra, a sneaky way to rewrite our equation is like this:

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

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

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

Of course, there's nothing special about the number 1.96. It just happens to be the multiplier you need to use if you want a 95% confidence interval. If I'd wanted a 70% confidence interval, I would have used 1.04 as the magic number rather than 1.96.

#### 5.5.1 A slight mistake in the formula

As usual, I lied. The formula that I've given above for the 95% confidence interval is approximately correct, but I glossed over an important detail in the discussion. Notice my formula requires you to use the standard error of the mean, SEM, which in turn requires you to use the true population standard deviation  $\sigma$ . Yet, in Section 5.4 I stressed the fact that we don't actually *know* the true population parameters. Because we don't know the true value of  $\sigma$  we have to use an estimate of the population standard deviation  $\hat{\sigma}$  instead. This is pretty straightforward to do, but this has the consequence that we need to use the percentiles of the  $t$ -distribution rather than the normal distribution to calculate our magic number, and the answer depends on the sample size. When  $N$  is very large, we get pretty much the same value using the  $t$ -distribution or the normal distribution: 1.96. But when  $N$  is small we get a much bigger number when we use the  $t$  distribution: 2.26.

There's nothing too mysterious about what's happening here. Bigger values mean that the confidence interval is wider, indicating that we're more uncertain about what the true value of  $\mu$  actually is. When we use the  $t$  distribution instead of the normal distribution we get

bigger numbers, indicating that we have more uncertainty. And why do we have that extra uncertainty? Well, because our estimate of the population standard deviation  $\hat{\sigma}$  might be wrong! If it's wrong, it implies that we're a bit less sure about what our sampling distribution of the mean actually looks like, and this uncertainty ends up getting reflected in a wider confidence interval.

### 5.5.2 Interpreting a confidence interval

The hardest thing about confidence intervals is understanding what they *mean*. Whenever people first encounter confidence intervals, the first instinct is almost always to say that “there is a 95% probability that the true mean lies inside the confidence interval”. It's simple and it seems to capture the common sense idea of what it means to say that I am “95% confident”. Unfortunately, it's not quite right. The intuitive definition relies very heavily on your own personal *beliefs* about the value of the population mean. I say that I am 95% confident because those are my beliefs. In everyday life that's perfectly okay, but if you remember back to Section ??, you'll notice that talking about personal belief and confidence is a Bayesian idea. However, confidence intervals are *not* Bayesian tools. Like everything else in this chapter, confidence intervals are *frequentist* tools, and if you are going to use frequentist methods then it's not appropriate to attach a Bayesian interpretation to them. If you use frequentist methods, you must adopt frequentist interpretations!

Okay, so if that's not the right answer, what is? Remember what we said about frequentist probability. The only way we are allowed to make “probability statements” is to talk about a sequence of events, and to count up the frequencies of different kinds of events. From that perspective, the interpretation of a 95% confidence interval must have something to do with replication. Specifically, if we replicated the experiment over and over again and computed a 95% confidence interval for each replication, then 95% of those *intervals* would contain the true mean. More generally, 95% of all confidence intervals constructed using this procedure should contain the true population mean. This idea is illustrated in Figure 5.13, which shows 50 confidence intervals constructed for a “measure 10 IQ scores” experiment (top panel) and another 50 confidence intervals for a “measure 25 IQ scores” experiment (bottom panel). A bit fortuitously, across the 100 replications that I simulated, it turned out that exactly 95 of them contained the true mean.

The critical difference here is that the Bayesian claim makes a probability statement about the population mean (i.e., it refers to our uncertainty about the population mean), which is not allowed under the frequentist interpretation of probability because you can't “replicate” a population! In the frequentist claim, the population mean is fixed and no probabilistic claims can be made about it. Confidence intervals, however, are repeatable so we can replicate experiments. Therefore a frequentist is allowed to talk about the probability that the *confidence interval* (a random variable) contains the true mean, but is not allowed to talk about the probability that the *true population mean* (not a repeatable event) falls within the confidence interval.

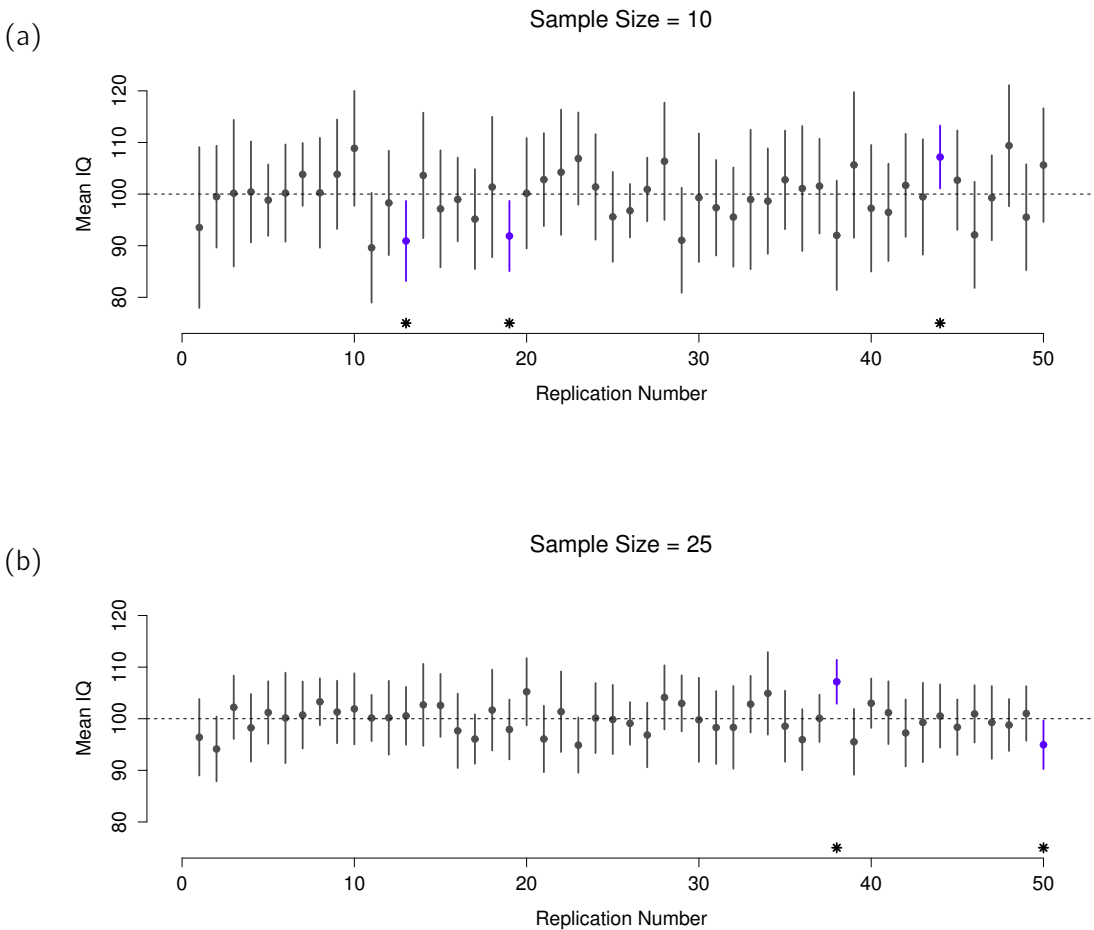


Figure 5.13: 95% confidence intervals. The top (panel a) shows 50 simulated replications of an experiment in which we measure the IQs of 10 people. The dot marks the location of the sample mean and the line shows the 95% confidence interval. In total 47 of the 50 confidence intervals do contain the true mean (i.e., 100), but the three intervals marked with asterisks do not. The lower graph (panel b) shows a similar simulation, but this time we simulate replications of an experiment that measures the IQs of 25 people.

I know that this seems a little pedantic, but it does matter. It matters because the difference in interpretation leads to a difference in the mathematics. There is a Bayesian alternative to confidence intervals, known as *credible intervals*. In most situations credible intervals are quite similar to confidence intervals, but in other cases they are drastically different. As promised, though, I'll talk more about the Bayesian perspective in Chapter ??.

### 5.5.3 Calculating confidence intervals in JASP

As of this edition, JASP does not (yet) include a simple way to calculate confidence intervals for the mean as part of the 'Descriptives' functionality. But the 'Descriptives' do have a check box for the S.E. Mean, so you can use this to calculate the lower 95% confidence interval as:

$\text{Mean} - (1.96 * \text{S.E. Mean})$  , and the upper 95% confidence interval as:

$\text{Mean} + (1.96 * \text{S.E. Mean})$

95% confidence intervals are the de facto standard in psychology. So, for example, if I load the `IQsim.jasp` file, check mean and S.E mean under 'Descriptives', I can work out the confidence interval associated with the simulated mean IQ:

Lower 95% CI =  $100.107 - (1.96 * 0.150) = 99.813$

Upper 95% CI =  $100.107 + (1.96 * 0.150) = 100.401$

So, in our simulated large sample data with  $N=10,000$ , the mean IQ score is 100.107 with a 95% CI from 99.813 to 100.401. Hopefully that's clear and fairly easy to interpret. So, although there currently is not a straightforward way to get JASP to calculate the confidence interval as part of the variable 'Descriptives' options, if we wanted to we could pretty easily work it out by hand.

Similarly, when it comes to plotting confidence intervals in JASP, this is also not (yet) available as part of the 'Descriptives' options. However, when we get onto learning about specific statistical tests, for example in Chapter ??, we will see that we can plot confidence intervals as part of the data analysis. That's pretty cool, so we'll show you how to do that later on.

## 5.6

---

### Summary

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

- Basic ideas about samples, sampling and populations (Section 5.1)
- Statistical theory of sampling: the law of large numbers (Section 5.2), sampling distri-



butions and the central limit theorem (Section 5.3).

- Estimating means and standard deviations (Section 5.4)
- Estimating a confidence interval (Section 5.5)

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

## 6. Hypothesis testing

---

*The process of induction is the process of assuming the simplest law that can be made to harmonize with our experience. This process, however, has no logical foundation but only a psychological one. It is clear that there are no grounds for believing that the simplest course of events will really happen. It is an hypothesis that the sun will rise tomorrow: and this means that we do not know whether it will rise.*

– Ludwig Wittgenstein<sup>1</sup>

In the last chapter I discussed the ideas behind estimation, which is one of the two “big ideas” in inferential statistics. It’s now time to turn our attention to the other big idea, which is *hypothesis testing*. In its most abstract form, hypothesis testing is really a very simple idea. The researcher has some theory about the world and wants to determine whether or not the data actually support that theory. However, the details are messy and most people find the theory of hypothesis testing to be the most frustrating part of statistics. The structure of the chapter is as follows. First, I’ll describe how hypothesis testing works in a fair amount of detail, using a simple running example to show you how a hypothesis test is “built”. I’ll try to avoid being too dogmatic while doing so, and focus instead on the underlying logic of the testing procedure.<sup>2</sup> Afterwards, I’ll spend a bit of time talking about the various dogmas, rules and heresies that surround the theory of hypothesis testing.

---

<sup>1</sup>The quote comes from Wittgenstein’s (1922) text, *Tractatus Logico-Philosophicus*.

<sup>2</sup>A technical note. The description below differs subtly from the standard description given in a lot of introductory texts. The orthodox theory of null hypothesis testing emerged from the work of Sir Ronald Fisher and Jerzy Neyman in the early 20th century; but Fisher and Neyman actually had very different views about how it should work. The standard treatment of hypothesis testing that most texts use is a hybrid of the two approaches. The treatment here is a little more Neyman-style than the orthodox view, especially as regards the meaning of the  $p$  value.

## A menagerie of hypotheses

Eventually we all succumb to madness. For me, that day will arrive once I'm finally promoted to full professor. Safely ensconced in my ivory tower, happily protected by tenure, I will finally be able to take leave of my senses (so to speak) and indulge in that most thoroughly unproductive line of medical research, evaluating the veracity of local folk medical tales. For me, the intrigue lies in the quaint Maltese folk tale that claims the shape of a pregnant woman's belly can predict the biological sex of her baby. Round bellies are said to foretell the arrival of a girl, while less round bellies are said to herald a boy. Safely ensconced in my academic nook, I find myself drawn to test this fascinating piece of cultural lore.<sup>3</sup>

Let's suppose that this glorious day has come. My first study on this subject aims to scientifically evaluate the accuracy of this belly shape method in predicting a baby's biological sex. Participants in the study are pregnant women in their third trimester, with the shape of their bellies categorized as either Round or Not Round. Each participant's baby's biological sex is then confirmed through ultrasound or after birth.

The setup is straightforward. A single, blind assessment is made for each participant, determining the shape of their belly and noting down the actual biological sex of the baby. In my dataset, I have information from  $N$  participants, with a certain number  $X$  having their baby's biological sex correctly predicted by the belly shape. For example, let's assume  $N = 100$  and  $X = 62$ .

This brings us to a pivotal question: Does the shape of a pregnant woman's belly provide any reliable indication of her baby's biological sex, or is this method no better than a random guess? This is where the power of hypothesis testing comes into play. But before we venture into the *testing* of these hypotheses, we need to clarify what we mean by hypotheses.

### 6.1.1 Research hypotheses versus statistical hypotheses

The first distinction that you need to keep clear in your mind is between research hypotheses and statistical hypotheses. In my homeopathy study, my overarching scientific objective is to demonstrate that homeopathic remedies can effectively lower blood pressure. In this situation, my research hypothesis is explicit: I aim to find evidence supporting the efficacy of homeopathy in blood pressure reduction. In other situations I might actually be a lot more neutral than that, so I might say that my research goal is to determine whether or not homeopathy has an impact on blood pressure levels. Regardless of how I want to portray myself, the basic point that I'm trying to convey here is that a research hypothesis involves making a substantive, testable scientific claim. Any of the following would count as **research hypotheses**:

---

<sup>3</sup>My apologies to anyone who actually believes in this stuff

- *Propranolol reduces systolic blood pressure.* This hypothesis posits a causal relationship between two medically relevant factors (homeopathic treatment and systolic blood pressure), making it a valid research hypothesis.
- *Blood pressure levels are related to adherence to medication.* Unlike the previous example, this is a weaker claim, suggesting a correlational relationship between two medical variables (blood pressure levels and medication adherence).
- *Diabetes is a disorder of insulin resistance.* This hypothesis is fundamentally different in nature from the others. It's not a relational claim, but an ontological one that seeks to define the intrinsic character of diabetes. Expanding on this is worthwhile. Generally, it's more straightforward to design experiments that test research hypotheses framed as "does X affect Y?" rather than addressing deeper questions like "what is X?" In practice, what usually happens is you create experiments to test relational hypotheses that are rooted in these more foundational ontological claims. For instance, if I believe that diabetes is essentially a disorder of insulin resistance, my experiments will likely involve looking for relationships between markers of diabetes and measures of insulin sensitivity. As a result, the bulk of day-to-day medical research questions are often relational in nature, but they are almost always underpinned by deeper ontological questions concerning medical conditions.

Notice that in practice, my research hypotheses could overlap a lot. My ultimate goal in the baby experiment might be to test an ontological claim like "belly shape determines biological sex", but I might operationally restrict myself to a narrower hypothesis like "Some people can predict a baby's biological sex after looking at the shape of a pregnant belly. That said, there are some things that really don't count as proper research hypotheses in any meaningful sense:

- *Health is wealth.* This is too vague to be testable. Whilst it's okay for a research hypothesis to have a degree of vagueness to it, it has to be possible to operationalise your theoretical ideas. Maybe I'm just not creative enough to see it, but I can't see how this can be converted into any concrete research design. If that's true then this isn't a scientific research hypothesis, it's a pop song. That doesn't mean it's not interesting. A lot of deep questions that humans have fall into this category. Maybe one day science will be able to construct testable theories of love, or to test to see if God exists, and so on. But right now we can't, and I wouldn't bet on ever seeing a satisfying scientific approach to either.
- *The first rule of tautology club is the first rule of tautology club.* This is not a substantive claim of any kind. It's true by definition. No conceivable state of nature could possibly be inconsistent with this claim. We say that this is an unfalsifiable hypothesis, and as such it is outside the domain of science. Whatever else you do in science your claims must have the possibility of being wrong.
- *More patients in my study will report feeling better" than worse".* This one fails as a research hypothesis because it's a claim about the data set, not about the anything

meaningful you might want to research (unless, of course, the actual research question investigates whether patients have a reporting bias toward feeling "better"). Actually, this hypothesis is starting to sound more like a statistical hypothesis than a research hypothesis.

As you can see, research hypotheses can be somewhat messy at times and ultimately they are *scientific* claims. **Statistical hypotheses** are neither of these two things. Statistical hypotheses must be mathematically precise and they must correspond to specific claims about the characteristics of the data generating mechanism (i.e., the "population"). Even so, the intent is that statistical hypotheses bear a clear relationship to the substantive research hypotheses that you care about! For instance, in my "belly" study my research hypothesis is that some people are able to predict a baby's biological sex or whatever. What I want to do is to "map" this onto a statement about how the data were generated. So let's think about what that statement would be. The quantity that I'm interested in within the experiment is  $P(\text{"correct"})$ , the true-but-unknown probability with which the participants in my experiment answer the question correctly. Let's use the Greek letter  $\theta$  (theta) to refer to this probability. Here are four different statistical hypotheses:

- If people cannot predict the baby's biological sex and if my experiment is well designed then my participants are just guessing. So I should expect them to get it right half of the time and so my statistical hypothesis is that the true probability of choosing correctly is  $\theta = 0.5$ .
- Alternatively, suppose people can predict the child's biological sex. If that's true people will perform better than chance and the statistical hypothesis is that  $\theta > 0.5$ .
- A third possibility is that people can predict the child's biological sex, but the guesses are all reversed and people don't realise it (okay, that's wacky, but you never know). If that's how it works then you'd expect people's performance to be *below* chance. This would correspond to a statistical hypothesis that  $\theta < 0.5$ .
- Finally, suppose people can predict the child's biological sex but I have no idea whether people are seeing the right sex or the wrong one. In that case the only claim I could make about the data would be that the probability of making the correct answer is *not* equal to 0.5. This corresponds to the statistical hypothesis that  $\theta \neq 0.5$ .

All of these are legitimate examples of a statistical hypothesis because they are statements about a population parameter and are meaningfully related to my experiment.

What this discussion makes clear, I hope, is that when attempting to construct a statistical hypothesis test the researcher actually has two quite distinct hypotheses to consider. First, he or she has a research hypothesis, and this then corresponds to a statistical hypothesis (a claim about the data generating population). In my example these might be:

Dani's **research** hypothesis: "belly shape determines biological sex"  
Dani's **statistical** hypothesis:  $\theta \neq 0.5$

And a key thing to recognise is this. *A statistical hypothesis test is a test of the statistical hypothesis, not the research hypothesis.* If your study is badly designed then the link between your research hypothesis and your statistical hypothesis is broken. To give a silly example, suppose that my belly study was conducted in a situation where the participant has an ultrasound available. If that happens I would be able to find very strong evidence that  $\theta \neq 0.5$ , but this would tell us nothing about whether “belly shape determines biological sex”.

### 6.1.2 Null hypotheses and alternative hypotheses

So far, so good. I have a research hypothesis that corresponds to what I want to believe about the world, and I can map it onto a statistical hypothesis that corresponds to what I want to believe about how the data were generated. It's at this point that things get somewhat counter-intuitive for a lot of people. Because what I'm about to do is invent a new statistical hypothesis (the “null” hypothesis,  $H_0$ ) that corresponds to the exact opposite of what I want to believe, and then focus exclusively on that almost to the neglect of the thing I'm actually interested in (which is now called the “alternative” hypothesis,  $H_1$ ). In our example, the null hypothesis is that  $\theta = 0.5$ , since that's what we'd expect if folk prediction *was not* true. My hope, of course, is that this age old tradition is totally real and so the *alternative* to this null hypothesis is  $\theta \neq 0.5$ . In essence, what we're doing here is dividing up the possible values of  $\theta$  into two groups: those values that I really hope aren't true (the null), and those values that I'd be happy with if they turn out to be right (the alternative). Having done so, the important thing to recognise is that the goal of a hypothesis test is *not* to show that the alternative hypothesis is (probably) true. The goal is to show that the null hypothesis is (probably) false. Most people find this pretty weird.

The best way to think about it, in my experience, is to imagine that a hypothesis test is a criminal trial<sup>4</sup>, *the trial of the null hypothesis*. The null hypothesis is the defendant, the researcher is the prosecutor, and the statistical test itself is the judge. Just like a criminal trial, there is a presumption of innocence. The null hypothesis is *deemed* to be true unless you, the researcher, can prove beyond a reasonable doubt that it is false. You are free to design your experiment however you like (within reason, obviously!) and your goal when doing so is to maximise the chance that the data will yield a conviction for the crime of being false. The catch is that the statistical test sets the rules of the trial and those rules are designed to protect the null hypothesis, specifically to ensure that if the null hypothesis is actually true the chances of a false conviction are guaranteed to be low. This is pretty important. After all, the null hypothesis doesn't get a lawyer, and given that the researcher is trying desperately to prove it to be false *someone* has to protect it.

---

<sup>4</sup>This analogy only works if you're from an adversarial legal system like UK/US/Australia. As I understand these things, the French inquisitorial system is quite different.

## Two types of errors

Before going into details about how a statistical test is constructed it's useful to understand the philosophy behind it. I hinted at it when pointing out the similarity between a null hypothesis test and a criminal trial, but I should now be explicit. Ideally, we would like to construct our test so that we never make any errors. Unfortunately, since the world is messy, this is never possible. Sometimes you're just really unlucky. For instance, suppose you flip a coin 10 times in a row and it comes up heads all 10 times. That feels like very strong evidence for a conclusion that the coin is biased, but of course there's a 1 in 1024 chance that this would happen even if the coin was totally fair. In other words, in real life we *always* have to accept that there's a chance that we made a mistake. As a consequence the goal behind statistical hypothesis testing is not to *eliminate* errors, but to *minimise* them.

At this point, we need to be a bit more precise about what we mean by "errors". First, let's state the obvious. It is either the case that the null hypothesis is true or that it is false, and our test will either retain the null hypothesis or reject it.<sup>5</sup> So, as the table below illustrates, after we run the test and make our choice one of four things might have happened:

	retain $H_0$	reject $H_0$
$H_0$ is true	correct decision	error (type I)
$H_0$ is false	error (type II)	correct decision

As a consequence there are actually *two* different types of error here. If we reject a null hypothesis that is actually true then we have made a **type I error**. On the other hand, if we retain the null hypothesis when it is in fact false then we have made a **type II error**.

Remember how I said that statistical testing was kind of like a criminal trial? Well, I meant it. A criminal trial requires that you establish "beyond a reasonable doubt" that the defendant did it. All of the evidential rules are (in theory, at least) designed to ensure that there's (almost) no chance of wrongfully convicting an innocent defendant. The trial is designed to

<sup>5</sup>An aside regarding the language you use to talk about hypothesis testing. First, one thing you really want to avoid is the word "prove". A statistical test really doesn't *prove* that a hypothesis is true or false. Proof implies certainty and, as the saying goes, statistics means never having to say you're certain. On that point almost everyone would agree. However, beyond that there's a fair amount of confusion. Some people argue that you're only allowed to make statements like "rejected the null", "failed to reject the null", or possibly "retained the null". According to this line of thinking you can't say things like "accept the alternative" or "accept the null". Personally I think this is too strong. In my opinion, this conflates null hypothesis testing with Karl Popper's falsificationist view of the scientific process. Whilst there are similarities between falsificationism and null hypothesis testing, they aren't equivalent. However, whilst I personally think it's fine to talk about accepting a hypothesis (on the proviso that "acceptance" doesn't actually mean that it's necessarily true, especially in the case of the null hypothesis), many people will disagree. And more to the point, you should be aware that this particular weirdness exists so that you're not caught unawares by it when writing up your own results.

protect the rights of a defendant, as the English jurist William Blackstone famously said, it is “better that ten guilty persons escape than that one innocent suffer.” In other words, a criminal trial doesn’t treat the two types of error in the same way. Punishing the innocent is deemed to be much worse than letting the guilty go free. A statistical test is pretty much the same. The single most important design principle of the test is to *control* the probability of a type I error, to keep it below some fixed probability. This probability, which is denoted  $\alpha$ , is called the **significance level** of the test. And I’ll say it again, because it is so central to the whole set-up, a hypothesis test is said to have significance level  $\alpha$  if the type I error rate is no larger than  $\alpha$ .

So, what about the type II error rate? Well, we’d also like to keep those under control too, and we denote this probability by  $\beta$ . However, it’s much more common to refer to the **power** of the test, that is the probability with which we reject a null hypothesis when it really is false, which is  $1 - \beta$ . To help keep this straight, here’s the same table again but with the relevant numbers added:

	retain $H_0$	reject $H_0$
$H_0$ is true	$1 - \alpha$ (probability of correct retention)	$\alpha$ (type I error rate)
$H_0$ is false	$\beta$ (type II error rate)	$1 - \beta$ (power of the test)

A “powerful” hypothesis test is one that has a small value of  $\beta$ , while still keeping  $\alpha$  fixed at some (small) desired level. By convention, scientists make use of three different  $\alpha$  levels: .05, .01 and .001. Notice the asymmetry here; the tests are designed to *ensure* that the  $\alpha$  level is kept small but there’s no corresponding guarantee regarding  $\beta$ . We’d certainly *like* the type II error rate to be small and we try to design tests that keep it small, but this is typically secondary to the overwhelming need to control the type I error rate. As Blackstone might have said if he were a statistician, it is “better to retain 10 false null hypotheses than to reject a single true one”. To be honest, I don’t know that I agree with this philosophy. There are situations where I think it makes sense, and situations where I think it doesn’t, but that’s neither here nor there. It’s how the tests are built.

## 6.3

### Test statistics and sampling distributions

At this point we need to start talking specifics about how a hypothesis test is constructed. To that end, let’s return to the pregnancy example. Let’s ignore the actual data that we obtained, for the moment, and think about the structure of the experiment. Regardless of what the actual numbers are, the *form* of the data is that  $X$  out of  $N$  people correctly identified the colour of the hidden card. Moreover, let’s suppose for the moment that the null hypothesis really is true, that the folklore is not correct and the true probability that anyone picks the correct colour is exactly  $\theta = 0.5$ . What would we *expect* the data to look like? Well, obviously



we'd expect the proportion of people who make the correct response to be pretty close to 50%. Or, to phrase this in more mathematical terms, we'd say that  $X/N$  is approximately 0.5. Of course, we wouldn't expect this fraction to be *exactly* 0.5. If, for example, we tested  $N = 100$  people and  $X = 53$  of them got the answer right, we'd probably be forced to concede that the data are quite consistent with the null hypothesis. On the other hand, if  $X = 99$  of our participants got the question right then we'd feel pretty confident that the null hypothesis is wrong. Similarly, if only  $X = 3$  people got the answer right we'd be similarly confident that the null was wrong. Let's be a little more technical about this. We have a quantity  $X$  that we can calculate by looking at our data. After looking at the value of  $X$  we make a decision about whether to believe that the null hypothesis is correct, or to reject the null hypothesis in favour of the alternative. The name for this thing that we calculate to guide our choices is a **test statistic**.

Having chosen a test statistic, the next step is to state precisely which values of the test statistic would cause us to reject the null hypothesis, and which values would cause us to keep it. In order to do so we need to determine what the **sampling distribution of the test statistic** would be if the null hypothesis were actually true (we talked about sampling distributions earlier in Section 5.3.1). Why do we need this? Because this distribution tells us exactly what values of  $X$  our null hypothesis would lead us to expect. And, therefore, we can use this distribution as a tool for assessing how closely the null hypothesis agrees with our data.

How do we actually determine the sampling distribution of the test statistic? For a lot of hypothesis tests this step is actually quite complicated, and later on in the book you'll see me being slightly evasive about it for some of the tests (some of them I don't even understand myself). However, sometimes it's very easy. And, fortunately for us, our belly shape example provides us with one of the easiest cases. Our population parameter  $\theta$  is just the overall probability that people respond correctly when asked the question, and our test statistic  $X$  is the *count* of the number of people who did so out of a sample size of  $N$ . We've seen a distribution like this before, in Section ??, and that's exactly what the binomial distribution describes! So, to use the notation and terminology that I introduced in that section, we would say that the null hypothesis predicts that  $X$  is binomially distributed, which is written

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

Since the null hypothesis states that  $\theta = 0.5$  and our experiment has  $N = 100$  people, we have the sampling distribution we need. This sampling distribution is plotted in Figure 6.1. No surprises really, the null hypothesis says that  $X = 50$  is the most likely outcome, and it says that we're almost certain to see somewhere between 40 and 60 correct responses.

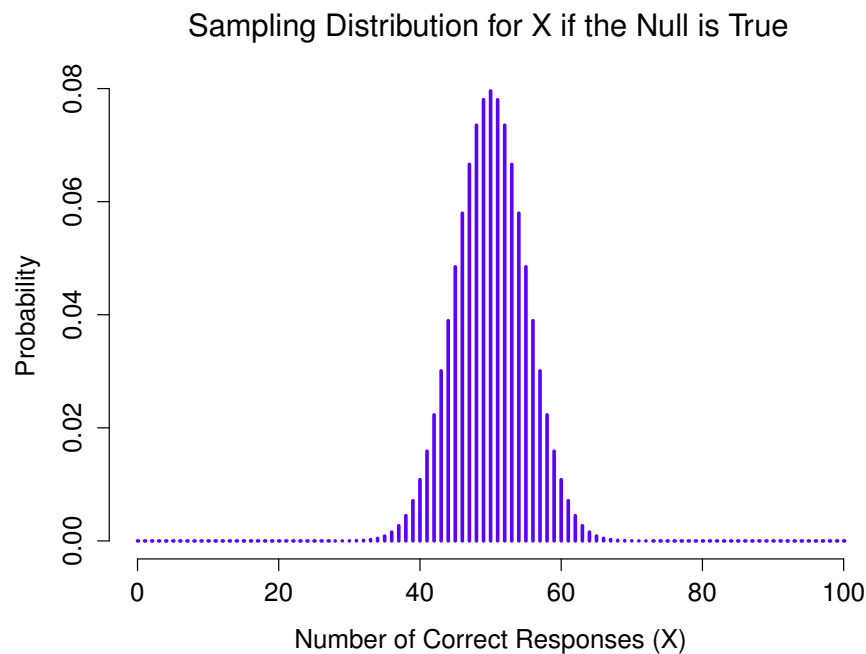


Figure 6.1: The sampling distribution for our test statistic  $X$  when the null hypothesis is true. For our scenario this is a binomial distribution. Not surprisingly, since the null hypothesis says that the probability of a correct response is  $\theta = .5$ , the sampling distribution says that the most likely value is 50 (out of 100) correct responses. Most of the probability mass lies between 40 and 60.

.....

## 6.4

### Making decisions

Okay, we're very close to being finished. We've constructed a test statistic ( $X$ ) and we chose this test statistic in such a way that we're pretty confident that if  $X$  is close to  $N/2$  then we should retain the null, and if not we should reject it. The question that remains is this. Exactly which values of the test statistic should we associate with the null hypothesis, and exactly which values go with the alternative hypothesis? In my pregnancy study, for example, I've observed a value of  $X = 62$ . What decision should I make? Should I choose to believe the null hypothesis or the alternative hypothesis?

#### 6.4.1 Critical regions and critical values

To answer this question we need to introduce the concept of a **critical region** for the test statistic  $X$ . The critical region of the test corresponds to those values of  $X$  that would lead us to reject null hypothesis (which is why the critical region is also sometimes called the rejection region). How do we find this critical region? Well, let's consider what we know:

- $X$  should be very big or very small in order to reject the null hypothesis.
- If the null hypothesis is true, the sampling distribution of  $X$  is Binomial(0.5,  $N$ ).
- If  $\alpha = .05$ , the critical region must cover 5% of this sampling distribution.

It's important to make sure you understand this last point. The critical region corresponds to those values of  $X$  for which we would reject the null hypothesis, and the sampling distribution in question describes the probability that we would obtain a particular value of  $X$  if the null hypothesis were actually true. Now, let's suppose that we chose a critical region that covers 20% of the sampling distribution, and suppose that the null hypothesis is actually true. What would be the probability of incorrectly rejecting the null? The answer is of course 20%. And, therefore, we would have built a test that had an  $\alpha$  level of 0.2. If we want  $\alpha = .05$ , the critical region is only *allowed* to cover 5% of the sampling distribution of our test statistic.

As it turns out those three things uniquely solve the problem. Our critical region consists of the most *extreme values*, known as the **tails** of the distribution. This is illustrated in Figure 6.2. If we want  $\alpha = .05$  then our critical regions correspond to  $X \leq 40$  and  $X \geq 60$ .<sup>6</sup> That is, if the number of people saying "true" is between 41 and 59, then we should retain the null hypothesis. If the number is between 0 to 40, or between 60 to 100, then we should reject the null hypothesis. The numbers 40 and 60 are often referred to as the **critical values** since they define the edges of the critical region.

---

<sup>6</sup>Strictly speaking, the test I just constructed has  $\alpha = .057$ , which is a bit too generous. However, if I'd chosen 39 and 61 to be the boundaries for the critical region then the critical region only covers 3.5% of the distribution. I figured that it makes more sense to use 40 and 60 as my critical values, and be willing to tolerate a 5.7% type I error rate, since that's as close as I can get to a value of  $\alpha = .05$ .

### Critical Regions for a Two-Sided Test

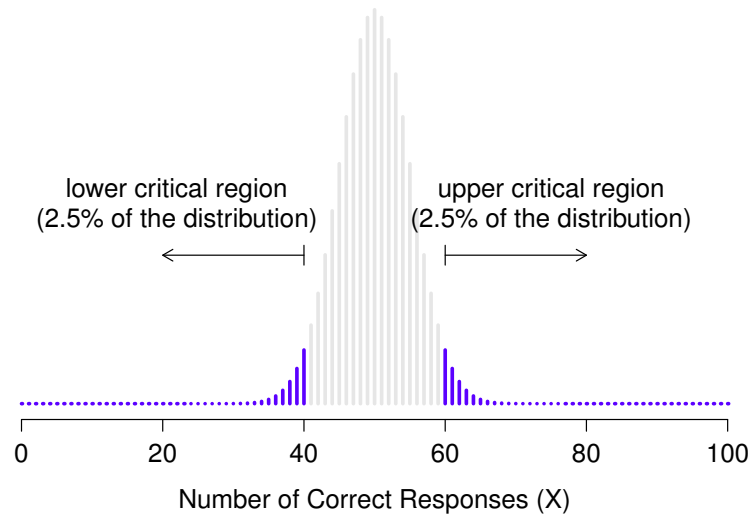


Figure 6.2: The critical region associated with the hypothesis test for the study, for a hypothesis test with a significance level of  $\alpha = .05$ . The plot shows the sampling distribution of  $X$  under the null hypothesis (i.e., same as Figure 6.1). The grey bars correspond to those values of  $X$  for which we would retain the null hypothesis. The blue (darker shaded) bars show the critical region, those values of  $X$  for which we would reject the null. Because the alternative hypothesis is two sided (i.e., allows both  $\theta < .5$  and  $\theta > .5$ ), the critical region covers both tails of the distribution. To ensure an  $\alpha$  level of .05, we need to ensure that each of the two regions encompasses 2.5% of the sampling distribution.

At this point, our hypothesis test is essentially complete:

1. (1) we choose an  $\alpha$  level (e.g.,  $\alpha = .05$ ;
2. (2) come up with some test statistic (e.g.,  $X$ ) that does a good job (in some meaningful sense) of comparing  $H_0$  to  $H_1$ ;
3. (3) figure out the sampling distribution of the test statistic on the assumption that the null hypothesis is true (in this case, binomial); and then
4. (4) calculate the critical region that produces an appropriate  $\alpha$  level (0-40 and 60-100).

All that we have to do now is calculate the value of the test statistic for the real data (e.g.,  $X = 62$ ) and then compare it to the critical values to make our decision. Since 62 is greater than the critical value of 60 we would reject the null hypothesis. Or, to phrase it slightly

differently, we say that the test has produced a statistically significant result.

#### 6.4.2 A note on statistical “significance”

*Like other occult techniques of divination, the statistical method has a private jargon deliberately contrived to obscure its methods from non-practitioners.*

– Attributed to G. O. Ashley<sup>7</sup>

A very brief digression is in order at this point, regarding the word “significant”. The concept of statistical significance is actually a very simple one, but has a very unfortunate name. If the data allow us to reject the null hypothesis, we say that “the result is *statistically significant*”, which is often shortened to “the result is significant”. This terminology is rather old and dates back to a time when “significant” just meant something like “indicated”, rather than its modern meaning which is much closer to “important”. As a result, a lot of modern readers get very confused when they start learning statistics because they think that a “significant result” must be an important one. It doesn’t mean that at all. All that “statistically significant” means is that the data allowed us to reject a null hypothesis. Whether or not the result is actually important in the real world is a very different question, and depends on all sorts of other things.

#### 6.4.3 The difference between one sided and two sided tests

There’s one more thing I want to point out about the hypothesis test that I’ve just constructed. If we take a moment to think about the statistical hypotheses I’ve been using,

$$H_0 : \theta = .5$$

$$H_1 : \theta \neq .5$$

we notice that the alternative hypothesis covers *both* the possibility that  $\theta < .5$  and the possibility that  $\theta > .5$ . This makes sense if I really think that the folk prediction could produce either better-than-chance performance *or* worse-than-chance performance (and there are some people who think that). In statistical language this is an example of a **two-sided test**. It’s called this because the alternative hypothesis covers the area on both “sides” of the null hypothesis, and as a consequence the critical region of the test covers both tails of the sampling distribution (2.5% on either side if  $\alpha = .05$ ), as illustrated earlier in Figure 6.2.

However, that’s not the only possibility. I might only be willing to believe in our little folkloristic ability if it produces better than chance performance. If so, then my alternative hypothesis would only covers the possibility that  $\theta > .5$ , and as a consequence the null hypothesis

---

<sup>7</sup>The internet seems fairly convinced that Ashley said this, though I can’t for the life of me find anyone willing to give a source for the claim.

## Critical Region for a One-Sided Test

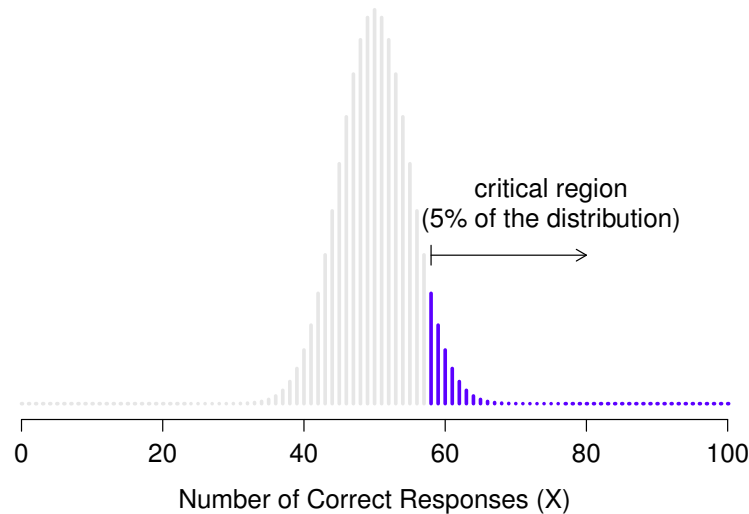


Figure 6.3: The critical region for a one sided test. In this case, the alternative hypothesis is that  $\theta > .5$  so we would only reject the null hypothesis for large values of  $X$ . As a consequence, the critical region only covers the upper tail of the sampling distribution, specifically the upper 5% of the distribution. Contrast this to the two-sided version in Figure 6.2.

.....

now becomes  $\theta \leq .5$

$$H_0 : \theta \leq .5$$

$$H_1 : \theta > .5$$

When this happens, we have what's called a **one-sided test** and the critical region only covers one tail of the sampling distribution. This is illustrated in Figure 6.3.

## 6.5

### The $p$ value of a test

In one sense, our hypothesis test is complete. We've constructed a test statistic, figured out its sampling distribution if the null hypothesis is true, and then constructed the critical region for the test. Nevertheless, I've actually omitted the most important number of all, **the  $p$  value**. It is to this topic that we now turn. There are two somewhat different ways of interpreting a  $p$  value, one proposed by Sir Ronald Fisher and the other by Jerzy Neyman. Both versions are

legitimate, though they reflect very different ways of thinking about hypothesis tests. Most introductory textbooks tend to give Fisher's version only, but I think that's a bit of a shame. To my mind, Neyman's version is cleaner and actually better reflects the logic of the null hypothesis test. You might disagree though, so I've included both. I'll start with Neyman's version.

#### 6.5.1 A softer view of decision making

One problem with the hypothesis testing procedure that I've described is that it makes no distinction at all between a result that is "barely significant" and those that are "highly significant". For instance, in my study the data I obtained only just fell inside the critical region, so I did get a significant effect but it was a pretty near thing. In contrast, suppose that I'd run a study in which  $X = 97$  out of my  $N = 100$  participants got the answer right. This would obviously be significant too but by a much larger margin, such that there's really no ambiguity about this at all. The procedure that I have already described makes no distinction between the two. If I adopt the standard convention of allowing  $\alpha = .05$  as my acceptable Type I error rate, then both of these are significant results.

This is where the  $p$  value comes in handy. To understand how it works, let's suppose that we ran lots of hypothesis tests on the same data set, but with a different value of  $\alpha$  in each case. When we do that for my original biological sex data what we'd get is something like this

Value of $\alpha$	0.05	0.04	0.03	0.02	0.01
Reject the null?	Yes	Yes	Yes	No	No

When we test the data ( $X = 62$  successes out of  $N = 100$  observations), using  $\alpha$  levels of .03 and above, we'd always find ourselves rejecting the null hypothesis. For  $\alpha$  levels of .02 and below we always end up retaining the null hypothesis. Therefore, somewhere between .02 and .03 there must be a smallest value of  $\alpha$  that would allow us to reject the null hypothesis for this data. This is the  $p$  value. As it turns out the data has  $p = .021$ . In short,

$p$  is defined to be the smallest Type I error rate ( $\alpha$ ) that you have to be willing to tolerate if you want to reject the null hypothesis.

If it turns out that  $p$  describes an error rate that you find intolerable, then you must retain the null. If you're comfortable with an error rate equal to  $p$ , then it's okay to reject the null hypothesis in favour of your preferred alternative.

In effect,  $p$  is a summary of all the possible hypothesis tests that you could have run, taken across all possible  $\alpha$  values. And as a consequence it has the effect of "softening" our decision process. For those tests in which  $p \leq \alpha$  you would have rejected the null hypothesis, whereas for those tests in which  $p > \alpha$  you would have retained the null. In my study I obtained  $X = 62$  and as a consequence I've ended up with  $p = .021$ . So the error rate I have to tolerate is

2.1%. In contrast, suppose my experiment had yielded  $X = 97$ . What happens to my  $p$  value now? This time it's shrunk to  $p = 1.36 \times 10^{-25}$ , which is a tiny, tiny<sup>8</sup> Type I error rate. For this second case I would be able to reject the null hypothesis with a lot more confidence, because I only have to be "willing" to tolerate a type I error rate of about 1 in 10 trillion trillion in order to justify my decision to reject.

### 6.5.2 The probability of extreme data

The second definition of the  $p$ -value comes from Sir Ronald Fisher, and it's actually this one that you tend to see in most introductory statistics textbooks. Notice how, when I constructed the critical region, it corresponded to the *tails* (i.e., extreme values) of the sampling distribution? That's not a coincidence, almost all "good" tests have this characteristic (good in the sense of minimising our type II error rate,  $\beta$ ). The reason for that is that a good critical region almost always corresponds to those values of the test statistic that are least likely to be observed if the null hypothesis is true. If this rule is true, then we can define the  $p$ -value as the probability that we would have observed a test statistic that is at least as extreme as the one we actually did get. In other words, if the data are extremely implausible according to the null hypothesis, then the null hypothesis is probably wrong.

### 6.5.3 A common mistake

Okay, so you can see that there are two rather different but legitimate ways to interpret the  $p$  value, one based on Neyman's approach to hypothesis testing and the other based on Fisher's. Unfortunately, there is a third explanation that people sometimes give, especially when they're first learning statistics, and it is *absolutely and completely wrong*. This mistaken approach is to refer to the  $p$  value as "the probability that the null hypothesis is true". It's an intuitively appealing way to think, but it's wrong in two key respects. First, null hypothesis testing is a frequentist tool and the frequentist approach to probability does *not* allow you to assign probabilities to the null hypothesis. According to this view of probability, the null hypothesis is either true or it is not, it cannot have a "5% chance" of being true. Second, even within the Bayesian approach, which does let you assign probabilities to hypotheses, the  $p$  value would not correspond to the probability that the null is true. This interpretation is entirely inconsistent with the mathematics of how the  $p$  value is calculated. Put bluntly, despite the intuitive appeal of thinking this way, there is no justification for interpreting a  $p$  value this way. Never do it.

[illegible]



## Reporting the results of a hypothesis test

When writing up the results of a hypothesis test there's usually several pieces of information that you need to report, but it varies a fair bit from test to test. Throughout the rest of the book I'll spend a little time talking about how to report the results of different tests (see Section ?? for a particularly detailed example), so that you can get a feel for how it's usually done. However, regardless of what test you're doing, the one thing that you always have to do is say something about the  $p$  value and whether or not the outcome was significant.

The fact that you have to do this is unsurprising, it's the whole point of doing the test. What might be surprising is the fact that there is some contention over exactly how you're supposed to do it. Leaving aside those people who completely disagree with the entire framework underpinning null hypothesis testing, there's a certain amount of tension that exists regarding whether or not to report the exact  $p$  value that you obtained, or if you should state only that  $p < \alpha$  for a significance level that you chose in advance (e.g.,  $p < .05$ ).

### 6.6.1 The issue

To see why this is an issue, the key thing to recognise is that  $p$  values are *terribly* convenient. In practice, the fact that we can compute a  $p$  value means that we don't actually have to specify any  $\alpha$  level at all in order to run the test. Instead, what you can do is calculate your  $p$  value and interpret it directly. If you get  $p = .062$ , then it means that you'd have to be willing to tolerate a Type I error rate of 6.2% to justify rejecting the null. If you personally find 6.2% intolerable then you retain the null. Therefore, the argument goes, why don't we just report the actual  $p$  value and let the reader make up their own minds about what an acceptable Type I error rate is? This approach has the big advantage of "softening" the decision making process. In fact, if you accept the Neyman definition of the  $p$  value, that's the whole point of the  $p$  value. We no longer have a fixed significance level of  $\alpha = .05$  as a bright line separating "accept" from "reject" decisions, and this removes the rather pathological problem of being forced to treat  $p = .051$  in a fundamentally different way to  $p = .049$ .

This flexibility is both the advantage and the disadvantage to the  $p$  value. The reason why a lot of people don't like the idea of reporting an exact  $p$  value is that it gives the researcher a bit *too much* freedom. In particular, it lets you change your mind about what error tolerance you're willing to put up with *after* you look at the data. For instance, consider my biological sex prediction experiment. Suppose I ran my test and ended up with a  $p$  value of .09. Should I accept or reject? Now, to be honest, I haven't yet bothered to think about what level of Type I error I'm "really" willing to accept. I don't have an opinion on that topic. But I *do* have an opinion about whether or not a baby's sex can be reliably predicted from belly shape, and I *definitely* have an opinion about whether my research should be published in a reputable

scientific journal. And amazingly, now that I've looked at the data I'm starting to think that a 9% error rate isn't so bad, especially when compared to how annoying it would be to have to admit to the world that my experiment has failed. So, to avoid looking like I just made it up after the fact, I now say that my  $\alpha$  is .1, with the argument that a 10% type I error rate isn't too bad and at that level my test is significant! I win.

In other words, the worry here is that I might have the best of intentions, and be the most honest of people, but the temptation to just "shade" things a little bit here and there is really, really strong. As anyone who has ever run an experiment can attest, it's a long and difficult process and you often get *very* attached to your hypotheses. It's hard to let go and admit the experiment didn't find what you wanted it to find. And that's the danger here. If we use the "raw"  $p$ -value, people will start interpreting the data in terms of what they *want* to believe, not what the data are actually saying and, if we allow that, why are we even bothering to do science at all? Why not let everyone believe whatever they like about anything, regardless of what the facts are? Okay, that's a bit extreme, but that's where the worry comes from. According to this view, you really *must* specify your  $\alpha$  value in advance and then only report whether the test was significant or not. It's the only way to keep ourselves honest.

#### 6.6.2 Two proposed solutions

In practice, it's pretty rare for a researcher to specify a single  $\alpha$  level ahead of time. Instead, the convention is that scientists rely on three standard significance levels: .05, .01 and .001. When reporting your results, you indicate which (if any) of these significance levels allow you to reject the null hypothesis. This is summarised in Table 6.1. This allows us to soften the decision rule a little bit, since  $p < .01$  implies that the data meet a stronger evidential standard than  $p < .05$  would. Nevertheless, since these levels are fixed in advance by convention, it does prevent people choosing their  $\alpha$  level after looking at the data.

Nevertheless, quite a lot of people still prefer to report exact  $p$  values. To many people, the advantage of allowing the reader to make up their own mind about how to interpret  $p = .06$  outweighs any disadvantages. In practice, however, even among those researchers who prefer exact  $p$  values it is quite common to just write  $p < .001$  instead of reporting an exact value for small  $p$ . This is in part because a lot of software doesn't actually print out the  $p$  value when it's that small (e.g., SPSS just writes  $p = .000$  whenever  $p < .001$ ), and in part because a very small  $p$  value can be kind of misleading. The human mind sees a number like .0000000001 and it's hard to suppress the gut feeling that the evidence in favour of the alternative hypothesis is a near certainty. In practice however, this is usually wrong. Life is a big, messy, complicated thing, and every statistical test ever invented relies on simplifications, approximations and assumptions. As a consequence, it's probably not reasonable to walk away from *any* statistical analysis with a feeling of confidence stronger than  $p < .001$  implies. In other words,  $p < .001$  is really code for "as far as *this test* is concerned, the evidence is overwhelming."

In light of all this, you might be wondering exactly what you should do. There's a fair

Table 6.1: A commonly adopted convention for reporting  $p$  values: in many places it is conventional to report one of four different things (e.g.,  $p < .05$ ) as shown below. I've included the "significance stars" notation (i.e., a \* indicates  $p < .05$ ) because you sometimes see this notation produced by statistical software. It's also worth noting that some people will write *n.s.* (not significant) rather than  $p > .05$ .

Usual notation	Signif. stars	English translation	The null is...
$p > .05$		The test wasn't significant	Retained
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$ .	Rejected
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$ .	Rejected
$p < .001$	***	The test was significant at all levels	Rejected

bit of contradictory advice on the topic, with some people arguing that you should report the exact  $p$  value, and other people arguing that you should use the tiered approach illustrated in Table 6.1. As a result, the best advice I can give is to suggest that you look at papers/reports written in your field and see what the convention seems to be. If there doesn't seem to be any consistent pattern, then use whichever method you prefer.

## 6.7

### Running the hypothesis test in practice

At this point some of you might be wondering if this is a "real" hypothesis test, or just a toy example that I made up. It's real. In the previous discussion I built the test from first principles, thinking that it was the simplest possible problem that you might ever encounter in real life. However, this test already exists. It's called the *binomial test*, and it's implemented by JASP as one of the statistical analyses available when you hit the 'Frequencies' button. To test the null hypothesis that the response probability is one-half  $p = .5$ ,<sup>9</sup> and using data in which  $x = 62$  of  $n = 100$  people made the correct response, available in the `binomialtest.jasp` data file, we get the results shown in Figure 6.4.

Right now, this output looks pretty unfamiliar to you, but you can see that it's telling you more or less the right things. Specifically, the  $p$ -value of 0.02 is less than the usual choice of

<sup>9</sup>Note that the  $p$  here has nothing to do with a  $p$  value. The  $p$  argument in the JASP binomial test corresponds to the probability of making a correct response, according to the null hypothesis. In other words, it's the  $\theta$  value.

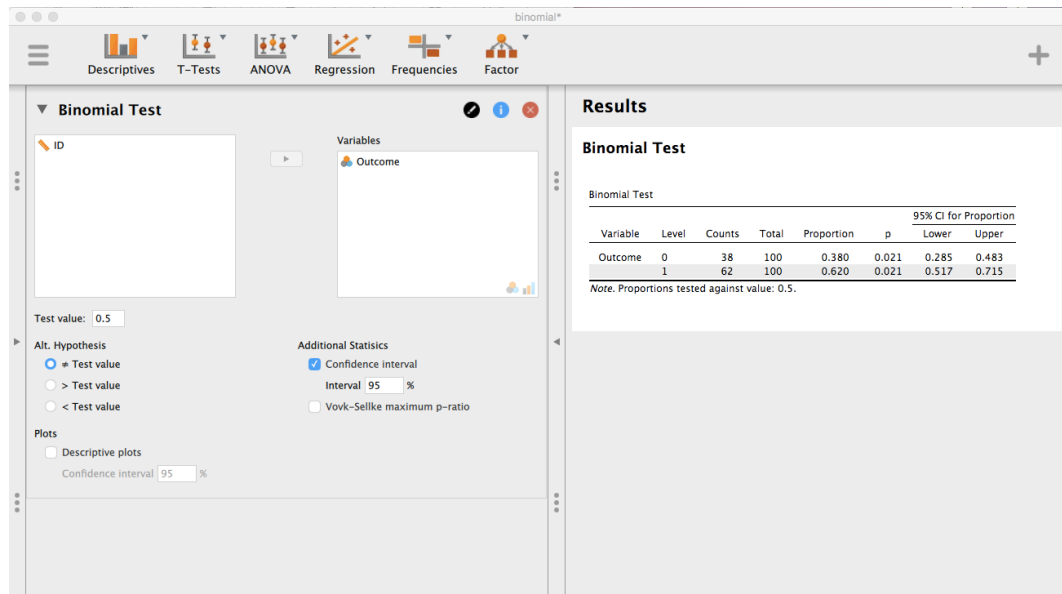


Figure 6.4: Binomial test analysis and results in JASP

$\alpha = .05$ , so you can reject the null. We'll talk a lot more about how to read this sort of output as we go along, and after a while you'll hopefully find it quite easy to read and understand.

## 6.8

### Effect size, sample size and power

In previous sections I've emphasised the fact that the major design principle behind statistical hypothesis testing is that we try to control our Type I error rate. When we fix  $\alpha = .05$  we are attempting to ensure that only 5% of true null hypotheses are incorrectly rejected. However, this doesn't mean that we don't care about Type II errors. In fact, from the researcher's perspective, the error of failing to reject the null when it is actually false is an extremely annoying one. With that in mind, a secondary goal of hypothesis testing is to try to minimise  $\beta$ , the Type II error rate, although we don't usually *talk* in terms of minimising Type II errors. Instead, we talk about maximising the *power* of the test. Since power is defined as  $1 - \beta$ , this is the same thing.

### 6.8.1 The power function

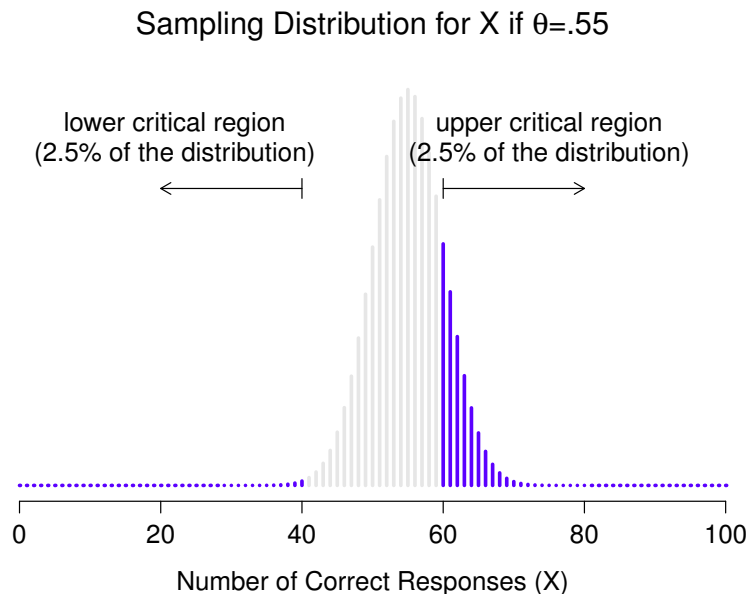


Figure 6.5: Sampling distribution under the *alternative* hypothesis for a population parameter value of  $\theta = 0.55$ . A reasonable proportion of the distribution lies in the rejection region.

.....

Let's take a moment to think about what a Type II error actually is. A Type II error occurs when the alternative hypothesis is true, but we are nevertheless unable to reject the null hypothesis. Ideally, we'd be able to calculate a single number  $\beta$  that tells us the Type II error rate, in the same way that we can set  $\alpha = .05$  for the Type I error rate. Unfortunately, this is a lot trickier to do. To see this, notice that in the study we have been discussing through this chapter, the alternative hypothesis actually corresponds to lots of possible values of  $\theta$ . In fact, the alternative hypothesis corresponds to every value of  $\theta$  *except* 0.5. Let's suppose that the true probability of someone choosing the correct response is 55% (i.e.,  $\theta = .55$ ). If so, then the *true* sampling distribution for  $X$  is not the same one that the null hypothesis predicts, as the most likely value for  $X$  is now 55 out of 100. Not only that, the whole sampling distribution has now shifted, as shown in Figure 6.5. The critical regions, of course, do not change. By definition the critical regions are based on what the null hypothesis predicts. What we're seeing in this figure is the fact that when the null hypothesis is wrong, a much larger proportion of the sampling distribution distribution falls in the critical region. And of course that's what should happen. The probability of rejecting the null hypothesis is larger when the null hypothesis is actually false! However  $\theta = .55$  is not the only possibility consistent with the

alternative hypothesis. Let's instead suppose that the true value of  $\theta$  is actually 0.7. What happens to the sampling distribution when this occurs? The answer, shown in Figure 6.6, is that almost the entirety of the sampling distribution has now moved into the critical region. Therefore, if  $\theta = 0.7$ , the probability of us correctly rejecting the null hypothesis (i.e., the power of the test) is much larger than if  $\theta = 0.55$ . In short, while  $\theta = .55$  and  $\theta = .70$  are both part of the alternative hypothesis, the Type II error rate is different.

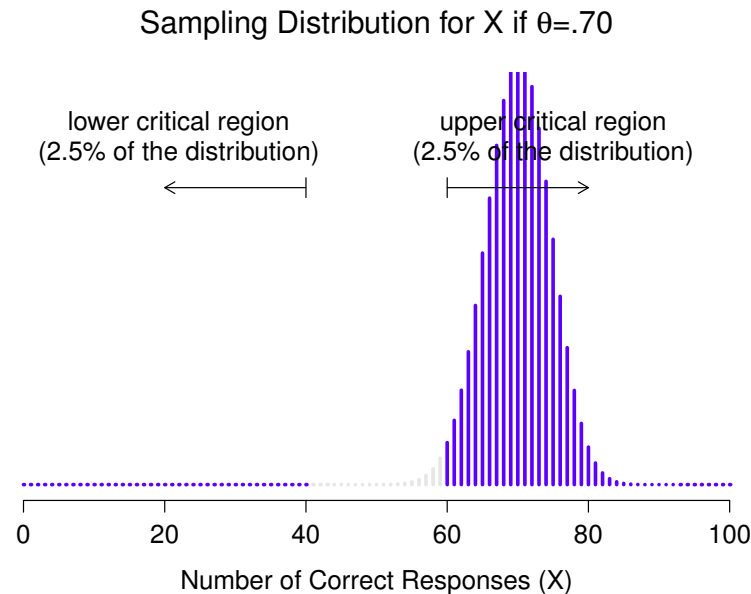


Figure 6.6: Sampling distribution under the *alternative* hypothesis for a population parameter value of  $\theta = 0.70$ . Almost all of the distribution lies in the rejection region.

What all this means is that the power of a test (i.e.,  $1 - \beta$ ) depends on the true value of  $\theta$ . To illustrate this, I've calculated the expected probability of rejecting the null hypothesis for all values of  $\theta$ , and plotted it in Figure 6.7. This plot describes what is usually called the **power function** of the test. It's a nice summary of how good the test is, because it actually tells you the power ( $1 - \beta$ ) for all possible values of  $\theta$ . As you can see, when the true value of  $\theta$  is very close to 0.5, the power of the test drops very sharply, but when it is further away, the power is large.

### 6.8.2 Effect size

*Since all models are wrong the scientist must be alert to what is importantly wrong.*

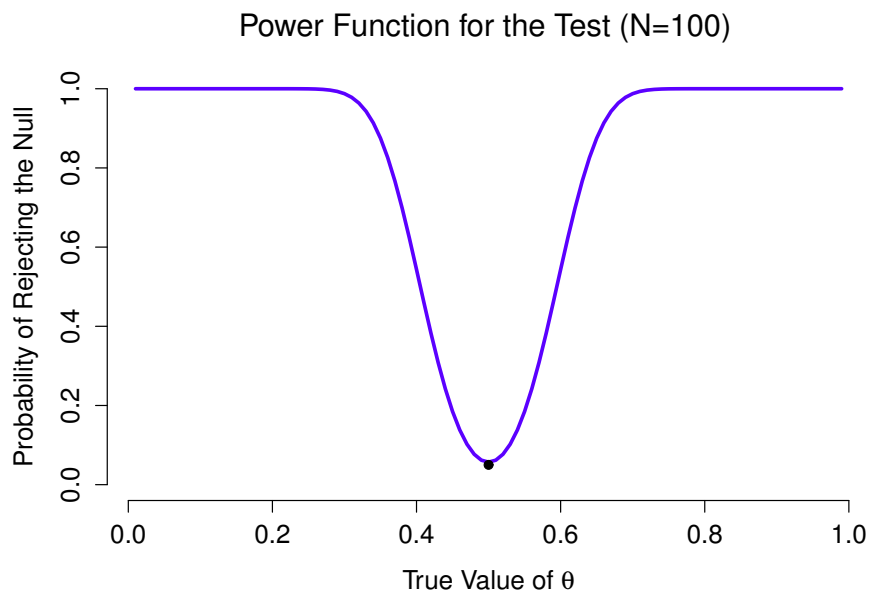


Figure 6.7: The probability that we will reject the null hypothesis, plotted as a function of the true value of  $\theta$ . Obviously, the test is more powerful (greater chance of correct rejection) if the true value of  $\theta$  is very different from the value that the null hypothesis specifies (i.e.,  $\theta = .5$ ). Notice that when  $\theta$  actually is equal to .5 (plotted as a black dot), the null hypothesis is in fact true and rejecting the null hypothesis in this instance would be a Type I error.

.....

*It is inappropriate to be concerned with mice when there are tigers abroad*

– George Box (Box 1976, p. 792)

The plot shown in Figure 6.7 captures a fairly basic point about hypothesis testing. If the true state of the world is very different from what the null hypothesis predicts then your power will be very high, but if the true state of the world is similar to the null (but not identical) then the power of the test is going to be very low. Therefore, it's useful to be able to have some way of quantifying how “similar” the true state of the world is to the null hypothesis. A statistic that does this is called a measure of **effect size** (e.g., Cohen 1988; Ellis 2010). Effect size is defined slightly differently in different contexts (and so this section just talks in general terms) but the qualitative idea that it tries to capture is always the same. How big is the difference between the *true* population parameters and the parameter values that are assumed by the null hypothesis? In our example, if we let  $\theta_0 = 0.5$  denote the value assumed by the null hypothesis and let  $\theta$  denote the true value, then a simple measure of effect size could be something like the difference between the true value and null (i.e.,  $\theta - \theta_0$ ), or possibly

Table 6.2: A crude guide to understanding the relationship between statistical significance and effect sizes. Basically, if you don't have a significant result then the effect size is pretty meaningless because you don't have any evidence that it's even real. On the other hand, if you do have a significant effect but your effect size is small then there's a pretty good chance that your result (although real) isn't all that interesting. However, this guide is very crude. It depends a lot on what exactly you're studying. Small effects can be of massive practical importance in some situations. So don't take this table too seriously. It's a rough guide at best.

	big effect size	small effect size
significant result	difference is real, and of practical importance	difference is real, but might not be interesting
non-significant result	no effect observed	no effect observed
.....		

just the magnitude of this difference,  $\text{abs}(\theta - \theta_0)$ .

Why calculate effect size? Let's assume that you've run your experiment, collected the data, and gotten a significant effect when you ran your hypothesis test. Isn't it enough just to say that you've gotten a significant effect? Surely that's the *point* of hypothesis testing? Well, sort of. Yes, the point of doing a hypothesis test is to try to demonstrate that the null hypothesis is wrong, but that's hardly the only thing we're interested in. If the null hypothesis claimed that  $\theta = .5$  and we show that it's wrong, we've only really told half of the story. Rejecting the null hypothesis implies that we believe that  $\theta \neq .5$ , but there's a big difference between  $\theta = .51$  and  $\theta = .8$ . If we find that  $\theta = .8$ , then not only have we found that the null hypothesis is wrong, it appears to be *very* wrong. On the other hand, suppose we've successfully rejected the null hypothesis, but it looks like the true value of  $\theta$  is only .51 (this would only be possible with a very large study). Sure, the null hypothesis is wrong but it's not at all clear that we actually *care* because the effect size is so small. In the context of my belly shape study we might still care since any demonstration of real psychic powers would actually be pretty cool<sup>10</sup>, but in other contexts a 1% difference usually isn't very interesting, even if it is a real difference. For instance, suppose we're looking at differences in high school exam scores between males and females and it turns out that the female scores are 1% higher on average than the males. If I've got data from thousands of students then this difference will almost certainly be *statistically significant*, but regardless of how small the  $p$  value is it's just not very interesting. You'd hardly want to go around proclaiming a crisis in boys education on the basis of such a tiny difference would you? It's for this reason that it is becoming more

<sup>10</sup>Although in practice a very small effect size is worrying because even very minor methodological flaws might be responsible for the effect, and in practice no experiment is perfect so there are always methodological issues to worry about.



standard (slowly, but surely) to report some kind of standard measure of effect size along with the the results of the hypothesis test. The hypothesis test itself tells you whether you should believe that the effect you have observed is real (i.e., not just due to chance), whereas the effect size tells you whether or not you should care.

### 6.8.3 Increasing the power of your study

Not surprisingly, scientists are fairly obsessed with maximising the power of their experiments. We want our experiments to work and so we want to maximise the chance of rejecting the null hypothesis if it is false (and of course we usually want to believe that it is false!). As we've seen, one factor that influences power is the effect size. So the first thing you can do to increase your power is to increase the effect size. In practice, what this means is that you want to design your study in such a way that the effect size gets magnified. For instance, in my study I might believe that psychic powers work best in a quiet, darkened room with fewer distractions to cloud the mind. Therefore I would try to conduct my experiments in just such an environment. If I can strengthen people's abilities somehow then the true value of  $\theta$  will go up<sup>11</sup> and therefore my effect size will be larger. In short, clever experimental design is one way to boost power, because it can alter the effect size.

Unfortunately, it's often the case that even with the best of experimental designs you may have only a small effect. Perhaps, for example, belly shape really is related to biological sex but even under the best of conditions it's very very weak. Under those circumstances your best bet for increasing power is to increase the sample size. In general, the more observations that you have available, the more likely it is that you can discriminate between two hypotheses. If I ran my experiment with 10 participants and 7 of them correctly guessed the newborn's sex you wouldn't be terribly impressed. But if I ran it with 10,000 participants, and 7,000 of them got the answer right, you would be much more likely to think I had discovered something. In other words, power increases with the sample size. This is illustrated in Figure 6.8, which shows the power of the test for a true parameter of  $\theta = 0.7$  for all sample sizes  $N$  from 1 to 100, where I'm assuming that the null hypothesis predicts that  $\theta_0 = 0.5$ .

Because power is important, whenever you're contemplating running an experiment it would be pretty useful to know how much power you're likely to have. It's never possible to know for sure since you can't possibly know what your real effect size is. However, it's often (well, sometimes) possible to guess how big it should be. If so, you can guess what sample size you need! This idea is called **power analysis**, and if it's feasible to do it then it's very helpful. It can tell you something about whether you have enough time or money to be able to run the experiment successfully. It's increasingly common to see people arguing that power analysis

---

<sup>11</sup>Notice that the true population parameter  $\theta$  doesn't necessarily correspond to an immutable fact of nature. In this context  $\theta$  is just the true probability that people would correctly guess the colour of the card in the other room. As such the population parameter can be influenced by all sorts of things. Of course, this is all on the assumption that such a phenomenon actually exists!

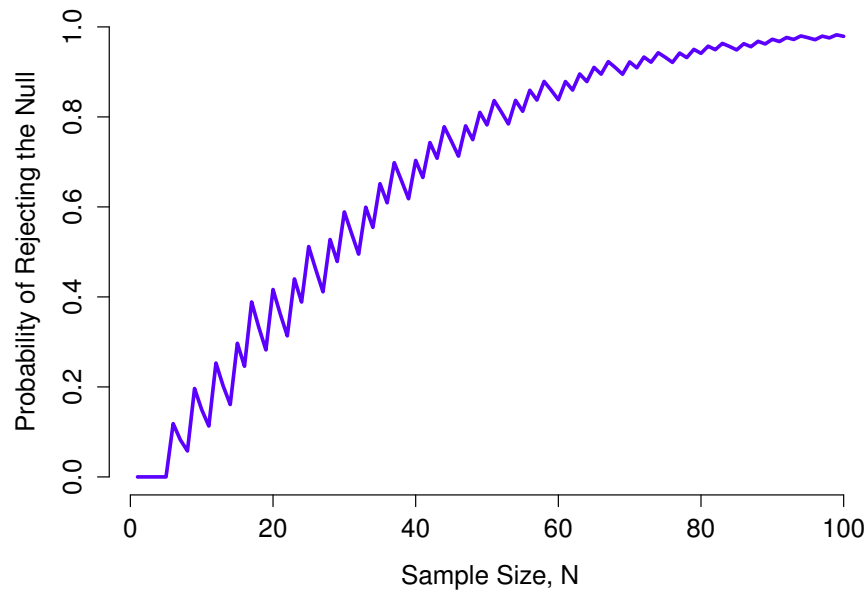


Figure 6.8: The power of our test plotted as a function of the sample size  $N$ . In this case, the true value of  $\theta$  is 0.7 but the null hypothesis is that  $\theta = 0.5$ . Overall, larger  $N$  means greater power. (The small zig-zags in this function occur because of some odd interactions between  $\theta$ ,  $\alpha$  and the fact that the binomial distribution is discrete, it doesn't matter for any serious purpose).

.....

should be a required part of experimental design, so it's worth knowing about. I don't discuss power analysis in this book, however. This is partly for a boring reason and partly for a substantive one. The boring reason is that I haven't had time to write about power analysis yet. The substantive one is that I'm still a little suspicious of power analysis. Speaking as a researcher, I have very rarely found myself in a position to be able to do one. It's either the case that (a) my experiment is a bit non-standard and I don't know how to define effect size properly, or (b) I literally have so little idea about what the effect size will be that I wouldn't know how to interpret the answers. Not only that, after extensive conversations with someone who does stats consulting for a living (my wife, as it happens), I can't help but notice that in practice the *only* time anyone ever asks her for a power analysis is when she's helping someone write a grant application. In other words, the only time any scientist ever seems to want a power analysis in real life is when they're being forced to do it by bureaucratic process. It's not part of anyone's day to day work. In short, I've always been of the view that whilst power is an

important concept, power *analysis* is not as useful as people make it sound, except in the rare cases where (a) someone has figured out how to calculate power for your actual experimental design and (b) you have a pretty good idea what the effect size is likely to be.<sup>12</sup> Maybe other people have had better experiences than me, but I've personally never been in a situation where both (a) and (b) were true. Maybe I'll be convinced otherwise in the future, and probably a future version of this book would include a more detailed discussion of power analysis, but for now this is about as much as I'm comfortable saying about the topic.

## 6.9

---

### Some issues to consider

What I've described to you in this chapter is the orthodox framework for null hypothesis significance testing (NHST). Understanding how NHST works is an absolute necessity because it has been the dominant approach to inferential statistics ever since it came to prominence in the early 20th century. It's what the vast majority of working scientists rely on for their data analysis, so even if you hate it you need to know it. However, the approach is not without problems. There are a number of quirks in the framework, historical oddities in how it came to be, theoretical disputes over whether or not the framework is right, and a lot of practical traps for the unwary. I'm not going to go into a lot of detail on this topic, but I think it's worth briefly discussing a few of these issues.

#### 6.9.1 Neyman versus Fisher

The first thing you should be aware of is that orthodox NHST is actually a mash-up of two rather different approaches to hypothesis testing, one proposed by Sir Ronald Fisher and the other proposed by Jerzy Neyman (see Lehmann 2011, for a historical summary). The history is messy because Fisher and Neyman were real people whose opinions changed over time, and at no point did either of them offer "the definitive statement" of how we should interpret their work many decades later. That said, here's a quick summary of what I take these two approaches to be.

First, let's talk about Fisher's approach. As far as I can tell, Fisher assumed that you only had the one hypothesis (the null) and that what you want to do is find out if the null hypothesis is inconsistent with the data. From his perspective, what you should do is check to see if the data are "sufficiently unlikely" according to the null. In fact, if you remember back to our earlier discussion, that's how Fisher defines the  $p$ -value. According to Fisher, if the null hypothesis provided a very poor account of the data then you could safely reject it. But, since you don't have any other hypotheses to compare it to, there's no way of "accepting the

---

<sup>12</sup>One possible exception to this is when researchers study the effectiveness of a new medical treatment and they specify in advance what an important effect size would be to detect, for example over and above any existing treatment. In this way some information about the potential value of a new treatment can be obtained.

alternative” because you don’t necessarily have an explicitly stated alternative. That’s more or less all there is to it.

In contrast, Neyman thought that the point of hypothesis testing was as a guide to action and his approach was somewhat more formal than Fisher’s. His view was that there are multiple things that you could *do* (accept the null or accept the alternative) and the point of the test was to tell you which one the data support. From this perspective, it is critical to specify your alternative hypothesis properly. If you don’t know what the alternative hypothesis is, then you don’t know how powerful the test is, or even which action makes sense. His framework genuinely requires a competition between different hypotheses. For Neyman, the  $p$  value didn’t directly measure the probability of the data (or data more extreme) under the null, it was more of an abstract description about which “possible tests” were telling you to accept the null, and which “possible tests” were telling you to accept the alternative.

As you can see, what we have today is an odd mishmash of the two. We talk about having both a null hypothesis and an alternative (Neyman), but usually<sup>13</sup> define the  $p$  value in terms of extreme data (Fisher), but we still have  $\alpha$  values (Neyman). Some of the statistical tests have explicitly specified alternatives (Neyman) but others are quite vague about it (Fisher). And, according to some people at least, we’re not allowed to talk about accepting the alternative (Fisher). It’s a mess, but I hope this at least explains why it’s a mess.

### 6.9.2 Bayesians versus frequentists

Earlier on in this chapter I was quite emphatic about the fact that you *cannot* interpret the  $p$  value as the probability that the null hypothesis is true. NHST is fundamentally a frequentist tool (see Chapter ??) and as such it does not allow you to assign probabilities to hypotheses. The null hypothesis is either true or it is not. The Bayesian approach to statistics interprets probability as a degree of belief, so it’s totally okay to say that there is a 10% chance that the null hypothesis is true. That’s just a reflection of the degree of confidence that you have in this hypothesis. You aren’t allowed to do this within the frequentist approach. Remember, if you’re a frequentist, a probability can only be defined in terms of what happens after a large number of independent replications (i.e., a long run frequency). If this is your interpretation of probability, talking about the “probability” that the null hypothesis is true is complete gibberish: a null hypothesis is either true or it is false. There’s no way you can talk about a long run frequency for this statement. To talk about “the probability of the null hypothesis” is as meaningless as “the colour of freedom”. It doesn’t have one!

Most importantly, this *isn’t* a purely ideological matter. If you decide that you are a Bayesian and that you’re okay with making probability statements about hypotheses, you have to follow the Bayesian rules for calculating those probabilities. I’ll talk more about this in Chapter ??, but for now what I want to point out to you is the  $p$  value is a *terrible* approximation

---

<sup>13</sup>Although this book describes both Neyman’s and Fisher’s definition of the  $p$  value, most don’t. Most introductory textbooks will only give you the Fisher version.

to the probability that  $H_0$  is true. If what you want to know is the probability of the null, then the  $p$  value is not what you're looking for!

### 6.9.3 Traps

As you can see, the theory behind hypothesis testing is a mess, and even now there are arguments in statistics about how it “should” work. However, disagreements among statisticians are not our real concern here. Our real concern is practical data analysis. And while the “orthodox” approach to null hypothesis significance testing has many drawbacks, even an unrepentant Bayesian like myself would agree that they can be useful if used responsibly. Most of the time they give sensible answers and you can use them to learn interesting things. Setting aside the various ideologies and historical confusions that we've discussed, the fact remains that the biggest danger in all of statistics is *thoughtlessness*. I don't mean stupidity, I literally mean thoughtlessness. The rush to interpret a result without spending time thinking through what each test actually says about the data, and checking whether that's consistent with how you've interpreted it. That's where the biggest trap lies.

To give an example of this, consider the following example (see Gelman and Stern 2006). Suppose I'm running my pregnant belly study and I've decided to analyse the data separately for the male participants and the female participants (by participants I people people with the power to guess correctly ... not the pregnant individuals!). Of the male participants, 33 out of 50 guessed the baby's sex correctly. This is a significant effect ( $p = .03$ ). Of the female participants, 29 out of 50 guessed correctly. This is not a significant effect ( $p = .32$ ). Upon observing this, it is extremely tempting for people to start wondering why there is a difference between males and females in terms of their psychic abilities. However, this is wrong. If you think about it, we haven't *actually* run a test that explicitly compares males to females. All we have done is compare males to chance (binomial test was significant) and compared females to chance (binomial test was non significant). If we want to argue that there is a real difference between the males and the females, we should probably run a test of the null hypothesis that there is no difference! We can do that using a different hypothesis test,<sup>14</sup> but when we do that it turns out that we have no evidence that males and females are significantly different ( $p = .54$ ). Now do you think that there's anything fundamentally different between the two groups? Of course not. What's happened here is that the data from both groups (male and female) are pretty borderline. By pure chance one of them happened to end up on the magic side of the  $p = .05$  line, and the other one didn't. That doesn't actually imply that males and females are different. This mistake is so common that you should always be wary of it. The difference between significant and not-significant is *not* evidence of a real difference. If you want to say that there's a difference between two groups, then you have to test for that difference!

The example above is just that, an example. I've singled it out because it's such a common one, but the bigger picture is that data analysis can be tricky to get right. Think about what

---

<sup>14</sup>In this case, the Pearson chi-square test of independence (Chapter ??)

it is you want to test, why you want to test it, and whether or not the answers that your test gives could possibly make any sense in the real world.

## 6.10

---

### Summary

Null hypothesis testing is one of the most ubiquitous elements to statistical theory. The vast majority of scientific papers report the results of some hypothesis test or another. As a consequence it is almost impossible to get by in science without having at least a cursory understanding of what a  $p$ -value means, making this one of the most important chapters in the book. As usual, I'll end the chapter with a quick recap of the key ideas that we've talked about:

- Research hypotheses and statistical hypotheses. Null and alternative hypotheses. (Section 6.1).
- Type 1 and Type 2 errors (Section 6.2)
- Test statistics and sampling distributions (Section 6.3)
- Hypothesis testing as a decision making process (Section 6.4)
- $p$ -values as “soft” decisions (Section 6.5)
- Writing up the results of a hypothesis test (Section 6.6)
- Running the hypothesis test in practice (Section 6.7)
- Effect size and power (Section 6.8)
- A few issues to consider regarding hypothesis testing (Section 6.9)

Later in the book, in Chapter ??, I'll revisit the theory of null hypothesis tests from a Bayesian perspective and introduce a number of new tools that you can use if you aren't particularly fond of the orthodox approach. But for now, though, we're done with the abstract statistical theory, and we can start discussing specific data analysis tools.

LEARNING STATISTICS  
**WITH JASP**