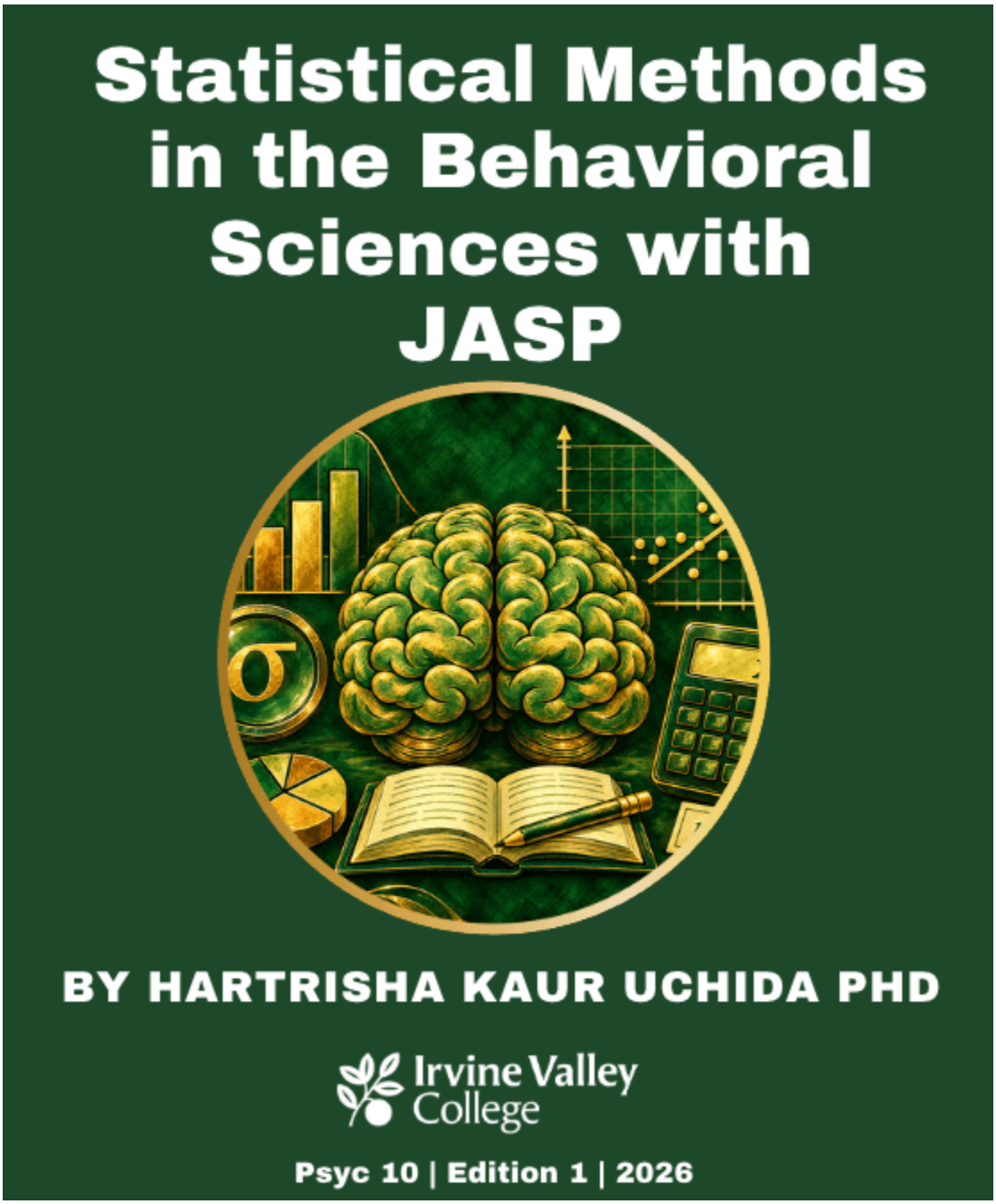
STATISTICAL METHODS IN THE BEHAVIORAL SCIENCES WITH JASP (PSYCH 10 AT IVC)



Hartrisha Kaur Uchida
Irvine Valley College

LibreTexts™



Statistical Methods in the Behavioral Sciences with JASP (Psych 10 at IVC)

Hartrisha Kaur Uchida PhD
Irvine Valley College

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden borne from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and were supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

The LibreTexts name, logo, and book covers are trademarked by LibreTexts, Inc. (a 501(c)(3) not-for-profit organization) and protected against unauthorized use.

This text was compiled on 11/05/2026

TABLE OF CONTENTS

Licensing

1: Introduction to Statistics

1.1: Introduction to this Textbook

1.2: Why do we learn statistics?

1.3: Why JASP?

1.4: Communicating Statistical Results

2: Research Design and Measurement

2.1: A brief introduction to research design

3: Getting Started with JASP

3.1: Getting started with JASP

3.2: Installing and Opening JASP

3.3: Accessibility and Other Settings

3.4: Creating Data Files in JASP

3.5: Importing Data from Excel into JASP

4: Descriptive Statistics

4.1: Descriptive statistics

4.2: Descriptive Statistics in JASP

5: Visualizing Data

5.1: Drawing graphs

5.2: Basic Plots

5.3: Tables

5.4: Exporting and Editing Plots from JASP

6: Probability and Distributions

6.1: Introduction to probability

7: Sampling and Estimation

7.1: Estimating unknown quantities from a sample

8: Hypothesis Testing

8.1: Hypothesis testing

9: Chi-Square Tests

9.1: Categorical data analysis

9.2: Chi-squared Test

10: Comparing Means - z-Tests and t-Tests

10.1: Comparing two means

10.2: Independent t-test and Mann-Whitney U

10.3: Paired t-test and Wilcoxon Signed Ranks

10.4: Worked Example- Calculating an Independent Samples t-Test

11: Correlation and Linear Regression

11.1: Correlation and linear regression

11.2: Pearson and Spearman Correlations

11.3: Calculating and Reporting Pearson's r

12: One-Way ANOVA

12.1: Comparing several means (one-way ANOVA)

13: Factorial ANOVA - Main Effects and Interactions

13.1: Factorial ANOVA

13.2: Worked Example - Calculating Sums of Squares for a Two-Factor ANOVA

13.3: Worked Example - Completing and Reporting a Two-Factor ANOVA

Index

Glossary

Detailed Licensing

This page titled Statistical Methods in the Behavioral Sciences with JASP (Psych 10 at IVC) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

Licensing

A detailed breakdown of this resource's licensing can be found in ***Back Matter/Detailed Licensing***.

This page titled Licensing is shared under a Public Domain license and was authored, remixed, and/or curated by LibreTexts.

1: INTRODUCTION TO STATISTICS

1.1: Introduction to this Textbook
1.2: Why do we learn statistics?
1.3: Why JASP?
1.4: Communicating Statistical Results

This page titled 1: Introduction to Statistics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

1.1: Introduction to this Textbook

Statistics helps us make sense of data and evaluate the claims we encounter in research and everyday life. In the behavioral sciences, we use statistics to study how people think, feel, and behave. Whether we are examining relationships between variables, comparing groups, or evaluating an intervention, statistics gives us tools for asking questions and interpreting evidence.

This textbook introduces the fundamentals of statistics for the behavioral sciences, with a particular emphasis on psychology. It is intended for first- and second-year college students, including students who have never taken a statistics course or used statistical software. You do not need to arrive already knowing how to analyze data. We will build those skills as we move through the book.

We begin with research design, measurement, descriptive statistics, and graphs. From there, we explore probability, sampling, estimation, and hypothesis testing before introducing specific analyses, including chi-square tests, t-tests, correlation, regression, and analysis of variance (ANOVA). Throughout these topics, we will connect statistical ideas to the questions researchers ask about behavior.

You will learn some calculations by hand so you can see how a statistical result is built from the data. You will also learn how to use JASP, a free statistical software program that allows you to analyze data without writing code. These approaches serve related purposes: hand calculations help explain what an analysis is doing, while software allows us to work efficiently with data and examine results in greater detail.

Calculating a statistic or producing a table of output is only part of the process. You will also learn to choose appropriate analyses, consider their assumptions, interpret results, and explain findings in context. A statistically significant result does not automatically mean that a finding is large, practically important, or evidence of causation. Learning to recognize these distinctions is part of becoming a thoughtful reader and user of statistics.

How This Textbook Was Developed

This textbook brings together material from several openly licensed sources, along with examples and explanations added for this course. Its primary sources are *Learning Statistics with JASP* by Navarro, Foxcroft, and Faulkenberry and *Statistics with JASP: First Steps for Psychology Students* by Denniss and Naneva. Additional sources provide selected worked examples and explanations.

Because the book combines contributions from different authors, you may notice changes in voice, terminology, notation, or formatting. Some sections explain concepts conversationally, while others provide more direct instructions or mathematical detail. You may also encounter more than one explanation of a concept. These differences reflect the sources from which the book was assembled. You may also notice regional differences in spelling. These differences do not affect the accuracy of the concepts being explained. When terminology or notation varies, focus on what each term or symbol represents and how it connects to what you have already learned.

Sources and licensing information are provided throughout the textbook to acknowledge the original authors and support future adaptation of their work. Openly licensed materials make it possible to share, revise, and build on educational resources while keeping them freely available to students.

Using This Textbook

Statistics takes practice. As you read, work through the examples, follow the calculations, and try the JASP procedures yourself. Pause to consider why a particular analysis was selected and what its results tell us about the original question. If an explanation does not make sense the first time, return to the example and work through it one step at a time.

The goal is for you to leave this course able to understand common statistical analyses, use JASP to examine data, and communicate findings accurately. These skills will support your work in psychology and related fields, and help you evaluate the evidence behind claims you encounter beyond the classroom.

This page titled 1.1: Introduction to this Textbook is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

1.2: Why do we learn statistics?

"Thou shalt not answer questionnaires
Or quizzes upon World Affairs,
Nor with compliance
Take any test. Thou shalt not sit
With statisticians nor commit
A social science" -W.H. Auden [1]

On the psychology of statistics

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

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

Why do you do statistics? Why don't scientists just use common sense?
It's a naive question in some ways, but most good questions are. There are a lot of good answers to it, [2] but for my money, the best answer is a really simple one: we don't trust ourselves enough. We worry that we're human, and susceptible to all of the biases, temptations and frailties that humans suffer from. Much of statistics is basically a safeguard. Using "common sense" to evaluate evidence means trusting gut instincts, relying on verbal arguments and on using the raw power of human reason to come up with the right answer. Most scientists don't think this approach is likely to work.

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

The curse of belief bias

People are mostly pretty smart. We're certainly smarter than the other species that we share the planet with (though many people might disagree). Our minds are quite amazing things, and we seem to be capable of the most incredible feats of thought and reason. That doesn't make us perfect though. And among the many things that psychologists have shown over the years is that we really do find it hard to be neutral, to evaluate evidence impartially and without being swayed by pre-existing biases. A good example of this is the belief bias effect in logical reasoning: if you ask people to decide whether a particular argument is logically valid (i.e., conclusion would be true if the premises were true), we tend to be influenced by the believability of the conclusion, even when we shouldn't. For instance, here's a valid argument where the conclusion is believable:

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"

If the psychological data looked like this (or even a good approximation to this), we might feel safe in just trusting our gut instincts. That is, it'd be perfectly okay just to let scientists evaluate data based on their common sense, and not bother with all this murky statistics stuff. However, you guys have taken psych classes, and by now you probably know where this is going.

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

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

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

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

Oh dear, that's not as good. Apparently, when people are presented with a strong argument that contradicts our pre-existing beliefs, we find it pretty hard to even perceive it to be a strong argument (people only did so 46% of the time). Even worse, when people are presented with a weak argument that agrees with our pre-existing biases, almost no-one can see that the argument is weak (people got that one wrong 92% of the time!). [3]

If you think about it, it's not as if these data are horribly damning. Overall, people did do better than chance at compensating for their prior biases, since about 60% of people's judgments were correct (you'd expect 50% by chance). Statistical methods give us a structured way to evaluate evidence rather than relying entirely on intuition. They do not eliminate bias or guarantee a correct conclusion, but they can help us identify patterns, account for uncertainty, and evaluate whether the data support our expectations.

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! [4] 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	
Department	Applicants	Percent Admitted	Applicants	Percent Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

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

Here's what's going on. First, notice that admission rates differ considerably across departments. Departments A and B have relatively high admission rates, while Department F has relatively low admission rates. Applicants were not distributed evenly across these departments. 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 $\mathbf{A} > \mathbf{B} > \mathrm{D} > \mathrm{C} > \mathrm{F} > \mathrm{E}$ (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 $\mathrm{C} > \mathrm{E} > \mathrm{D} > \mathrm{F} > \mathrm{A} > \mathrm{B}$. In other words, what these data seem to be suggesting is that the female applicants tended to apply to "harder" departments. And in fact, if we look at Figure 1.1 we see that this trend is systematic, and quite striking. This effect is known as Simpson's paradox. It's not common, but it does happen in real life, and most people are very surprised by it when they first encounter it, and many people refuse to even believe that it's real. It is very real. And while there are lots of very subtle statistical lessons buried in there, I want to use it to make a much more important point: doing research is hard, and there are lots

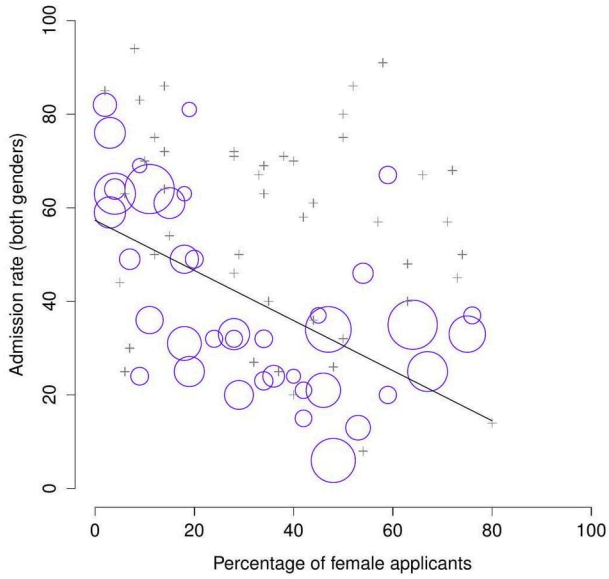


Figure 1.1: The Berkeley 1973 college admissions data. This figure plots the admission rate for the 85 departments that had at least one female applicant, as a function of the percentage of applicants that were female. The plot is a redrawing of Figure 1 from Bickel, Hammel, and O'Connell (1975). Circles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

Looking at admission rates within departments changes the interpretation of the overall difference. However, these descriptive results alone do not establish that the admissions process was fair or explain why applicants were distributed differently across departments. Those questions require additional evidence and careful attention to context.

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

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

Statistics in psychology

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

- Why does psychology have so much statistics?
  To be perfectly honest, there's a few different reasons, some of which are better than others. The most important reason is

that psychology is a statistical science. What I mean by that is that the "things" that we study are people. Real, complicated, gloriously messy, infuriatingly perverse people. The "things" of physics include objects like electrons, and while there are all sorts of complexities that arise in physics, electrons don't have minds of their own. They don't have opinions, they don't differ from each other in weird and arbitrary ways, they don't get bored in the middle of an experiment, and they don't get angry at the experimenter and then deliberately try to sabotage the data set (not that I've ever done that!). At a fundamental level psychology is harder than physics. [5]

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

- Can't someone else do the statistics?
  To some extent, but not completely. It's true that you don't need to become a fully trained statistician just to do psychology, but you do need to reach a certain level of statistical competence. In my view, there's three reasons that every psychological researcher ought to be able to do basic statistics:
  - Firstly, there's the fundamental reason: statistics is deeply intertwined with research design. If you want to be good at designing psychological studies, you need to at the very least understand the basics of stats.
  - Secondly, if you want to be good at the psychological side of the research, then you need to be able to understand the psychological literature, right? But almost every paper in the psychological literature reports the results of statistical analyses. So if you really want to understand the psychology, you need to be able to understand what other people did with their data. And that means understanding a certain amount of statistics.
  - Thirdly, understanding statistics helps you work independently and collaborate effectively with others. Even when a specialist conducts an analysis, you need enough statistical knowledge to explain your research question, understand the methods used, and evaluate the results.

- Note that a lot of these reasons generalize beyond researchers. If you want to be a practicing psychologist and stay on top of the field, it helps to be able to read the scientific literature, which relies pretty heavily on statistics.
- I don't care about jobs, research, or clinical work. Do I need statistics?
  Okay, now you're just messing with me. Still, I think it should matter to you too. Statistics should matter to you in the same way that statistics should matter to everyone. We live in the 21st century, and data are everywhere. Frankly, given the world in which we live these days, a basic knowledge of statistics is pretty damn close to a survival tool! Which is the topic of the next section.

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.

There's more to research methods than statistics

So far, most of what I've talked about is statistics, and so you'd be forgiven for thinking that statistics is all I care about in life. To be fair, you wouldn't be far wrong, but research methodology is a
broader concept than statistics. So most research methods courses will cover a lot of topics that relate much more to the pragmatics of research design, and in particular the issues that you encounter when trying to do research with humans. However, about 99% of student fears relate to the statistics part of the course, so I've focused on the stats in this discussion, and hopefully I've convinced you that statistics matters, and more importantly, that it's not to be feared. That being said, it's pretty typical for introductory research methods classes to be very stats-heavy. This is not (usually) because the lecturers are evil people. Quite

the contrary, in fact. Introductory classes focus a lot on the statistics because you almost always find yourself needing statistics before you need the other research methods training. Why? Because almost all of your assignments in other classes will rely on statistical training, to a much greater extent than they rely on other methodological tools. It's not common for undergraduate assignments to require you to design your own study from the ground up (in which case you would need to know a lot about research design), but it is common for assignments to ask you to analyse and interpret data that were collected in a study that someone else designed (in which case you need statistics). In that sense, from the perspective of allowing you to do well in all your other classes, the statistics is more urgent.

But note that "urgent" is different from "important" - they both matter. I really do want to stress that research design is just as important as data analysis, and this book does spend a fair amount of time on it. However, while statistics has a kind of universality, and provides a set of core tools that are useful for most types of psychological research, the research methods side isn't quite so universal. There are some general principles that everyone should think about, but a lot of research design is very idiosyncratic, and is specific to the area of research that you want to engage in. To the extent that it's the details that matter, those details don't usually show up in an introductory stats and research methods class.

---

1. [1] The quote comes from Auden's 1946 poem Under Which Lyre: A Reactionary Tract for the Times, delivered as part of a commencement address at Harvard University. The history of the poem is kind of interesting: http://harvardmagazine.com/2007/11/a-poets-warning.html
2. [2] Including the suggestion that common sense is in short supply among scientists.
3. [3] In my more cynical moments I feel like this fact alone explains 95% of what I read on the internet.
4. [4] 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.
5. [5] Which might explain why physics is just a teensy bit further advanced as a science than we are.

This page titled 1.2: Why do we learn statistics? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 1: Why do we learn statistics? by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

1.3: Why JASP?

This textbook uses JASP to introduce statistical software without requiring previous experience with coding. The practical instructions adapted from Denniss provide step-by-step guidance, while the surrounding chapters explain the concepts, assumptions, and interpretation of each analysis. Because the source materials were developed using different versions of JASP, some screenshots or menu options may differ slightly from the version you are using.

'Why JASP? What is wrong with SPSS?'

There is nothing particularly 'wrong' with SPSS as it is an excellent statistics package that does everything that you would want, however for many students it is confusing and creates a barrier to learning. The barriers arise because of the complex menu structure that is not particularly intuitive, and from the lack of a clear link between which menu items you select and the output that appears. JASP addresses both barriers; the menu structure is very straightforward, and the output immediately appears on the right-hand side of the screen. Learners can 'play' in JASP and try things out, knowing that any button they click can be unclicked and the output updated, without ending up with pages of outputs (like in SPSS) which leave learners feeling lost as to which is the output they should be looking at. In JASP outputs update automatically and so you can easily see the links between the functions and the output.

As use of JASP is more intuitive that SPSS you can focus more on the 'why' of the analysis, rather than the 'how'.

What is JASP?

JASP is a graphical user interface (GUI) built on an R platform, R being another package for statistical analysis that you code; you can see the R code behind each of the functions in JASP. So why not just teach R? For some of the same pedagogical reasons that SPSS is problematic – here students would need to first learn how to code before they could get to grips with the analysis (the 'how'), rather than focusing on the 'why' of the analysis. This could act as a barrier to learning for those for whom coding just doesn't gel. The accessibility of the R code in JASP means that learners interested in R coding can get a little of a springboard into learning this, and learners who just want to be able to do the analysis in the most easily digestible way possible can completely ignore the R code. Everybody wins!

We hope that you enjoy using this guide, and if you have found it useful, please let us know (contact details can be found in About the Authors).

Copyright statement

Statistics with JASP: First Steps for Psychology Students copyright © 2025 by The Authors and the University of Sheffield. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except as noted otherwise here or in the text. Software shown in screenshots copyright the respective owners, for reuse permissions contact the owner(s). JASP software copyright © 2025 The JASP Team, images reproduced by permission. Microsoft Excel copyright © Microsoft Corporation.

This page titled 1.3: Why JASP? is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss via source content that was edited to the style and standards of the LibreTexts platform.

1.4: Communicating Statistical Results

Communicating Statistical Results

Throughout this course, you will learn how to calculate statistics by hand and conduct analyses in JASP. You will also learn how to explain what those results mean. A calculation or a table of output is only part of the answer. We need to connect the results back to the question we are trying to answer.

Psychologists typically use American Psychological Association (APA) style to report statistical findings. This gives us a consistent way to communicate the analysis we conducted, the results we obtained, and their meaning in the context of the study.

You will encounter several statistical symbols throughout this textbook:

Symbol	What it represents
M	Mean
SD	Standard deviation
t	Test statistic for a t-test
F	Test statistic for an ANOVA
r	Pearson's correlation coefficient
p	p value
df	Degrees of freedom

Most of these symbols are italicized in APA style. Greek symbols, such as χ^2 and η^2, are not italicized. You do not need to understand every statistic in this table yet. We will introduce each one as we work through the relevant chapters.

A statistical write-up generally identifies the analysis, reports the relevant numerical results, and explains the finding in context. Depending on the analysis, this may include group means and standard deviations, a test statistic, degrees of freedom, a p value, and an effect size.

For example, if we are comparing two groups, we should explain which group had the higher mean. If we are examining a correlation, we should explain whether higher scores on one variable were associated with higher or lower scores on the other. Saying that a result was "statistically significant" does not give the reader enough information to understand the finding.

We typically report exact p values to three decimal places, such as $p = .032$. When a p value is smaller than .001, report $p < .001$. If JASP displays a rounded value of .000, do not write $p = .000$.

You will find APA-style reporting examples alongside the analyses in this textbook. Use these examples as guides, adapting the variable names, numerical results, and interpretation to your own analysis. As you write, ask yourself: **Have I told the reader what I analyzed, what I found, and what that finding means?**

This page titled 1.4: Communicating Statistical Results is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

2: RESEARCH DESIGN AND MEASUREMENT

2.1: A brief introduction to research design

This page titled 2: Research Design and Measurement is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

2.1: A brief introduction to research design

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

In this chapter, we're going to start thinking about the basic ideas that go into designing a study, collecting data, checking whether your data collection works, and so on. It won't give you enough information to allow you to design studies of your own, but it will give you a lot of the basic tools that you need to assess the studies done by other people. However, since the focus of this book is much more on data analysis than on data collection, I'm only giving a very brief overview. Note that this chapter is "special" in two ways. Firstly, it's much more psychology-specific than the later chapters. Secondly, it focuses much more heavily on the scientific problem of research methodology, and much less on the statistical problem of data analysis. Nevertheless, the two problems are related to one another, so it's traditional for stats textbooks to discuss the problem in a little detail. This chapter relies heavily on Campbell and Stanley (1963) for the discussion of study design, and Stevens (1946) for the discussion of scales of measurement.

2.1 Introduction to psychological measurement

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

2.1.1 Some thoughts about psychological measurement

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

- My age is 33 years.
- I do not like anchovies.
- My chromosomal gender is male.
- My self-identified gender is male. [2]

In the short list above, the bolded part is "the thing to be measured", and the italicised part is "the measurement itself". In fact, we can expand on this a little bit, by thinking about the set of possible measurements that could have arisen in each case:

- My age (in years) could have been $0, 1, 2, 3 \ldots$, etc. The upper bound on what my age could possibly be is a bit fuzzy, but in practice you'd be safe in saying that the largest possible age is 150, since no human has ever lived that long.
- When asked if I like anchovies, I might have said that I do, or I do not, or I have no opinion, or I sometimes do.
- My chromosomal gender is almost certainly going to be male (XY) or female (XX), but there are a few other possibilities. I could also have Klinfelter's syndrome (XXY), which is more similar to male than to female. And I imagine there are other possibilities too.
- My self-identified gender is also very likely to be male or female, but it doesn't have to agree with my chromosomal gender. I may also choose to identify with neither, or to explicitly call myself transgender.

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

Looking at this a bit more closely, you might also realise that the concept of "age" isn't actually all that precise. In general, when we say "age" we implicitly mean "the length of time since birth". But that's not always the right way to do it. Suppose you're interested in how newborn babies control their eye movements. If you're interested in kids that young, you might also start to worry that "birth" is not the only meaningful point in time to care about. If Baby Alice is born 3 weeks premature and Baby Bianca is born 1 week late, would it really make sense to say that they are the "same age" if we encountered them "2 hours after birth"? In one sense, yes. By social convention we use birth as our reference point for talking about age in everyday life, since it defines

the amount of time the person has been operating as an independent entity in the world. But from a scientific perspective that's not the only thing we care about. When we think about the biology of human beings, it's often useful to think of ourselves as organisms that have been growing and maturing since conception, and from that perspective Alice and Bianca aren't the same age at all. So you might want to define the concept of "age" in two different ways: the length of time since conception and the length of time since birth. When dealing with adults it won't make much difference, but when dealing with newborns it might.

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

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

2.1.2 Operationalisation: defining your measurement

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

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

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

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

- A theoretical construct. This is the thing that you're trying to take a measurement of, like "age", "gender" or an "opinion". A theoretical construct can't be directly observed, and often they're actually a bit vague.
- A measure. The measure refers to the method or the tool that you use to make your observations. A question in a survey, a behavioural observation or a brain scan could all count as a measure.
- An operationalisation. The term "operationalisation" refers to the logical connection between the measure and the theoretical construct, or to the process by which we try to derive a measure from a theoretical construct.

- A variable. Finally, a new term. A variable is what we end up with when we apply our measure to something in the world. That is, variables are the actual "data" that we end up with in our data sets.

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

2.2 Scales of measurement

As the previous section indicates, the outcome of a psychological 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 (also referred to as a categorical variable) is one in which there is no particular relationship between the different possibilities. For these kinds of variables it doesn't make any sense to say that one of them is "bigger' or "better" than any other one, and it absolutely doesn't make any sense to average them. The classic example for this is "eye colour". Eyes can be blue, green or brown, amongst other possibilities, but none of them is any "bigger" than any other one. As a result, it would feel really weird to talk about an "average eye colour". Similarly, gender is nominal too: male isn't better or worse than female. Neither does it make sense to try to talk about an "average gender". In short, nominal scale variables are those for which the only thing you can say about the different possibilities is that they are different. That's it.

Let's take a slightly closer look at this. Suppose I was doing research on how people commute to and from work. One variable I would have to measure would be what kind of transportation people use to get to work. This "transport type" variable could have quite a few possible values, including: "train", "bus", "car", "bicycle". For now, let's suppose that these four are the only possibilities. Then imagine that I ask 100 people how they got to work today, with this result:

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

So, what's the average transportation type? Obviously, the answer here is that there isn't one.

It's a silly question to ask. You can say that travel by car is the most popular method, and travel by train is the least popular method, but that's about all. Similarly, notice that the order in which I list the options isn't very interesting. I could have chosen to display the data like this...

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

...and nothing really changes.

2.2.2 Ordinal scale

Ordinal scale variables have a bit more structure than nominal scale variables, but not by a lot. An ordinal scale variable is one in which there is a natural, meaningful way to order the different possibilities, but you can't do anything else. The usual example given of an ordinal variable is "finishing position in a race". You can say that the person who finished first was faster than the person who finished second, but you don't know how much faster. As a consequence we know that 1st > 2nd, and we know that 2nd > 3rd, but the difference between 1st and 2nd might be much larger than the difference between 2nd and 3rd.

Here's a more psychologically interesting example. Suppose I'm interested in people's attitudes to climate change. I then go and ask some people to pick the statement (from four listed statements) that most closely matches their beliefs:

- (1)Temperatures are rising because of human activity
- (2)Temperatures are rising but we don't know why
- (3)Temperatures are rising but not because of humans
- (4)Temperatures are not rising

Notice that these four statements actually do have a natural ordering, in terms of "the extent to which they agree with the current science". Statement 1 is a close match, statement 2 is a reasonable match, statement 3 isn't a very good match, and statement 4 is in strong opposition to current science. So, in terms of the thing I'm interested in (the extent to which people endorse the science), I can order the items as 1 > 2 > 3 > 4. Since this ordering exists, it would be very weird to list the options like this...

- (3)Temperatures are rising but not because of humans
- (1)Temperatures are rising because of human activity
- (4)Temperatures are not rising
- (2)Temperatures are rising but we don't know why

...because it seems to violate the natural "structure" to the question.

So, let's suppose I asked 100 people these questions, and got the following answers:

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

When analysing these data it seems quite reasonable to try to group (1), (2) and (3) together, and say that 81 out of 100 people were willing to at least partially endorse the science. And it's also quite reasonable to group (2), (3) and (4) together and say that 49 out of 100 people registered at least some disagreement with the dominant scientific view. However, it would be entirely bizarre to try to group (1), (2) and (4) together and say that 90 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. It is not exactly clear what this average means; certainly, the value 1.97 is not a valid response to the survey.

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 is measuring temperature in degrees celsius. For instance, if it was 15° yesterday and 18° today, then the 3° difference between the two is genuinely meaningful. Moreover, that 3° difference is exactly the same as the 3° difference between 7° and 10°. In short, addition and subtraction are meaningful for interval scale variables. [3]

However, notice that the 0° does not mean "no temperature at all". It actually means "the temperature at which water freezes", which is pretty arbitrary. As a consequence it becomes pointless to try to multiply and divide temperatures. It is wrong to say that 20° is twice as hot as 10°, just as it is weird and meaningless to try to claim that 20° is negative two times as hot as -10°.

Again, lets look at a more psychological example. Suppose I'm interested in looking at how the attitudes of first-year university students have changed over time. Obviously, I'm going to want to record the year in which each student started. This is an interval scale variable. A student who started in 2003 did arrive 5 years before a student who started in 2008. However, it would be completely 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 fourth and final type of variable to consider is a ratio scale variable, in which zero really means zero, and it's okay to multiply and divide. A good psychological example of a ratio scale variable is response time (RT). In a lot of tasks it's very common to record the amount of time somebody takes to solve a problem or answer a question, because it's an indicator of how difficult the task is. Suppose that Alan takes 2.3 seconds to respond to a question, whereas Ben takes 3.1 seconds. As with an interval scale variable, addition and subtraction are both meaningful here. Ben really did take $3.1 - 2.3 = 0.8$ seconds longer than Alan did. However, notice that multiplication and division also make sense here too: Ben took $3.1/2.3 = 1.35$ times as long as Alan did to answer the question. And the reason why you can do this is that for a ratio scale variable such as RT "zero seconds" really does mean "no time at all".

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're pretty simple once you see some examples. For instance, response time is continuous. If Alan takes 3.1 seconds and Ben takes 2.3 seconds to respond to a question, then Cameron's response time will lie in between if he took 3.0 seconds. And of course it would also be possible for David to take 3.031 seconds to respond, meaning that his RT would lie in between Cameron's and Alan's. And while in practice it might be impossible to measure RT that precisely, it's certainly possible in principle. Because we can always find a new value for RT in between any two other ones we regard RT as a continuous measure.

Discrete variables occur when this rule is violated. For example, nominal scale variables are always discrete. There isn't a type of transportation that falls "in between" trains and bicycles, not in the strict mathematical way that 2.3 falls in between 2 and 3. So transportation type is discrete. Similarly, ordinal scale variables are always discrete. Although "2nd place" does fall between "1st place" and "3rd place", there's nothing that can logically fall in between "1st place" and "2nd place". Interval scale and ratio scale variables can go either way. As we saw above, response time (a ratio scale variable) is continuous. Temperature in degrees celsius (an interval scale variable) is also continuous. However, the year you went to school (an interval scale variable) is discrete. There's no year in between 2002 and 2003. The number of questions you get right on a true-or-false test (a ratio scale variable) is also discrete. Since a true-or-false question doesn't allow you to be "partially correct", there's nothing in between 5/10 and 6/10. 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". It's very unfortunate.

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	✓	✓

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.

So let's take a classic example, maybe the classic example, of a psychological measurement tool: the Likert scale. The humble Likert scale is the bread and butter tool of all survey design. You yourself have filled out hundreds, maybe thousands, of them and odds are you've even used one yourself. Suppose we have a survey question that looks like this:

Which of the following best describes your opinion of the statement that "all pirates are freaking awesome"?
and then the options presented to the participant are these:

- (1)Strongly disagree
- (2)Disagree
- (3)Neither agree nor disagree
- (4)Agree
- (5)Strongly agree

This set of items is an example of a 5-point Likert scale, in which people are asked to choose among
one of several (in this case 5) clearly ordered possibilities, generally with a verbal descriptor given in each case. However, it's not necessary that all items are explicitly described. This is a perfectly good example of a 5-point Likert scale too:

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

Likert scales are very handy, if somewhat limited, tools. The question is what kind of variable are they? They're obviously discrete, since you can't give a response of 2.5. They're obviously not nominal scale, since the items are ordered; and they're not ratio scale either, since there's no natural zero.

But are they ordinal scale or interval scale? One argument says that we can't really prove that the difference between "strongly agree" and "agree" is of the same size as the difference between "agree" and "neither agree nor disagree". In fact, in everyday life it's pretty obvious that they're not the same at all. So this suggests that we ought to treat Likert scales as ordinal variables. On the other hand, in practice most participants do seem to take the whole "on a scale from 1 to 5" part fairly seriously, and they tend to act as if the differences between the five response options were fairly similar to one another. As a consequence, a lot of researchers treat Likert scale data as interval scale. [4] It's not interval scale, but in practice it's close enough that we usually think of it as being quasi-interval scale.

2.3 Assessing the reliability of a measurement

At this point we've thought a little bit about how to operationalize a theoretical construct and thereby create a psychological measure. And we've seen that by applying psychological measures we end up with variables, which can come in many different types. At this point, we should start discussing the obvious question: is the measurement any good? We'll do this in terms of two related ideas: reliability and validity. Put simply, the reliability of a measure tells you how precisely you are measuring something, whereas the validity of a measure tells you how accurate the measure is. 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. The measurement of my weight by means of a "bathroom scale" is very reliable. If I step on and off the scales over and over again, it'll keep giving me the same answer. Measuring my intelligence by means of "asking my mum" is very unreliable. Some days she tells me I'm a bit thick, and other days she tells me I'm a complete idiot. Notice that this concept of reliability is

different to the question of whether the measurements are correct (the correctness of a measurement relates to it's validity). If I'm holding a sack of potatoes when I step on and off the bathroom scales the measurement will still be reliable: it will always give me the same answer. However, this highly reliable answer doesn't match up to my true weight at all, therefore it's wrong. In technical terms, this is a reliable but invalid measurement. Similarly, whilst my mum's estimate of my intelligence is a bit unreliable, she might be right. Maybe I'm just not too bright, and so while her estimate of my intelligence fluctuates pretty wildly from day to day, it's basically right. That would be an unreliable but valid measure. Of course, if my mum's estimates are too unreliable it's going to be very hard to figure out which one of her many claims about my intelligence is actually the right one. To some extent, then, a very unreliable measure tends to end up being invalid for practical purposes; so much so that many people would say that reliability is necessary (but not sufficient) to ensure validity.

Okay, now that we're clear on the distinction between reliability and validity, let's have a think about the different ways in which we might measure reliability:

- -Test-retest reliability. This relates to consistency over time. If we repeat the measurement at a later date do we get a the same answer?
- -Inter-rater reliability. This relates to consistency across people. If someone else repeats the measurement (e.g., someone else rates my intelligence) will they produce the same answer?
- -Parallel forms reliability. This relates to consistency across theoretically-equivalent measurements. If I use a different set of bathroom scales to measure my weight does it give the same answer?
- -Internal consistency reliability. If a measurement is constructed from lots of different parts that perform similar functions (e.g., a personality questionnaire result is added up across several questions) do the individual parts tend to give similar answers.

Not all measurements need to possess all forms of reliability. For instance, educational assessment can be thought of as a form of measurement. One of the subjects that I teach, Computational Cognitive Science, has an assessment structure that has a research component and an exam component (plus other things). The exam component is intended to measure something different from the research component, so the assessment as a whole has low internal consistency. However, within the exam there are several questions that are intended to (approximately) measure the same things, and those tend to produce similar outcomes. So the exam on its own has a fairly high internal consistency. Which is as it should be. You should only demand reliability in those situations where you want to be measuring the same thing!

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

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

2.4 The "role" of variables: predictors and outcomes

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

When we are doing an analysis we have different names for X and Y, since they play different roles in the analysis. The classical names for these roles are independent variable (IV) and dependent variable (DV). The IV is the variable that you use to do the explaining (i.e., X) and the DV is the variable being explained (i.e., Y). The logic behind these names goes like this: if there really is a relationship between X and Y then we can say that Y depends on X, and if we have designed our study "properly" then X isn't dependent on anything else. However, I personally find those names horrible. They're hard to remember and they're highly misleading because (a) the IV is never actually "independent of everything else", and (b) if there's no relationship then the DV doesn't actually depend on the IV. And in fact, because I'm not the only person who thinks that IV and DV are just awful names, there are a number of alternatives that I find more appealing. The terms that I'll use in this book are predictors and outcomes. The idea here is that what you're trying to do is use X (the predictors) to make guesses about Y (the outcomes). [5] 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 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.

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 examine lung cancer among smokers and nonsmokers without assigning people to smoke, is an observational study. The researcher observes existing differences rather than assigning the predictor variable. We can still use statistics to analyze 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 analyze the results of case studies and it's usually very hard to draw any general conclusions about "people in general" from a few isolated examples. However, case studies are very useful in some situations. Firstly, there are situations where you don't have any alternative. Neuropsychology has this issue a lot. Sometimes, you just can't find a lot of people with brain damage in a specific brain area, so the only thing you can do is describe those cases that you do have in as much detail and with as much care as you can. However, there's also some genuine advantages to case studies. Because you don't have as many people to study you have the ability to invest lots of time and effort trying to understand the specific factors at play in each case.

This is a very valuable thing to do. As a consequence, case studies can complement the more statistically-oriented approaches that you see in experimental and quasi-experimental designs. We won't talk much about case studies in this book, but they are nevertheless very valuable tools. Note that case studies are sometimes discussed in the same sentence with anecdotal evidence, but
in general these are two very different things. Whereas case studies are a systematic research method designed to generate generalizable knowledge, anecdotal evidence is usually the result of an informal, personal account or single example. For example, a case study might involve a year-long investigation into how one school implemented a new mathematics learning program, using interviews, test scores, and classroom observations as multiple sources of data to support a generalizable conclusion. On the other hand, anecdotal evidence might involve a teacher saying "I tried that program with my students last year and it didn't work". Certainly, the teacher has some data to support their claim.

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 concerns how findings apply to real-world situations and is related to external validity.

2.6.1 Internal validity

Internal validity refers to the extent to which you are able draw the correct conclusions about the causal relationships between variables. It's called "internal" because it refers to the relationships between things "inside" the study. Let's illustrate the concept with a simple example. Suppose you're interested in finding out whether a university education makes you write better. To do so, you get a group of first year students, ask them to write a 1000 word essay, and count the number of spelling and grammatical errors they make. Then you find some third-year students, who obviously have had more of a university education than the first-years, and repeat the exercise. And let's suppose it turns
out that the third-year students produce fewer errors. And so you conclude that a university education improves writing skills. Right? Except that the big problem with this experiment is that the third-year students are older and they've had more experience with writing things. So it's hard to know for sure what the causal relationship is. Do older people write better? Or people who have had more writing experience? Or people who have had more education? Which of the above is the true cause of the superior performance of the third-years? Age? Experience? Education? You can't tell. This is an example of a failure of internal validity, because your study doesn't properly tease apart the causal relationships between the different variables.

2.6.2 External validity

External validity relates to the generalizability or applicability of your findings. That is, to what extent do you expect to see the same pattern of results in "real life" as you saw in your study. To put it a bit more precisely, any study that you do in psychology will involve a fairly specific set of questions or tasks, will occur in a specific environment, and will involve participants that are drawn from a particular subgroup (disappointingly often it is college students!). So, if it turns out that the results don't actually generalize or apply to people and situations beyond the ones that you studied, then what you've got is a lack of external validity.

The classic example of this issue is the fact that a very large proportion of studies in psychology will use undergraduate psychology students as the participants. Obviously, however, the researchers don't care only about psychology students. They care about people in general. Given that, a study that uses only psychology students as participants always carries a risk of lacking external validity. That is, if there's something "special" about psychology students that makes them different to the general

population in some relevant respect, then we may start worrying about a lack of external validity.

That said, it is absolutely critical to realize that a study that uses only psychology students does not necessarily have a problem with external validity. I'll talk about this again later, but it's such a common mistake that I'm going to mention it here. The external validity of a study is threatened by the choice of population if (a) the population from which you sample your participants is very narrow (e.g., psychology students), and (b) the narrow population that you sampled from is systematically different from the general population in some respect that is relevant to the psychological phenomenon that you intend to study. The italicized part is the bit that lots of people forget. It is true that psychology undergraduates differ from the general population in lots of ways, and so a study that uses only psychology students may have problems with external validity. However, if those differences aren't very relevant to the phenomenon that you're studying, then there's nothing to worry about. To make this a bit more concrete here are two extreme examples:

- You want to measure "attitudes of the general public towards psychotherapy", but all of your participants are psychology students. This study would almost certainly have a problem with external validity.
- You want to measure the effectiveness of a visual illusion, and your participants are all psychology students. This study is 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:

- People might answer a "psychology questionnaire" in a manner that doesn't reflect what they would do in real life.
- Your lab experiment on (say) "human learning" has a different structure to the learning problems people face in real life.

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

It is worth noting that the idea of construct validity has had some recent criticism. For example, Borsboom and colleagues [6] argue that construct validity is no longer a tenable concept, because it can fluctuate over time (e.g., the case of the phlogiston concept). Instead, Borsboom and colleagues recommend the term causal validity, defined as follows: A measure of an attribute is valid if and only if (a) the attribute exists, and (b) manipulations of the attribute lead to corresponding changes in the measure.

2.6.4 Face validity

Face validity simply refers to whether or not a measure "looks like" it's doing what it's supposed to, nothing more. If I design a test of intelligence, and people look at it and they say "no, that test doesn't measure intelligence", then the measure lacks face validity. It's as simple as that. Obviously,

face validity isn't very important from a pure scientific perspective. After all, what we care about is whether or not the measure actually does what it's supposed to do, not whether it looks like it does what it's supposed to do. As a consequence, we generally don't care very much about face validity. That said, the concept of face validity serves three useful pragmatic purposes:

- -Sometimes, an experienced scientist will have a "hunch" that a particular measure won't work. While these sorts of hunches have no strict evidentiary value, it's often worth paying attention to them. Because often times people have knowledge that they can't quite verbalize, so there might be something to worry about even if you can't quite say why. In other words, when someone you trust criticizes the face validity of your study, it's worth taking the time to think more carefully about your design to see if you can think of reasons why it might go awry. Mind you, if you don't find any reason for concern, then you should probably not worry. After all, face validity really doesn't matter very much.

- -Often (very often), completely uninformed people will also have a "hunch" that your research is crap. And they'll criticize it on the internet or something. On close inspection you may notice that these criticisms are actually focused entirely on how the study "looks", but not on anything deeper. The concept of face validity is useful for gently explaining to people that they need to substantiate their arguments further.
- -Expanding on the last point, if the beliefs of untrained people are critical (e.g., this is often the case for applied research where you actually want to convince policy makers of something or other) then you have to care about face validity. Simply because, whether you like it or not, a lot of people will use face validity as a proxy for real validity. If you want the government to change a law on scientific psychological grounds, then it won't matter how good your studies "really" are. If they lack face validity you'll find that politicians ignore you. Of course, it's somewhat unfair that policy often depends more on appearance than fact, but that's how things go.

2.6.5 Ecological validity

Ecological validity is a different notion of validity, which is similar to external validity, but less important. The idea is that, in order to be ecologically valid, the entire set up of the study should closely approximate the real world scenario that is being investigated. Ecological validity is related to external validity. It relates mostly to whether the study "looks" right, but with a bit more rigor to it. To be ecologically valid the study has to look right in a fairly specific way. The idea behind it is the intuition that a study that is ecologically valid is more likely to be externally valid. It's no guarantee, of course. But the nice thing about ecological validity is that it's much easier to check whether a study is ecologically valid than it is to check whether a study is externally valid. A simple example would be eyewitness identification studies. Most of these studies tend to be done in a university setting, often with a fairly simple array of faces to look at, rather than a line up. The length of time between seeing the "criminal" and being asked to identify the suspect in the "line up" is usually shorter. The "crime" isn't real so there's no chance of the witness being scared, and there are
no police officers present so there's not as much chance of feeling pressured. These things all mean that the study definitely lacks ecological validity. They might (but might not) mean that it also lacks external validity.

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 [7] 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 generalize 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 not necessarily evenly, between different groups.

However, there are always swings and roundabouts and when we start thinking about artefacts rather than confounders the shoe is very firmly on the other foot. For the most part, artefactual results tend to be more of a concern for experimental studies than for non-experimental studies. To see this, it helps to realize that the reason that a lot of studies are non-experimental is precisely because what the researcher is trying to do is examine human behaviour in a more naturalistic context. By working in a more real-world context you lose experimental control (making yourself vulnerable to confounders), but because you tend to be studying human psychology "in the wild" you reduce the chances of getting an artefactual result. Or, to put it another way, when you take psychology out of the wild and bring it into the lab (which we usually have to do to gain our experimental control), you always run the risk of accidentally studying something different to what you wanted to study.

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 characterized as a confounder or an artefact, they're pretty vague concepts. So let's have a look at some of the most common examples.

2.7.1 History effects

History effects refer to the possibility that specific events may occur during the study that might influence the outcome measure. For instance, something might happen in between a pre-test and a post-test. Or in-between testing participant 23 and participant 24. Alternatively, it might be that you're looking at a paper from an older study that was perfectly valid for its time, but the world has changed enough since then that the conclusions are no longer trustworthy. Examples of things that would count as history effects are:

- -You're interested in how people think about risk and uncertainty. You started your data collection in December 2010. But finding participants and collecting data takes time, so you're still finding new people in February 2011. Unfortunately for you (and even more unfortunately for others), the Queensland floods occurred in January 2011 causing billions of dollars of damage and killing many people. Not surprisingly, the people tested in February 2011 express quite different beliefs about handling risk than the people tested in December 2010. Which (if any) of these reflects the "true" beliefs of participants? I think the answer is probably both. The Queensland floods genuinely changed the beliefs of the Australian public, though possibly only temporarily. The key thing here is that the "history" of the people tested in February is quite different to people tested in December.
- -You're testing the psychological effects of a new anti-anxiety drug. So what you do is measure anxiety before administering the drug (e.g., by self-report, and taking physiological measures). Then you administer the drug, and afterwards you take the same measures. In the middle however, because your lab is in Los Angeles, there's an earthquake which increases the anxiety of the participants.

2.7.2 Maturation effects

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

- -When doing developmental psychology research you need to be aware that children grow up quite rapidly. So, suppose that you want to find out whether some educational trick helps with vocabulary size among 3 year olds. One thing that you need to be aware of is that the vocabulary size of children that age is growing at an incredible rate (multiple words per day) all on its own. If you design your study without taking this maturational effect into account, then you won't be able to tell if your educational trick works.
- -When running a very long experiment in the lab (say, something that goes for 3 hours) it's very likely that people will begin to get bored and tired, and that this maturational effect will cause performance to decline regardless of anything else going on in the experiment

2.7.3 Repeated testing effects

An important type of history effect is the effect of repeated testing. Suppose I want to take two measurements of some psychological construct (e.g., anxiety). One thing I might be worried about is if the first measurement has an effect on the second measurement. In other words, this is a history effect in which the "event" that influences the second measurement is the first measurement itself! This is not at all uncommon. Examples of this include:

- -Learning and practice: e.g., "intelligence" at time 2 might appear to go up relative to time 1 because participants learned the general rules of how to solve "intelligence-test-style" questions during the first testing session.
- -Familiarity with the testing situation: e.g., if people are nervous at time 1, this might make performance go down. But after sitting through the first testing situation they might calm down a lot precisely because they've seen what the testing looks like.
- -Auxiliary changes caused by testing: e.g., if a questionnaire assessing mood is boring then mood rating at measurement time 2 is more likely to be "bored" precisely because of the boring measurement made at time 1.

2.7.4 Selection bias

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

If any of those characteristics turns out to be relevant (say, your treatment works better on females than males) then you're in a lot of trouble.

2.7.5 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 generalizability 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. Suppose that, for the first time ever in the history of psychology, I manage to find the perfectly balanced and representative sample of people. I start running "Dani's incredibly long and tedious experiment" on my perfect sample but then, because my study is incredibly long and tedious, lots of people start dropping out. I can't stop this. Participants absolutely have the right to stop doing any experiment, any time, for whatever reason they feel like, and as researchers we are morally (and professionally) obliged to remind people that they do have this right. So, suppose that "Dani's incredibly long and tedious experiment" has a very high drop out rate. What do you suppose the odds are that this drop out is random? Answer: zero. Almost certainly the people who remain are more conscientious, more tolerant of boredom, etc., than those that leave. To the extent that (say) conscientiousness is relevant to the psychological phenomenon that I care about, this attrition can decrease the validity of my results.

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

2.7.6 Non-response bias

Non-response bias is closely related to selection bias and to differential attrition. The simplest version of the problem goes like this. You mail out a survey to 1000 people 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 generalize from those 300 people who replied to the population at large, since you now have a very non-random sample. The issue of non-response bias is more general than this, though. Among the (say) 300 people that did respond to the survey, you might find that not everyone answers every question. If (say) 80 people chose not to answer one of your questions, does this introduce problems? As always, the answer is maybe. If the question that wasn't answered was on the last page of the questionnaire, and those 80 surveys were returned with the last page missing, there's a good chance that the missing data isn't a big deal; probably the pages just fell off. However, if the question that 80 people didn't answer was the most confrontational or invasive personal question in the questionnaire, then almost certainly you've got a problem. In essence, what you're dealing with here is what's called the problem of missing data. If the data that is missing was "lost" randomly, then it's not a big problem. If it's

missing systematically, then it can be a big problem.

2.7.7 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.

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

Regression to the mean is surprisingly common. For instance, if two very tall people have kids their children will tend to be taller than average but not as tall as the parents. The reverse happens with very short parents. Two very short parents will tend to have short children, but nevertheless those kids will tend to be taller than the parents. It can also be extremely subtle. For instance, there have been studies done that suggested that people learn better from negative feedback than from positive feedback. However, the way that people tried to show this was to give people positive reinforcement whenever they did good, and negative reinforcement when they did bad. And what you see is that after the positive reinforcement people tended to do worse, but after the negative reinforcement they tended to do better. But notice that there's a selection bias here! When people do very well, you're selecting for "high" values, and so you should expect, because of regression to the mean, that performance on the next trial should be worse regardless of whether reinforcement is given. Similarly, after a bad trial, people will tend to improve all on their own. The apparent superiority of negative feedback is an artefact caused by regression to the mean (see Kahneman and Tversky 1973, for discussion).

2.7.8 Experimenter bias

Experimenter bias can come in multiple forms. The basic idea is that the experimenter, despite the best of intentions, can accidentally end up influencing the results of the experiment by subtly communicating the "right answer" or the "desired behaviour" to the participants. Typically, this occurs because the experimenter has special knowledge that the participant does not, 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.

The general solution to the problem of experimenter bias is to engage in double blind studies, where neither the experimenter nor the participant knows which condition the participant is in or knows what the desired behaviour is. This provides a very good solution to the problem, but it's important to recognize that it's not quite ideal, and hard to pull off perfectly. For instance, the obvious way that I could try to construct a double blind study is to have one of my Ph.D. students (one who doesn't know anything about the experiment) run the study. That feels like it should be enough. The only person (me) who knows all the details (e.g., correct answers to the questions, assignments of participants to conditions) has no interaction with the participants, and the person who does all the talking to
people (the Ph.D. student) doesn't know anything. Except 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 Ph.D. student also knows me and knows a bit about my general beliefs about people and psychology (e.g., I tend to think humans are much smarter than psychologists give them credit for). As a result of all this, it's almost impossible for the experimenter to avoid knowing a little bit about what expectations I have. And even a little bit of knowledge can have an effect. Suppose the experimenter accidentally conveys the fact that the participants are expected to do well in this task. Well, there's a thing called the "Pygmalion effect", 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.

2.7.9 Demand effects and reactivity

When talking about experimenter bias, the worry is that the experimenter's knowledge or desires for the experiment are communicated to the participants, and that these 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 psychological 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 think like a social psychologist and look at some of the roles that people might adopt during an experiment but might not adopt if the corresponding events were occurring in the real world:

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

2.7.10 Placebo effects

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

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. Of interest to psychology is that the strongest evidence for at least some placebo effect is in self-reported outcomes, most notably in treatment of pain (Hróbjartsson and Gøtzsche 2010).

2.7.11 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 generalize to a wider array of people, places and measures.

2.7.12 Fraud, deception and self-deception

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

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. [8] Not only that, as I mentioned earlier, scientists are not immune to belief bias. It's easy for a researcher to end up deceiving themselves into believing the wrong thing, and this can lead them to conduct subtly flawed research and then hide those flaws when they write it up. So you need to consider not only the (probably unlikely) possibility of outright fraud, but also the (probably quite common) possibility that the research is unintentionally "slanted".

I opened a few standard textbooks and didn't find much of a discussion of this problem, so here's my own attempt to list a few ways in which these issues can arise:

- Data fabrication. Sometimes, people just make up the data. This is occasionally done with "good" intentions. For instance, the researcher believes that the fabricated data do reflect the truth, and may actually reflect "slightly cleaned up" versions of actual data. On other occasions, the fraud is deliberate and malicious. Some high-profile examples where data fabrication has been alleged or shown include Cyril Burt (a psychologist who is thought to have fabricated some of his data), Andrew Wakefield (who has been accused of fabricating his data connecting the MMR vaccine to autism) and Hwang Woo-suk (who falsified a lot of his data on stem cell research).
- Hoaxes. Hoaxes share a lot of similarities with data fabrication, but they differ in the intended purpose. A hoax is often a joke, and many of them are intended to be (eventually) discovered. Often, the point of a hoax is to discredit someone or some field. There's quite a few well known scientific hoaxes that have occurred over the years (e.g., Piltdown man) 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 analyzed, 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 [9] & 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. [10] This is an example of a publication bias. Since no-one ever published the 19 studies that didn't find an effect, a naive reader would never know that they existed. Worse yet, most researchers "internalise" this bias and end up self-censoring their research. Knowing that negative results aren't going to be accepted for publication, they never even try to report them. As a friend of mine says "for every experiment that you get published, you also have 10 failures". And she's right. The catch is, while some (maybe most) of those studies are failures for boring reasons (e.g. you stuffed something up) others might be genuine "null" results that you ought to acknowledge when you write up the "good" experiment. And telling which is which is often hard to do. A good place to start is a paper by Ioannidis (2005) with the depressing title "Why most published research findings are false". I'd also suggest taking a look at work by Kühberger, Fritz, and Scherndl (2014) presenting statistical evidence that this actually happens in psychology.

There's probably a lot more issues like this to think about, but that'll do to start with. What I really want to point out is the blindingly obvious truth that real world science is conducted by actual humans, and only the most gullible of people automatically assumes that everyone else is honest and impartial. Actual scientists aren't usually that naive, but for some reason the world likes to pretend that we are, and the textbooks we usually write seem to reinforce that stereotype.

2.8 Summary

This chapter isn't really meant to provide a comprehensive discussion of psychological research methods. It would require another volume just as long as this one to do justice to the topic. However, in real life statistics and study design are 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 psychological 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.

---

1. [1] Presidential Address to the First Indian Statistical Congress, 1938. Source: http://en.wikiquote.org/wiki/Ro nald_Fisher
2. [2] Well... now this is awkward, isn't it? This section is one of the oldest parts of the book, and it's outdated in a rather embarrassing way. I wrote this in 2010, at which point all of those facts were true. Revisiting this in 2018, well I'm not 33 any more, but that's not surprising I suppose. I can't imagine my chromosomes have changed, so I'm going to guess my karyotype was then and is now XY. The self-identified gender, on the other hand...ah. I suppose the fact that the title page now refers to me as Danielle rather than Daniel might possibly be a giveaway, but I don't typically identify as "male" on a gender questionnaire these days, and I prefer "she/her" pronouns as a default (it's a long story)! I did think a little about how I was going to handle this in the book, actually. The book has a somewhat distinct authorial voice to it, and I feel like it would be a rather different work if I went back and wrote everything as Danielle and updated all the pronouns in the work. Besides, it would be a lot of work, so I've left my name as "Dan" throughout the book, and in any case "Dan" is a perfectly good nickname for "Danielle", don't you think? In any case, it's not a big deal. I only wanted to mention it to make life a little easier for readers who aren't sure how to refer to me. I still don't like anchovies though :-)
3. [3] 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° 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.

4. [4] Ah, psychology ... never an easy answer to anything!
5. [5] 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". Sigh. This sort of terminological confusion is very common, I'm afraid.
6. [6] see Borsboom, D., Cramer, A. O. J., Kievit, R. A., Zand Scholten, A., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), The concept of validity: Revisions, new directions, and applications (pp. 135-170). Charlotte, NC: Information Age Publishing and Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. Psychological Review, 111(4), 1061-1071.
7. [7] 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.
8. [8] 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.

9. [9] Since the original writing of this book, the term "data mining" has become more popular in a lot of different contexts, particularly related to machine learning, artificial intelligence, etc. To be clear, the discussion in this section is objecting to the notion of data mining as it relates to post hoc searches for patterns without a strong theoretical basis for prediction.
[10] Clearly, the real effect is that only insane people would even try to read Finnegans Wake.

This page titled 2.1: A brief introduction to research design is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 2: A brief introduction to research design by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

3: GETTING STARTED WITH JASP

3.1: Getting started with JASP
3.2: Installing and Opening JASP
3.3: Accessibility and Other Settings
3.4: Creating Data Files in JASP
3.5: Importing Data from Excel into JASP

This page titled 3: Getting Started with JASP is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

3.1: Getting started with JASP

> Robots are nice to work with. - Roger Zelazny [1]

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.
- 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.

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:

```
01 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 [2], the current version of JASP is 0.19.3, but they usually issue updates every few months, so you'll probably have a newer version. [3]
3.1.1 Starting up JASP

One way or another, regardless of what operating system you're using, 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.

Users of earlier versions of JASP were often a bit dismayed to see that there was no place to begin entering data. This was a deliberate decision on the part of the JASP developers; their philosophy was that users should be allowed to use the editor they are most comfortable with [4]. However, starting with version 0.18.0, JASP now includes a data editor. Even so, the preferred method for getting data

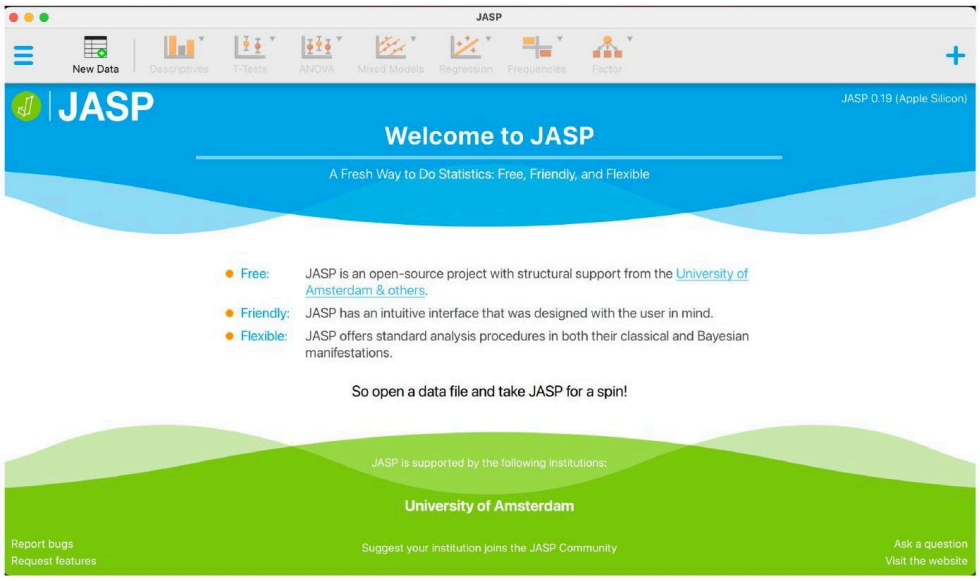


Figure 3.1: JASP looks like this when you start it.

into JASP is still to load a CSV file (.csv), which is a text-based data format that can be created by (and opened in) any spreadsheet program. We'll talk more about this a bit later.

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

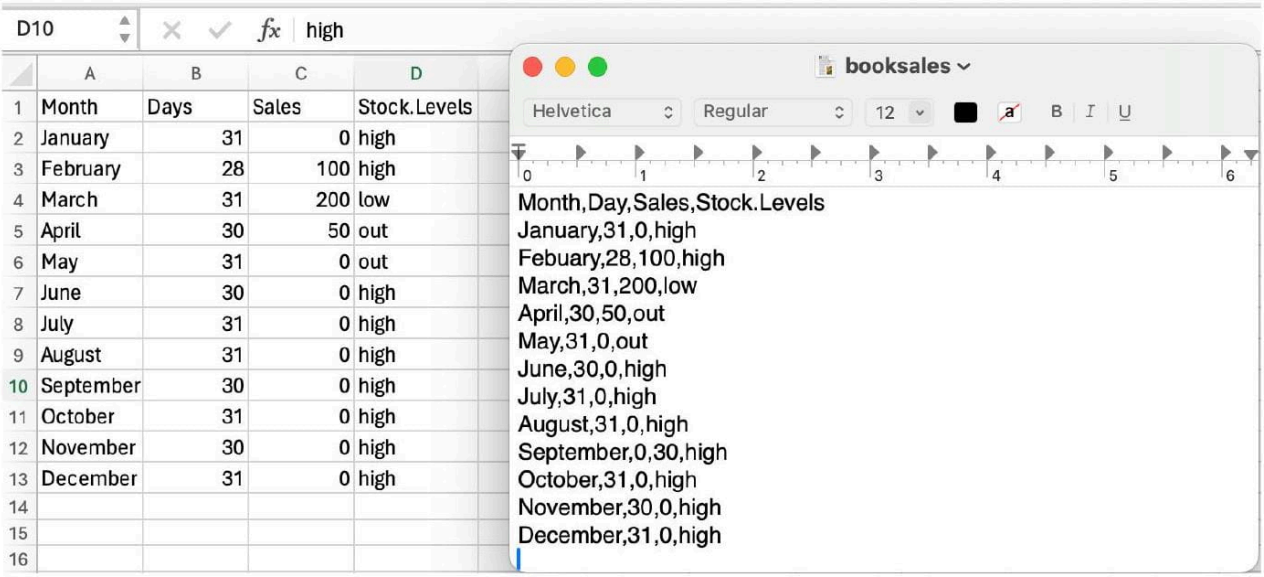
One quite commonly used data format is the humble "comma separated value" file, also called a CSV file, and usually bearing the file extension .csv. CSV files are just plain old-fashioned text files and what they store is basically just a table of data. This is illustrated in Figure 3.2, which shows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



	A	B	C	D
1	Month	Days	Sales	Stock.Levels
2	January	31	0	high
3	February	28	100	high
4	March	31	200	low
5	April	30	50	out
6	May	31	0	out
7	June	30	0	high
8	July	31	0	high
9	August	31	0	high
10	September	30	0	high
11	October	31	0	high
12	November	30	0	high
13	December	31	0	high
14				
15				
16				

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.

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.

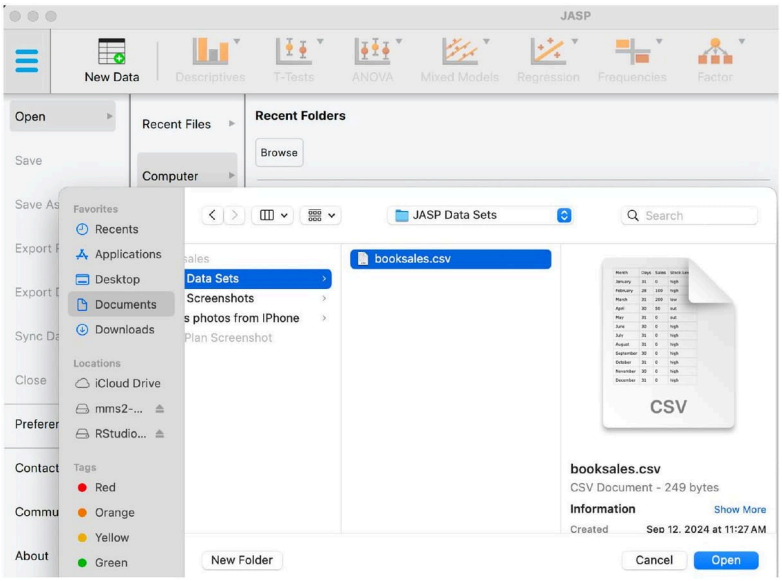


Figure 3.3: A dialog box on a Mac asking you to select the CSV file JASP should try to import. Mac users will recognise this immediately - it's the usual way in which a Mac asks you to find a file. Windows users won't see this, but instead will see the usual explorer window that Windows always gives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

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 LaTeX document preparation system, you can select "Copy special" and "LaTeX code"; doing so will place the LATEXsyntax into your clipboard.

3.5 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.

3.6 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.

3.7 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!

---

1. [1] Source: Dismal Light (1968).

2. [2] Version 2.0, May 21, 2025
   [3] 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.
   [4] See https://jasp-stats.org/2018/05/15/data-editing-in-jasp/ for a discussion of this very issue.

This page titled 3.1: Getting started with JASP is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 3: Getting started with JASP by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

3.2: Installing and Opening JASP

If your institution has approved JASP for use, you can typically download it from a dedicated Software Centre on managed devices such as university desktop computers. If you would like to install JASP on your own computer, you can download it for free from jasp-stats.org by clicking on the 'Download JASP' button.

Note: Make sure to read the installation guide for your device's operating system before installing the software. Especially important for Mac devices which have stricter requirements for running JASP.

Once downloaded you can start the program via the 'Start' menu in Windows, the 'Finder' icon on a Mac, or the 'App Launcher' in Chromebook.

Below is a screenshot of the JASP start-up screen which should be the first thing you see after opening the program.

screenshot of the Welcome to JASP page

To **open an existing data file** you need to click on the collapsed menu icon.

the 'burger' or 'collapsed menu' option.

To **start a new data file** click on the New Data icon.

New Data icon

This page titled 3.2: Installing and Opening JASP is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss & Stanislava Naneva via source content that was edited to the style and standards of the LibreTexts platform.

3.3: Accessibility and Other Settings

JASP offers a number of **adjustable settings that support accessibility**. Different people have different needs and preferences so we recommend that you try out different options. Settings can be changed at any time so don't be afraid to try out different options! We have made some general recommendations below.

You can find more information about JASP's accessibility at jasp-stats.org/accessibility

Interface Settings

1. Click on the collapsed menu icon. A side menu should appear on the left.

2. Select **Preferences > Interface** where you will find a number of options. Here are some recommendations:

    a. Font (Interface, Results, and Help) – we strongly recommend using **APA-approved sans serif fonts** such as Calibri, Arial, or Lucida Sans Unicode. The default is Arial.

    b. Themes – try both Light and Dark theme to see which is better for your eyes.

    c. Zoom (%) – we recommend 130% but optimal settings may vary depending on the size and resolution of your display.

image

Results Settings

1. Click on the collapsed menu icon. A side menu should appear on the left.

2. Select **Preferences > Results** where you will find a number of options. Here are some recommendations:

    a. Table options – select only 'Fix the number of decimals'. We recommend 3 decimal places (dp) as you will usually be rounding down to 2dp but some values (e.g., *p* values) need to be reported to 3dp.

    b. Plot options – leave as default options.

    c. Miscellaneous options – if you are interested in coding, enabling the R syntax will give you more information about how JASP operations run. [1]

image

---

1. R is an open-source program for data analysis and visualisation that is widely used in scientific disciplines. You can find out more at r-project.org. 1.02: Accessibility and Other Settings#return-footnote-217-1

This page titled 3.3: Accessibility and Other Settings is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss & Stanislava Naneva via source content that was edited to the style and standards of the LibreTexts platform.

3.4: Creating Data Files in JASP

Opening an Existing JASP Data File

1. Click on the collapsed menu icon. A side menu should appear on the left.

2. Select **Open > Computer > Browse** then find the folder where the JASP data file is saved.

3. Click **Open**.

Alternatively, you can directly open a JASP data file from the computer folder in which it is saved.

image

Creating a New JASP Data File

If you select the 'New Data' button, you will notice that some of the menu items in the toolbar at the top will change. You will learn about these later.

Each **row** contains information for one participant. So, if you have multiple pieces of data about one participant (e.g., age, gender, and height) they will need to be inserted in the same row.

Each **column** represents one variable (e.g., age).

You will need to enter the relevant information for each of your variables (i.e., **column**) per each participant (i.e., **row**).

How do I do that?

Double-click on the header of the column you wish to edit (e.g., Column 1). This will open a new section above your data which contains the information for that variable.

image

When you are **setting up a new variable**, the most important information you need to enter is:

image

1. Name: Enter the name of your variable (e.g., Age). Try to keep this name relatively short.

2. Long Name: If your variable is slightly more complex (e.g., Age of participants at the start of the study) then you can provide that information in this box.

3. Description: This box allows you to give even more information about the variable. This can be useful for data sets with a large number of variables.

4. Column type: Select the type of variable. This will affect the analysis that JASP will allow you to conduct, as some analyses are specific to the variable type.

- **Scale** (for continuous variables; e.g., age)
- **Nominal** (for categorical variables; e.g., gender)
- **Ordinal**(for variables that use a scale; e.g., Likert scale)

5. Computed type: This allows you to calculate new variables (e.g., a sum of other variables).

Entering Scale Variables

How do I do that?

1. Enter the name of the variable and select ‘Scale’ as the Column Type.

2. Enter your data by clicking into the empty cell for each row/column.

Entering Categorical (Nominal) Variables – Assigning Labels to Groups

For categorical data, you will need to specify which values in your dataset relate to which group or category (e.g., for *Gender*, 1 = male and 2 = female). You can have as many categories as you need. For example, for *Gender* you may also have 3 = non-binary and 4 = other.

How do I do that?

1. Enter the variable name (e.g., *Gender*) and select ‘Nominal’ as the column type.

2. Enter the numerical values for the category or group that each participant belongs to in the variable column. In the example, we have 1 = male and 2 = female (see **A**). This information will now appear under the ‘Label editor’.

A

image

3. Once you have entered the data for your participants go to the ‘Label editor’ and double-click on the ‘Label’ next to the ‘Value’ to change the number to a more meaningful name for your category or group (see **B**). This change will be reflected in the variable column where you should now be able to see the label.

B
image

Note: JASP will let you use ‘string’ variables (i.e., words). However, you need to enter the words in exactly the same way otherwise JASP will think you have additional categories. If you choose to enter the data in this way, check the ‘Label editor’ to make sure you have the correct number of categories.

Entering Ranked (Ordinal) Variables

In the same way that you assign numerical values to categorical variables, you will also have numerical values for ranked variables (e.g., liking ratings on a questionnaire). Usually, ordinal variables will have more than two values.

image

How do I do that?

1. Enter the variable name (e.g., *Liking*) and enter ‘Ordinal’ as the column type.

2. Enter the numerical value for each of your values (e.g., 1 = strongly disagree, 2 = disagree, etc.).

3. Once you have entered these values for your participants go to the ‘Label editor’ and double-click on the ‘Label’ next to the ‘Value’ to change the number to the label for each of the ranked variables. This change will be reflected in the variable column where you should now be able to see the label.

4. It might be the case that when you are entering your data the values are not in numerical order. In the ‘Label editor’ you can use the arrows on the right-hand side of the box to move the labels so that they are in a logical order (e.g., 1, 2, 3, 4, etc.).

image

Saving JASP Data Files

The best place to save your work when you are working from a university computer is a dedicated drive, folder, or cloud storage which you can access from any managed university device. We strongly recommend saving your work periodically to avoid losing progress.

How do I do that?

1. Click on the collapsed menu icon. A side menu should appear on the left. Select ‘Save As’.

2. Once provided with the option, click on the little arrow at the top of the Save Menu. A drop-down menu will appear listing all the possible places on your computer where you could save your file.

image

3. In the list of drives find the one which has (U:) at the end or you know is a safe place to store your work.

4. Click on your selected drive. This should then bring up the folders and files currently listed in that drive.

5. Give your file an appropriate file name and click **Save**.

Note: If you’re not working on a university computer, we recommend that you create a dedicated folder on your own device to save your JASP files. There is nothing worse than losing your data file or analysis!

This page titled 3.4: Creating Data Files in JASP is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss & Stanislava Naneva via source content that was edited to the style and standards of the LibreTexts platform.

3.5: Importing Data from Excel into JASP

JASP can open Excel spreadsheets directly. You can also save a spreadsheet as a comma-separated values (.csv) file before opening it in JASP.

How do I do that?

1. Open the Excel spreadsheet that contains your data.

2. Once in the spreadsheet, click on **File**.

image

3. Select **Save As** from the left-hand-side menu, choose where you wish to save your file, and then click on the drop-down menu that lists different file formats. You need to select **CSV (Comma delimited) (*.csv)**. Note that there are other CSV options but you need to ignore them.

image

4. Finally, click **Save** and your .csv file should now be ready to import into JASP. image

5. Open JASP and select **Open** from the left-hand-side menu. Click **Browse** and locate the folder where you have saved your .csv file.

image

6. Select the .csv file you wish to import and click **Open**. Your data should now be available within JASP.

image

This page titled 3.5: Importing Data from Excel into JASP is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss & Stanislava Naneva via source content that was edited to the style and standards of the LibreTexts platform.

4: DESCRIPTIVE STATISTICS

4.1: Descriptive statistics
4.2: Descriptive Statistics in JASP

This page titled 4: Descriptive Statistics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

4.1: 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 aflsmall_margins file and see what variables are stored in the file.

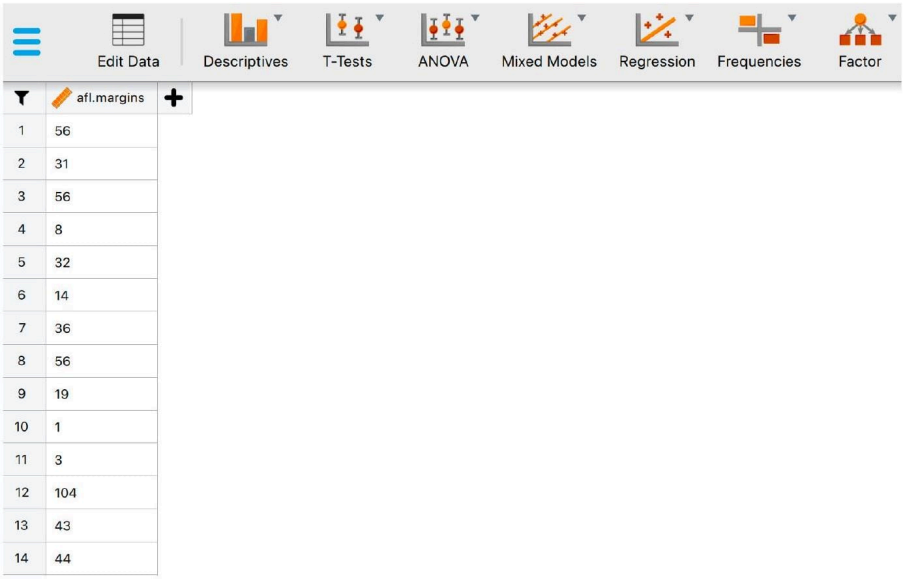


Figure 4.1: A screenshot of JASP showing the variables stored in the aflsmall_margins.csv file

In fact, there is just one variable here, afl.margins. We'll focus a bit on this variable in this chapter, so I'd better tell you what it is. Unlike most of the data sets in this book, this is actually real data, relating to the Australian Football League (AFL). [1] The afl.margins variable contains the winning margin (number of points) for all 176 home and away games played during the 2010 season.

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 5). 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 afl.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 5.1. For now, it's enough to look at the histogram and note that it provides a fairly interpretable representation of the afl.margins data.

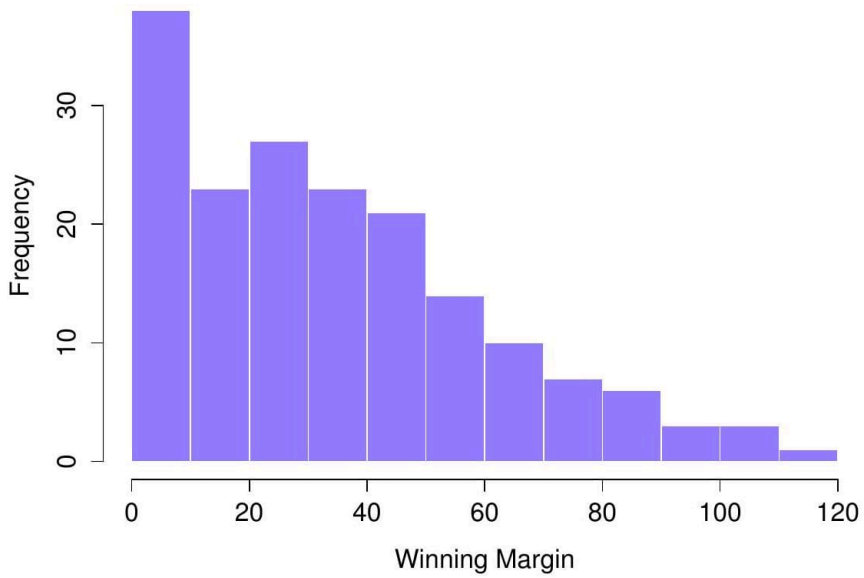


Figure 4.2: A histogram of the AFL 2010 winning margin data (the afl.margins varable). As you might expect, the larger the winning margin the less frequently you tend to see it.

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 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 AFL winning margins 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 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 afl.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
winning margin, game 1	X_1	56 points
winning margin, game 2	X_2	31 points

the observation	its symbol	the observed value
winning margin, game 3	X_3	56 points
winning margin, game 4	X_4	8 points
winning margin, game 5	X_5	32 points

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:

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

This formula is entirely correct but it's terribly long, so we make use of the summation symbol Σ to shorten it. [a] 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, Σ 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".

[a] The choice to use Σ 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 Π 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.

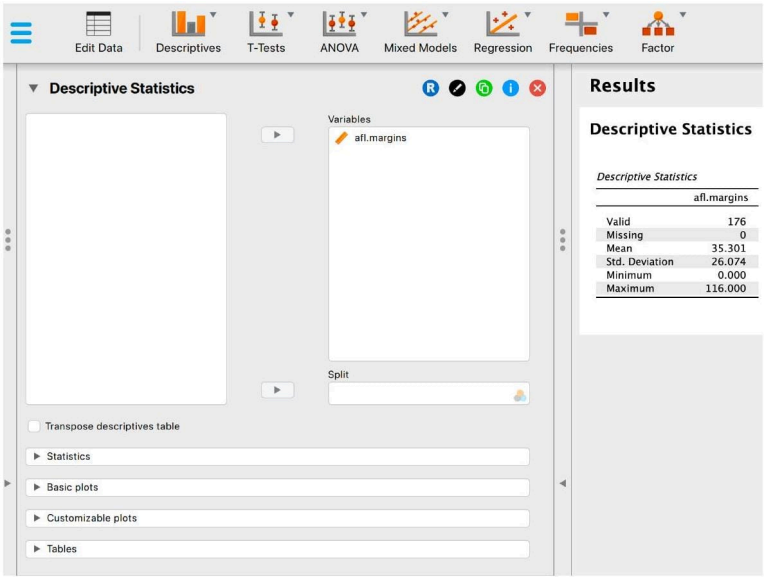


Figure 4.3: Default descriptives for the AFL 2010 winning margin data (the afl.margins variable).

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 $(\mathrm{N} = 176)$, 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.

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, suppose Bob (income $50,000), Kate (income $60,000) and Jane (income $65,000) are sitting at a table. The average income at the table is $58,333 and the median income is $60,000. Then Bill sits down with them (income $100,000,000). The average income has now jumped to $25,043,750 but the median rises only to $62,500. If you're interested in looking at the overall income at the table the mean might be the right answer. But if you're interested in what counts as a typical income at the table the median would be a better choice here.

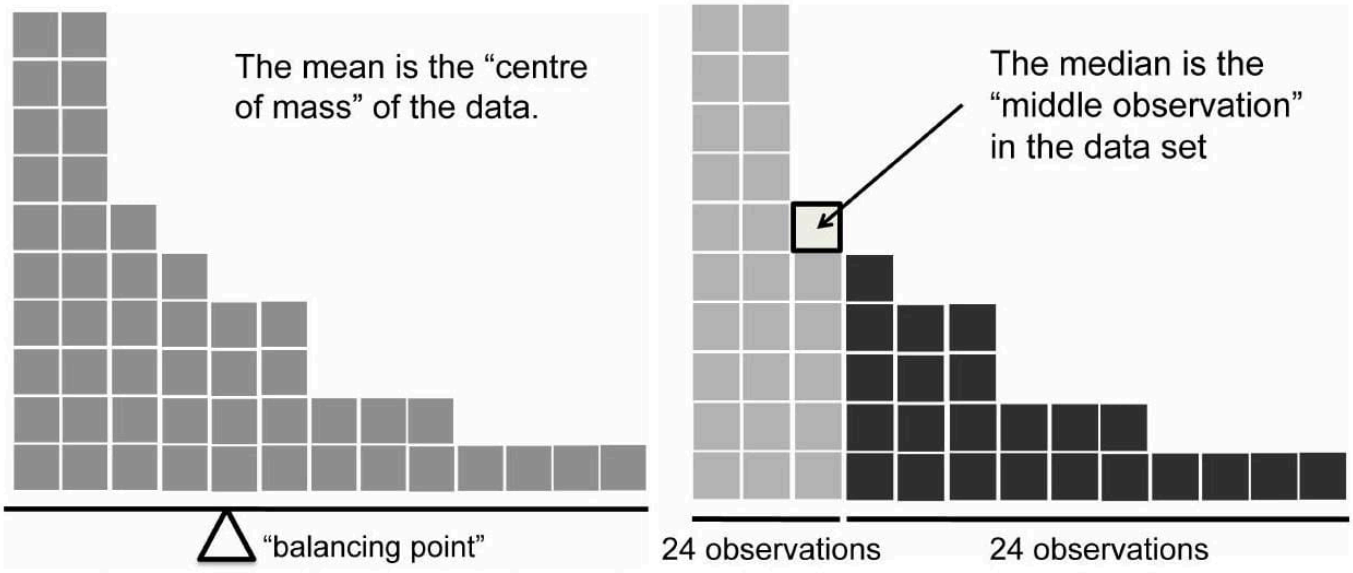


Figure 4.4: An illustration of the difference between how the mean and the median should be interpreted. The mean is basically the "centre of gravity" of the data set. If you imagine that the histogram of the data is a solid object, then the point on which you could balance it (as if on a see-saw) is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

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. Since I tend to mock journalists for their poor scientific and statistical knowledge, I should give credit where credit is due. This is an excellent article on the ABC news website [2] from 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

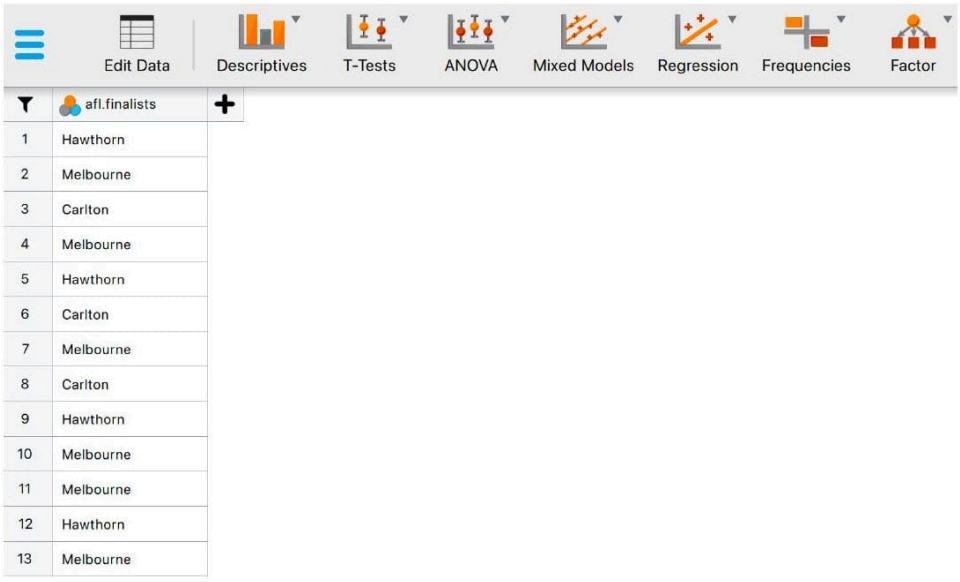
That's, um, a rather serious discrepancy. One group of people say 9, another says 4-5. Should we just split the difference and say the truth lies somewhere in between? Absolutely not! This is a situation where there is a right answer and a wrong answer. Demographia is correct, and the Commonwealth Bank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

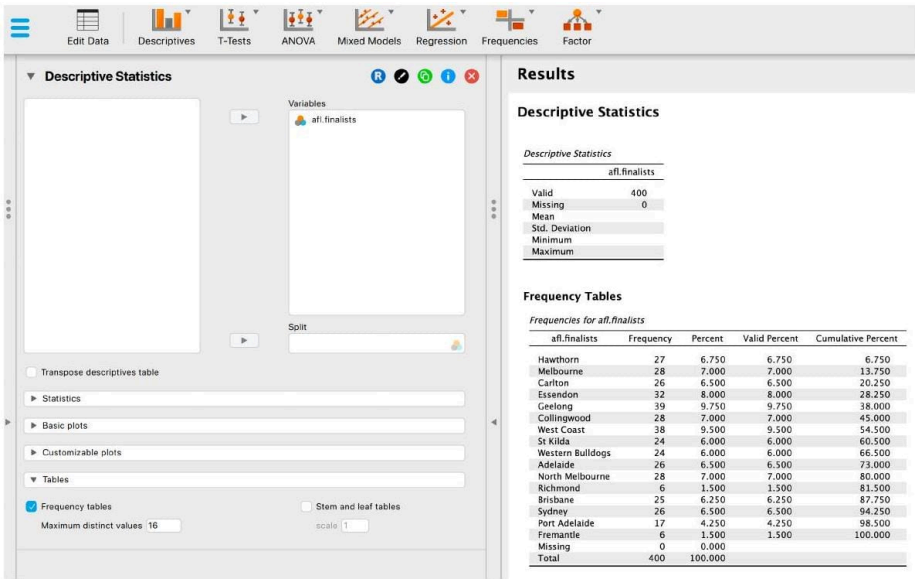


Figure 4.6: A screenshot of JASP showing the frequency table for the afl.finalists variable

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

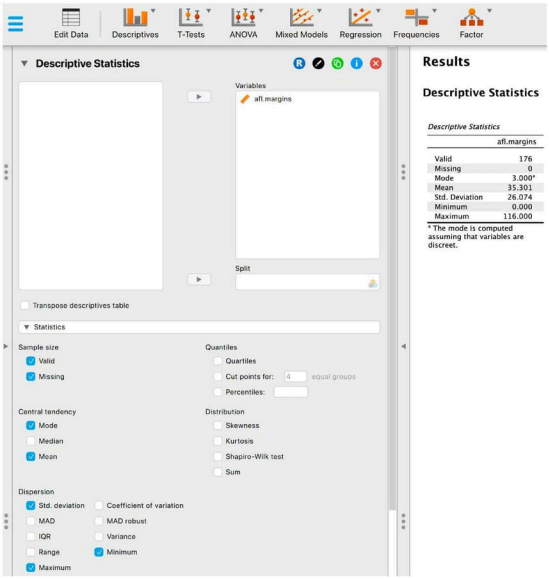


Figure 4.7: A screenshot of JASP showing the modal value for the afl.margins variable

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 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
Mode	3.000^{a}
Mean	35.301

	afl.margins
Std. Deviation	26.074
Minimum	0.000
Maximum	116.000
25th percentile	12.750
50th percentile	30.500
75th percentile	50.500
[a] **The mode is computed assuming that variables are discreet.**	

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 $50.50 - 12.25 = 37.75$, 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 psychology. The measure based on the mean does occasionally show up in psychology 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	value	deviation from mean	absolute deviation
	i	X_i	$X_i - \bar{X}$	$\left\|X_i - \bar{X}\right\|$
	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 do to 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:

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

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, [3] 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:

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

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:

English: maths:	which game i	value X_i	deviation from mean $X_i - \bar{X}$	squared deviation $\left(X_i - \bar{X} \right)^2$
	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:

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

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 7, 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 7.

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 324.64 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:

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

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 7 I'll refer to this new quantity as $\hat{\sigma}$ (read as: "sigma hat"), and the formula for this is:

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

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". [4] 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.

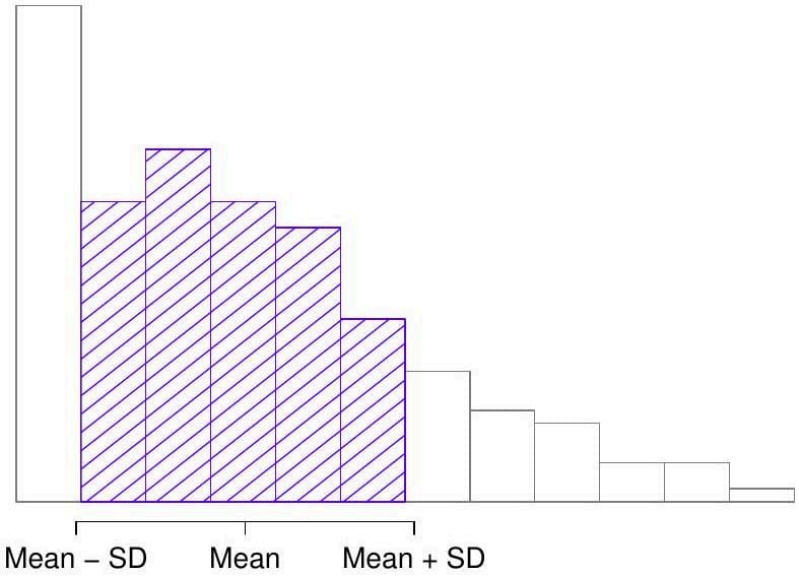


Figure 4.9: An illustration of the standard deviation from the AFL winning margins data. The shaded bars in the histogram show how much of the data fall within one standard deviation of the mean. In this case, 65.3% of the data set lies within this range, which is pretty consistent with the "approximately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

4.2.6 Which measure to use?

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

- terpretable 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 psychological 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.

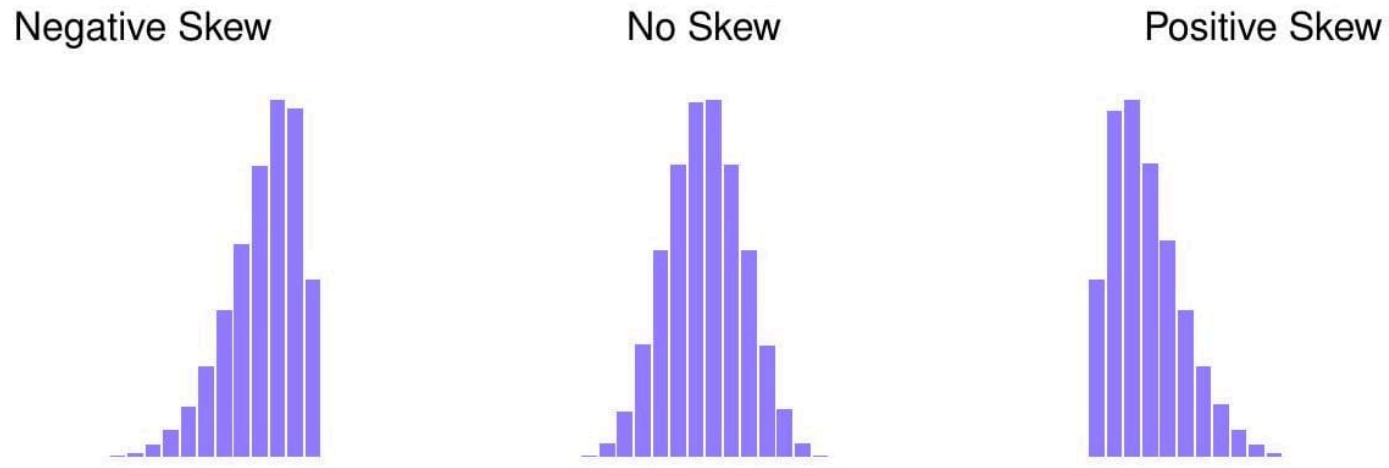


Figure 4.10: An illustration of skewness. On the left we have a negatively skewed data set (skewness $= -.93$), in the middle we have a data set with no skew (well, hardly any: skewness $= -.006$), and on the right we have a positively skewed data set (skewness = .93).

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

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

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 afl.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 an absolute value of 2 or less can mean that the data is not very skewed, and an absolute 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. Put simply, kurtosis describes the tendency for extreme values in a distribution's tails.

Negative excess kurtosis is called platykurtic and indicates lighter tails than a normal distribution. Positive excess kurtosis is called leptokurtic and indicates heavier tails. Zero excess kurtosis is called mesokurtic.

informal term	technical name	kurtosis value
"too flat"	platykurtic	negative
"just pointy enough"	mesokurtic	zero
"too pointy"	leptokurtic	positive

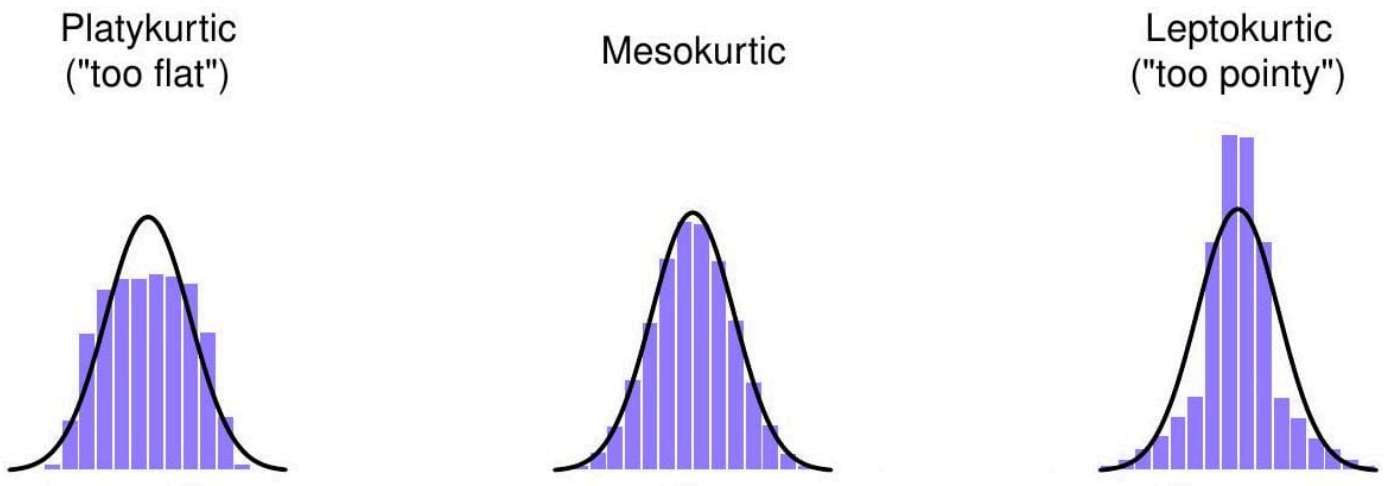


Figure 4.11: An illustration of kurtosis. On the left, we have a "platykurtic" data set (kurtosis = -.95) meaning that the data set is "too flat". In the middle we have a "mesokurtic" data set (kurtosis is almost exactly 0) which means that the pointiness of the data is just about right. Finally, on the right, we have a "leptokurtic" data set (kurtosis $= 2.12$) indicating that the data set is "too pointy". Note that kurtosis is measured with respect to a normal curve (black line).

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: [a]

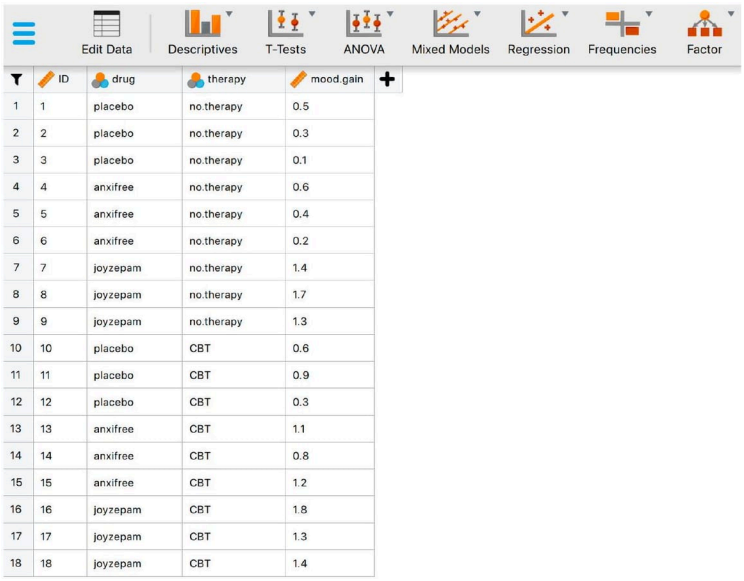
$$\text{kurtosis}(X) = \frac{1}{N\hat{\sigma}^4} \sum_{i=1}^{N} \left(X_i - \bar{X}\right)^4 - 3.$$

[a] 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.

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 just pointy 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 12 (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:



	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

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

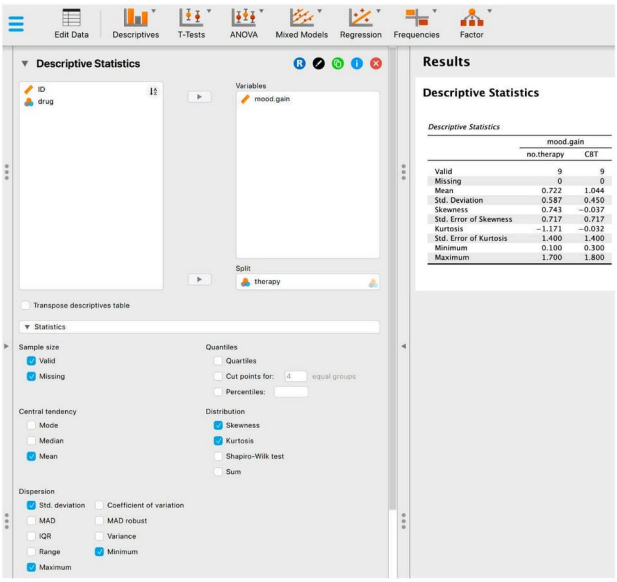


Figure 4.13: A screenshot of JASP showing Descriptives split by therapy type

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 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.

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. [5] 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.

4.6 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 its own story, was its 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 done 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.

---

1. [1] Note for non-Australians: the AFL is an Australian rules football competition. You don't need to know anything about Australian rules in order to follow this section.
2. [2] www.abc.net.au/news/stories/2010/09/24/3021480.htm
3. [3] 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 for uncorrelated variables. Here's what that means. Suppose I have two variables X and Y, whose variances are $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ 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 $\mathrm{Var}(X) + \mathrm{Var}(Y)$. This is a very useful property, but it's not true of the other measures that I talk about in this section.
4. [4] Strictly, the assumption is that the data are normally distributed, which is an important concept that we'll discuss more in Chapter 6 and will turn up over and over again later in the book.
5. [5] 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.

This page titled 4.1: Descriptive statistics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 4: Descriptive statistics by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

4.2: Descriptive Statistics in JASP

Getting Started – Entering 'Analysis Mode'

JASP has two 'modes' that allow you to access different options:

1. **Data View Mode** – allows you to edit your data; create, modify, and delete variables; compute new variables from existing ones

2. **Analysis Mode** – allows you to 'work' with the data. This includes obtaining descriptive statistics, performing different analyses, and generating tables and graphs.

When working with a dataset you first want to get an overview of the data. For example, what is the distribution of the data? What are the key descriptive statistics that encapsulate this data? All of the tools for analysing your data can be found by clicking on the green **Analyses** button which will take you out of the 'Data View Mode' and into 'Analysis Mode'.

image

When you click 'Analyses' the menu options at the top change, If you need to go back into 'Data View Mode', you simply need to click the **Edit Data** button which has replaced the 'Analyses' button.

image

Descriptive Statistics

To find the tools for descriptive statistics, plots, and graphs you need to click the **Descriptives** button which is the first option in the top menu.

When you click the **Descriptives** button, a new screen will appear, like so:

image

Your new screen is now **split into two:**

- **Left-hand side** has the variables that you can select with a number of expandable menus: 'Statistics', 'Basic plots', 'Customizable plots', and 'Tables' (you will need these later).
- **Right-hand side** shows you the output from the options you can select on the left-hand side.

There are two main options for extracting descriptive statistics from your data depending upon the kind of information you want to obtain. These have been explained in the next two sections.

Continuous Variables for a Single Group of Participants

1. Move the continuous variable(s) you are interested in into the **Variables** box. Continuous (scale) variables are indicated by a ruler symbol.

2. As soon as you move the variable(s) across, the **Descriptive Statistics table** on the right-hand side of the screen will populate with information (e.g., mean and standard deviation).

image

Continuous Variables for Different Groups of Participants

1. Move the continuous variable(s) you are interested in into the **Variables** box.

2. Move the categorical variable you want to use to create separate groups into the **Split** box below **Variables**. Categorical (nominal) variables are indicated by a Venn diagram symbol.

3. As soon as you move the variable(s) across, the **Descriptive Statistics table** on the right-hand side of the screen will populate with information (e.g., mean and standard deviation for **each group of participants**).

image

4. You can change the statistics provided in the Descriptives Statistics table. To do so, click on the **Statistics** button to expand the menu and select the statistics you wish to see in the table. The most common statistics have already been selected in the screenshot below.

image

Note: The best thing about JASP is that you can click and unclick as many options as you like! You will be able to see the output from your selected options immediately on the right-hand side of the screen. Feel free to play around and experiment.

Investigating Outliers – Calculating the Z-Score

Before you can start analysing your data, you will need to check some of the most common assumptions for inferential parametric tests. For example, whether there are any outliers (i.e., unusual values) in the data.

In this section, we show you how to **detect extreme outliers for continuous variables using z-scores.** [1]

How do I do that?

1. Create a new variable next to the variable you want to calculate the z-scores for.

2. Name this variable 'z-[name of original variable]', then in the **Computed type menu** select 'Computed with drag-and-drop'.

image

image

3. Scroll down the menu on the right-hand side of the computed column box until you see **zScores(y)**.

image

4. Drag and drop **zScores(y)** into the large box in the centre. Then drag the original variable so that it falls in between the brackets.

image image

5. The new variable will now be populated with the z-scores.

image

6. **Optional:** You can sort the z-scores into descending or ascending order to more easily see extreme values. Remember that outliers can be positive or negative so you should check both ends of the column.

---

1. The z-score is a standardised statistic that tells you how many standard deviations away from the mean each data point is. A general rule of thumb is that z-scores above 3 or below -3 are extreme outliers. 3.01: Descriptive Statistics#return-footnote-244-1

This page titled 4.2: Descriptive Statistics in JASP is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss & Stanislava Naneva via source content that was edited to the style and standards of the LibreTexts platform.

5: VISUALIZING DATA

5.1: Drawing graphs
5.2: Basic Plots
5.3: Tables
5.4: Exporting and Editing Plots from JASP

This page titled 5: Visualizing Data is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

5.1: Drawing graphs

> Above all else show the data.
> -Edward Tufte [1]

Visualising data is one of the most important tasks facing the data analyst. It's important for two distinct but closely related reasons. Firstly, there's the matter of drawing "presentation graphics", displaying your data in a clean, visually appealing fashion makes it easier for your reader to understand what you're trying to tell them. Equally important, perhaps even more important, is the fact that drawing graphs helps you to understand the data. To that end, it's important to draw "exploratory graphics" that help you learn about the data as you go about analysing it. These points might seem pretty obvious but I cannot count the number of times I've seen people forget them.

To give a sense of the importance of this chapter, I want to start with a classic illustration of just how powerful a good graph can be. To that end, Figure 5.1 shows a redrawing of one of the most famous data visualisations of all time. This is John Snow's 1854 map of cholera deaths. The map is elegant in its simplicity. In the background we have a street map which helps orient the viewer. Over the top we see a large number of small dots, each one representing the location of a cholera case. The larger symbols show the location of water pumps, labelled by name. Even the most casual inspection of the graph makes it very clear that the source of the outbreak is almost certainly the Broad Street pump. Upon viewing this graph Dr Snow arranged to have the handle removed from the pump and ended the outbreak that had killed over 500 people. Such is the power of a good data visualisation.

The goals in this chapter are twofold. First, to discuss several fairly standard graphs that we use a lot when analysing and presenting data, and second to show you how to create these graphs in JASP. The graphs themselves tend to be pretty straightforward, so in one respect this chapter is pretty simple. Where people usually struggle is learning how to produce graphs, and especially learning how to produce good graphs. Fortunately, learning how to draw graphs in JASP is reasonably simple as long as you're not too picky about what your graph looks like. What I mean when I say this is that JASP has very good default graphs, or plots, that most of the time produce a clean, high-quality graphic. However, on those occasions when you do want to do something non-standard, or if you need to make highly specific changes to the figure, then the graphics functionality in JASP is not yet capable of supporting advanced work or detail editing.

Snow's cholera map of London

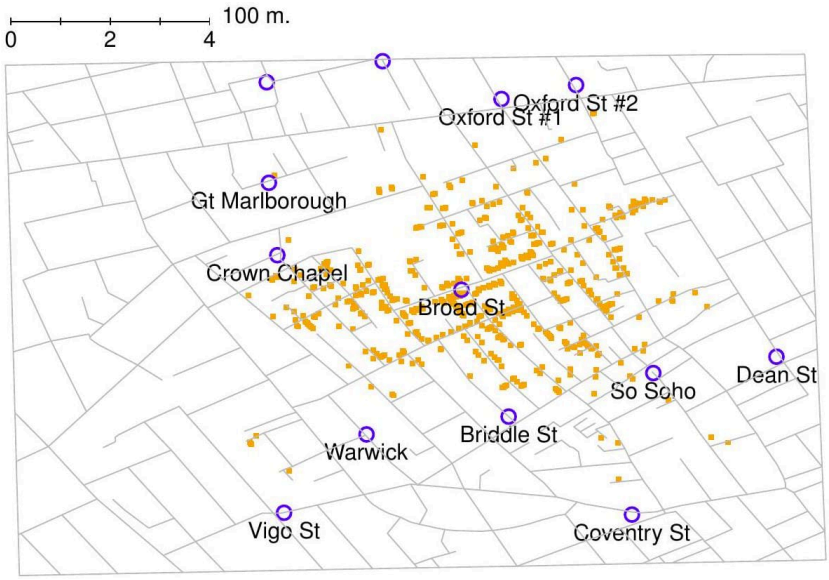


Figure 5.1: A stylised redrawing of John Snow's original cholera map. Each small dot represents the location of a cholera case and each large circle shows the location of a well. As the plot makes clear, the cholera outbreak is centred very closely on the Broad St pump.

5.1 Histograms

Let's begin with the humble histogram. Histograms are one of the simplest and most useful ways of visualising data. They make most sense when you have an interval or ratio scale variable (e.g., the afl.margins data from Chapter 4) and what you want to do is get an overall impression of the variable. Most of you probably know how histograms work, since they're so widely used, but for the sake of completeness I'll describe them. All you do is divide up the possible values into bins and then count the number of observations that fall within each bin. This count is referred to as the frequency of the bin and is displayed as a vertical bar. Ihe AFL winning margins data there are 33 games in which the winning margin was less than 10 points and it is this fact that is represented by
the height of the leftmost bar that we showed earlier in Chapter 4, Figure 4.2. The earlier graphs were generated using an advanced plotting package in R which, for now, is beyond the capability of JASP. But JASP gets us close, and drawing this histogram in JASP is pretty straightforward. Open up the 'Plots' menu under 'Descriptives' - 'Descriptive Statistics' and click the 'Distribution plots' check box, as in Figure 5.2. JASP defaults to labelling the y-axis as 'Counts' and the x-axis with the variable name. The bins are selected automatically. Note that while counts are displayed, the actual values do not matter too much. Rather, what we are really interested in is our impression of the shape of the distribution: is it normally distributed or is there a skew or kurtosis? Our first impressions of these characteristics come from drawing a histogram.

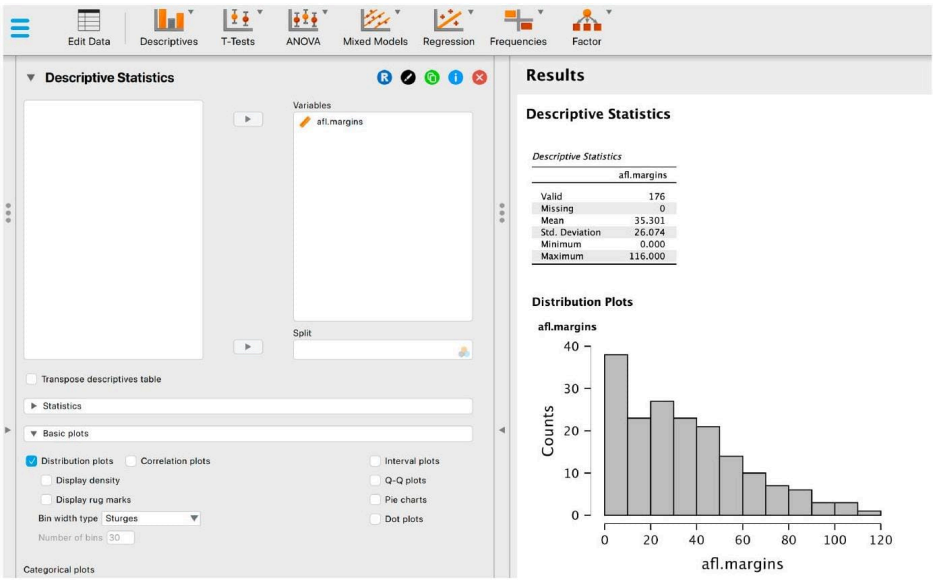


Figure 5.2: JASP screenshot showing the 'Distribution plots' option and accompanying histogram.

One additional feature that JASP provides is the ability to plot a 'Density' curve on top of this histogram. You can do this by clicking the 'Display density' check box under the 'Plots' options, and this gives us the plot shown in Figure 5.3. A density plot visualises the distribution of data over a continuous interval or time period. This chart is a variation of a histogram that uses kernel smoothing to plot values, allowing for smoother distributions by smoothing out the noise. The peaks of a density plot help display where values are concentrated over the interval. An advantage density plots have over histograms is that they are better at determining the distribution shape because they're not affected by the number of bins used (each bar used in a typical histogram). A histogram comprising of only 4 bins wouldn't produce a distinguishable enough shape of distribution as a 20-bin histogram would. However, with density plots, this isn't an issue.

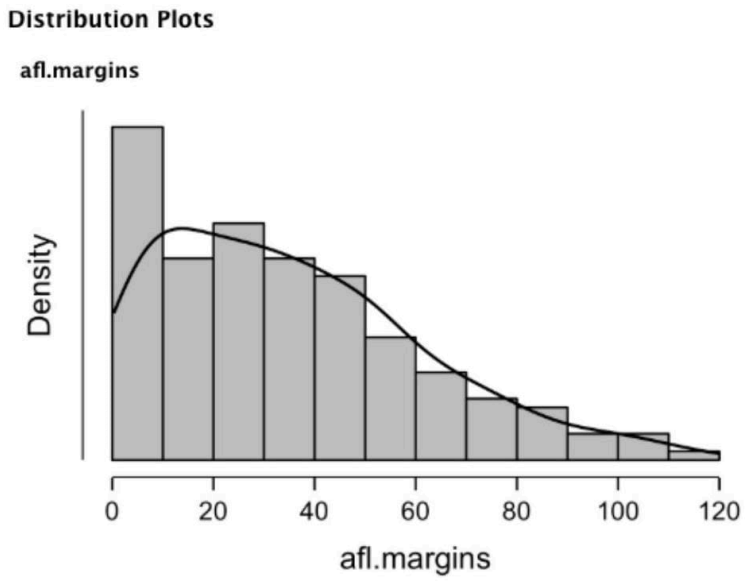


Figure 5.3: A density plot of the afl.margins variable plotted in JASP

Although this image would need a lot of cleaning up in order to make a good presentation graphic (i.e., one you'd include in a report), it nevertheless does a pretty good job of describing the data. In fact, the big strength of a histogram or density plot is that (properly used) it does show the entire spread of the data, so you can get a pretty good sense about what it looks like. The downside to histograms is that they aren't very compact. Unlike some of the other plots I'll talk about it's hard to cram 20-30 histograms into a single image without overwhelming the viewer. And of course, if your data are nominal scale then histograms are useless.

5.2 Boxplots

Another alternative to histograms is a boxplot, sometimes called a "box and whiskers" plot. Like histograms they're most suited to interval or ratio scale data. The idea behind a boxplot is to provide a simple visual depiction of the median, the interquartile range, and the range of the data. And because they do so in a fairly compact way boxplots have become a very popular statistical graphic, especially during the exploratory stage of data analysis when you're trying to understand the data yourself. Let's have a look at how they work, again using the afl.margins data as our example.

The easiest way to describe what a boxplot looks like is just to draw one. Click on the 'Boxplots'

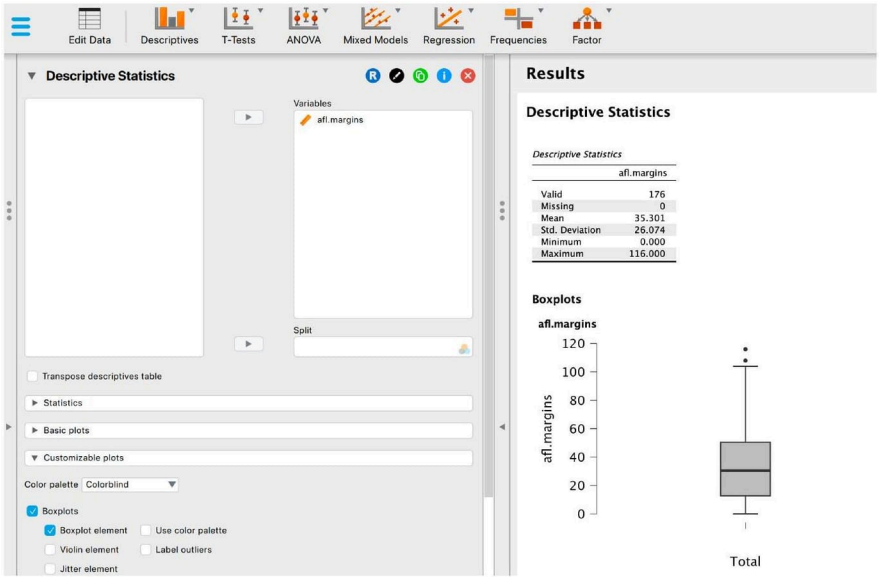


Figure 5.4: A box plot of the afl.margins variable plotted in JASP

check box and you will get the plot shown on the lower right of Figure 5.4. By default, JASP has drawn the most basic boxplot possible. When you look at this plot this is how you should interpret it: the thick line in the middle of the box is the median; the box itself spans the range from the 25th percentile to the 75th percentile; and the "whiskers" go out to the most extreme data point that doesn't exceed a certain bound. By default, this value is 1.5 times the interquartile range (IQR), calculated as 25th percentile - (1.5*IQR) for the lower boundary, and 75th percentile + (1.5*IQR) for the upper boundary. Any observation whose value falls outside this range is plotted as a circle or dot instead of being covered by the whiskers, and is commonly referred to as an outlier. For our AFL margins data there are two observations that fall outside this range, and these observations are plotted as dots (the upper boundary is 107, and looking over the data column in the spreadsheet there are two observations with values higher than this, 108 and 116, so these are the dots).

5.2.1 Violin plots

A variation to the traditional box plot is the violin plot. Violin plots are similar to box plots except that they also show the kernel probability density of the data at different values. Typically, violin plots will include a marker for the median of the data and a box indicating the interquartile range, as in

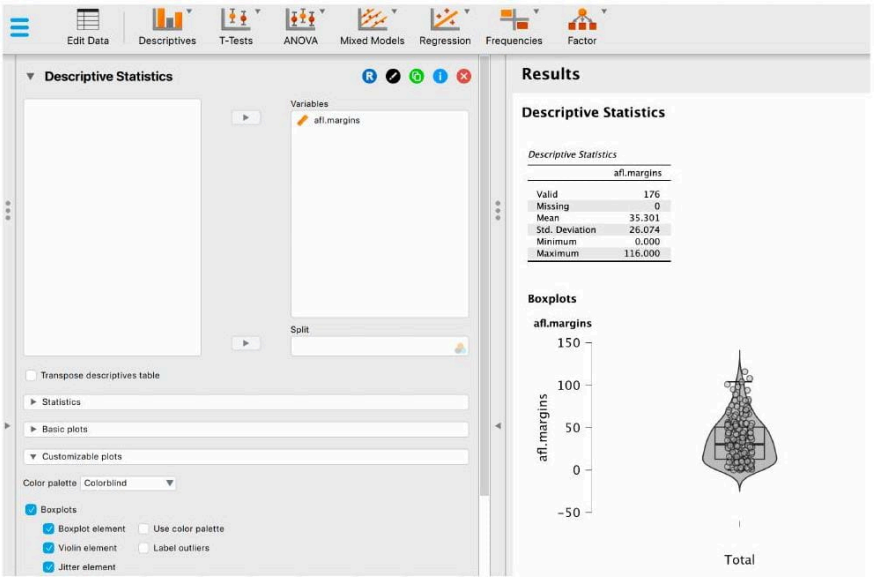


Figure 5.5: A violin plot of the afl.margins variable plotted in JASP, also showing a box plot and data points

standard box plots. In JASP you can achieve this sort of functionality by checking both the 'Violin element' and the 'Boxplot element' check boxes. See Figure 5.5, which also has the data plotted (select the 'Jitter element' check box to show the actual data points on the plot).

5.2.2 Drawing multiple boxplots

One last thing. What if you want to draw multiple boxplots at once? Suppose, for instance, I wanted separate boxplots showing the AFL margins not just for 2010 but for every year between 1987 and 2010. To do that the first thing we'll have to do is find the data. These are stored in the aflsmall2.csv file. So let's load it into JASP and see what is in it. You will see that it is a pretty big data set. It contains 4296 games and the variables that we're interested in. What we want to do is have JASP draw boxplots for the margin variable, but plotted separately for each year. The way to do this is to first change year to a nominal variable, then year across into the 'Split' box.

The result is shown in Figure 5.6. This version of the box plot, split by year, gives a sense of why it's sometimes useful to choose box plots instead of histograms. It's possible to get a good sense of what the data look like from year to year without getting overwhelmed with too much detail. Now imagine what would have happened if I'd tried to cram 24 histograms into this space: no chance at all that the reader is going to learn anything useful.

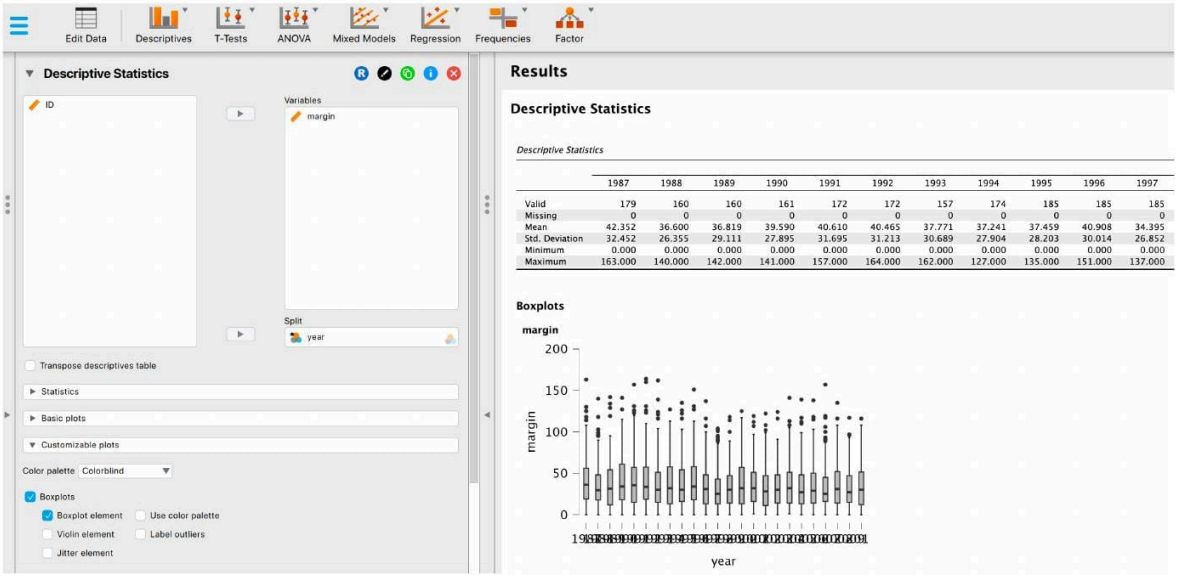


Figure 5.6: Multiple boxplots plotted in JASP, for the margin by year variables in the aflsmall2 data set

5.3 Saving image files using JASP

Hold on, you might be thinking. What's the good of being able to draw pretty pictures in JASP if I can't save them and send them to friends to brag about how awesome my data is? How do I save the picture? Simples. Just click on the triangle beside the header of your plot and select 'Save Image As'. You can save it in one of several formats, including 'png', 'pdf', 'eps', or 'tif'. These formats all produce nice images that you can the send to your friends, or (perhaps more importantly) include in your assignments or papers.

5.4 Summary

Perhaps I'm a simple minded person, but I love pictures. Every time I write a new scientific paper one of the first things I do is sit down and think about what the pictures will be. In my head an article is really just a sequence of pictures linked together by a story. All the rest of it is just window dressing.

What I'm really trying to say here is that the human visual system is a very powerful data analysis tool. Give it the right kind of information and it will supply a human reader with a massive amount of knowledge very quickly. Not for nothing do we have the saying "a picture is worth a thousand words". With that in mind, I think that this is one of the most important chapters in the book. The topics covered were:

- -Common plots. Much of the chapter was focused on some of the standard graphs that statisticians like to produce: histograms (Section 5.1) and boxplots (Section 5.2)
- -Saving image files. Importantly, we also covered how to export your pictures (Section 5.3)

One final thing to point out. While JASP produces some really neat default graphics, editing the plots is currently not possible. For more advanced graphics and plotting capability the packages available in R are much more powerful. One of the most popular graphics systems is provided by the ggplot2 package (see http://ggplot2.org/), which is loosely based on "The grammar of graphics" (Wilkinson, Wills, Rope, Norton, and Dubbs 2006). It's not for novices. You need to have a pretty good grasp of R before you can start using it, and even then it takes a while to really get the hang of it. But when you're ready it's worth taking the time to teach yourself, because it's a much more powerful and cleaner system.

---

1. [1] The origin of this quote is Tufte's lovely book The Visual Display of Quantitative Information.

This page titled 5.1: Drawing graphs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 5: Drawing graphs by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

5.2: Basic Plots

The second expandable menu in 'Descriptive Statistics' is called **Basic plots**.

To see the options in this menu, click on 'Basic plots'. You can then select which types of plots (i.e., graphs) you want to appear on the **right-hand side** of the screen. Remember that you can always unselect options in JASP so do not be afraid to experiment.

image

Here are some of the most common plots/graphs you are likely to use as a Psychology student:

1. Distribution plots (i.e., histograms)

When selecting distribution plots you may also want to select **Display density** as this overlays a line that will help you assess the normality of the distribution visually.

image

2. Correlation plots (i.e., scatterplots)

Selecting correlation plots will give you both a scatterplot and a histogram. There are other options for creating scatterplots which are covered in Section 11.

image

3. Interval plots (i.e., error bars)

Interval plots show the mean and its confidence interval. This is particularly useful when comparing two groups or conditions.

image

This page titled 5.2: Basic Plots is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss & Stanislava Naneva via source content that was edited to the style and standards of the LibreTexts platform.

5.3: Tables

The fourth and final expandable menu in 'Descriptive Statistics' is called **Tables**.

This menu has two options: **Frequency tables** and **Stem and leaf tables**. You are only likely to use the 'Frequency tables' option as it allows you to obtain information about the frequencies for categorical variables which is usually the first step in a chi-squared analysis.

image

image

The most important information in the Frequency Table output is:

- **Total** number of people in your sample. In statistics, we denote the total sample size with the capital letter *N* in italics.
- The **Frequency** (i.e., number) of people in each group of the sample. This is only really of interest when you have a categorical variable.

This page titled 5.3: Tables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss & Stanislava Naneva via source content that was edited to the style and standards of the LibreTexts platform.

5.4: Exporting and Editing Plots from JASP

While you can create some excellent graphs with Excel, JASP allows you to export your plots (e.g., histograms) as images or directly into Microsoft PowerPoint.

This section will show you how to export plots directly from JASP and how you can edit them.

Exporting Plots as Images

The main benefit of exporting your plots directly from JASP is that you do not need to move your data to an Excel spreadsheet and spend time creating graphs from scratch.

Note: Using the method below will not allow you to edit your plots outside of JASP. This may be ok but many times there are requirements for how specific plots should look. An alternative method of exporting is covered further down in this section.

How do I do that?

1. Once you have generated your plots(s), you can edit their dimensions by dragging at the corner of each graph, like so:

image

2. Once you have adjusted your graph to the desired size, click on the downward triangle (▼) next to the title of the graph and select **Save Image As**.

image

2. Save your image in the desired location. The default format is .png but you can select different file formats. We recommend the default file type.

image

3. You can then insert the image in another document such as Microsoft Word.

Exporting Plots as Editable Files

While exporting graphs as images is quick and easy, those images cannot be edited which may sometimes be required or preferable. For example, you may wish to add colour or adjust the graph so that it fully complies with the American Psychological Association (APA) formatting style. [1]

How do I do that?

1. Click on the downward triangle (▼) next to the title of the graph and select **Save Image As**.

image

2. Save the graph in PowerPoint (.pptx).

image

3. Open your graph in Microsoft PowerPoint. It should look something like this:

image

4. Now you can edit your graph as you see fit or copy it directly into Word (or another document) by right-clicking on the graph, selecting **Copy**, and then pasting it into another document.

image

Editing JASP Plots in PowerPoint/Word

You can also edit JASP plots and tables in Microsoft PowerPoint and Word. To learn how to export plots, see the previous section.

An explanation of how plots/graphs (also referred to as figures) should be edited to be compliant with the APA style can be found on the APA Style Website.

Example

A final-year psychology student needs to create a simple bar chart for their dissertation to represent their findings from a *t*-test. Their bar chart will be labelled as *Figure 1* and needs to be compliant with the APA style.

How do I do that?

1. Open your saved JASP plot in PowerPoint (or copy it from PowerPoint into Word).

2. **Remove the background/grid** – click twice anywhere on the background of the bar chart and press 'Delete'. **DO NOT click twice fast!**

3. Click on the graph once. You should see a border around it, like so:

image

4. Next, **adjust the font** if it is not already APA-compliant. Select *Calibri* or *Arial* from the fonts drop-down menu. Adjust the size of the font as required.

image

5. You can move and edit any text on the graph by double-clicking the text boxes (invisible until clicked). You can also move the text anywhere on the graph.

Note: Depending on how you input your variables in JASP, this step may not be necessary. See the example below.

image

6. You can also **change the colour** of any element on the graph. For example, the colour of the bars. Click twice on the bar (or element) you want to change until you see a border around it. You should now be able to see a menu called **Shape Format** at the top of the document. Click on it and select your chosen colour under **Shape Fill**.

image

7. After you have done this for all elements, your graph should look something like this:

image

8. The final step is to **label and title your graph**. Under the **Home** menu, there should be a subsection called 'Drawings'. Find it and select **Text Box**. Draw a text box anywhere. You can adjust font size and placement later. Label and name your graph, like so:

image

Note: In APA, all graphs are labelled numerically in the order in which they are presented in-text and are called 'Figures'. The figures also require titles that describe the content of the graph so that the reader can understand what information is being presented.

---

1. You can find out more about the APA style here: apastyle.apa.org ↩

This page titled 5.4: Exporting and Editing Plots from JASP is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss & Stanislava Naneva via source content that was edited to the style and standards of the LibreTexts platform.

6: PROBABILITY AND DISTRIBUTIONS

6.1: Introduction to probability

This page titled 6: Probability and Distributions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

6.1: Introduction to probability

6. Introduction to probability

> [God] has afforded us only the twilight . . . of Probability.
> - John Locke

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

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

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

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

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

Inferential statistics provides the tools that we need to answer these sorts of questions, and since these kinds of questions lie at the heart of the scientific enterprise, they take up the lions share of every introductory course on statistics and research methods. However, the theory of statistical inference is built on top of probability theory. And it is to probability theory that we must now turn. This discussion of probability theory is basically background detail. There's not a lot of statistics per se in this chapter, and you don't need to understand this material in as much depth as the other chapters in this part of the book. Nevertheless, because probability theory does underpin so much of statistics, it's worth covering some of the basics.

6.1 How are probability and statistics different?

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

- What are the chances of a fair coin coming up heads 10 times in a row?
- If I roll a six sided dice twice, how likely is it that I'll roll two sixes?

- How likely is it that five cards drawn from a perfectly shuffled deck will all be hearts?
- What are the chances that I'll win the lottery?

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

$$P(\text{ heads }) = 0.5 \tag{6.1.1}$$

which you can read as "the probability of heads is 0.5 ". As we'll see later, in the same way that percentages are numbers that range from 0\% to 100\%, probabilities are just numbers that range from 0 to 1. When using this probability model to answer the first question I don't actually know exactly what's going to happen. Maybe I'll get 10 heads, like the question says. But maybe I'll get three heads. That's the key thing. In probability theory the model is known but the data are not.

So that's probability. What about statistics? Statistical questions work the other way around. In statistics we do not know the truth about the world. All we have is the data and it is from the data that we want to learn the truth about the world. Statistical questions tend to look more like these:

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

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

H H H H H H H H H H

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

6.2 What does probability mean?

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

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

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

Each of these seems sensible. However, they're not identical and not every statistician would endorse all of them. The reason is

that there are different statistical ideologies (yes, really!) and depending on which one you subscribe to, you might say that some of those statements are meaningless or irrelevant. In this section I give a brief introduction the two main approaches that exist in the literature. These are by no means the only approaches, but they're the two big ones.

6.2.1 The frequentist view

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

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

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

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

Notice that at the start of the sequence the proportion of heads fluctuates wildly, starting at .00 and rising as high as .80. Later on, one gets the impression that it dampens out a bit, with more and more of the values actually being pretty close to the “right” answer of .50. This is the frequentist definition of probability in a nutshell. Flip a fair coin over and over again, and as N grows large...(approaches infinity, denoted $N \rightarrow \infty$) the proportion of heads will converge to 50%. There are some subtle technicalities that the mathematicians care about, but qualitatively speaking that's how the frequentists define probability. Unfortunately, I don't have an infinite number of coins or the infinite patience required to flip a coin an infinite number of times. However, I do have a computer and computers excel at mindless repetitive tasks. So I asked my computer to simulate flipping a coin 1000 times and then drew a picture of what happens to the proportion N_H/N as N increases. Actually, I did it four times just to make sure it wasn't a fluke. The results are shown in Figure 6.1. As you can see, the proportion of observed heads eventually stops fluctuating and settles down. When it does, the number at which it finally settles is the true probability of heads.

The frequentist definition of probability has some desirable characteristics. First, it is objective. The probability of an event is necessarily grounded in the world. The only way that probability statements can make sense is if they refer to (a sequence of) events that occur in the physical universe. [1] Secondly, it is unambiguous. Any two people watching the same sequence of events unfold, trying to calculate the probability of an event, must inevitably come up with the same answer.

However, it also has undesirable characteristics. First, infinite sequences don't exist in the physical world. Suppose you picked up a coin from your pocket and started to flip it. Every time it lands it impacts on the ground. Each impact wears the coin down a bit. Eventually the coin will be destroyed. So, one might ask whether it really makes sense to pretend that an "infinite" sequence of coin flips is even a meaningful concept, or an objective one. We can't say that an "infinite sequence" of events is a real thing in the physical universe, because the physical universe doesn't allow infinite anything. More seriously, the frequentist definition has a narrow scope. There are lots of things out there that human beings are happy to assign probability to in everyday language, but cannot (even in theory) be mapped onto a hypothetical sequence of events. For instance, if a meteorologist comes on TV and says "the probability of rain in Adelaide on 2 November 2048 is 60\%" we humans are happy to accept this. But it's not clear how to define this in frequentist terms. There's only one city of Adelaide, and only one 2 November 2048. There's no infinite sequence of events here, just a one-off thing. Frequentist probability genuinely forbids us from making probability statements about a single event. From the frequentist perspective it will either rain tomorrow or it will not. There is no "probability" that attaches to a single non-repeatable event. Now, it should be said that there are some very clever tricks that frequentists can use to get around this. One possibility is that what the meteorologist means is something like "There is a category of days for which I predict a 60\% chance of rain, and if we look only across those days for which I make this prediction, then on 60\% of those days it will actually rain". It's very weird and counterintuitive to think of it this way, but you do see frequentists do this sometimes. And it will come up

later in this book.

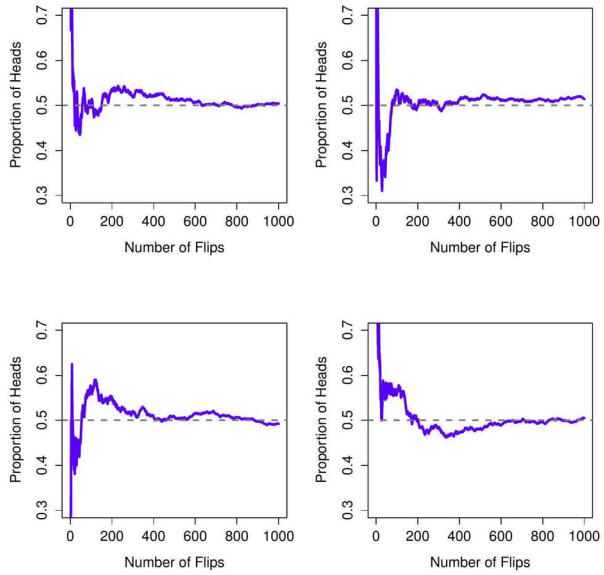


Figure 6.1.1: An illustration of how frequentist probability works. If you flip a fair coin over and over again the proportion of heads that you've seen eventually settles down and converges to the true probability of 0.5. Each panel shows four different simulated experiments. In each case we pretend we flipped a coin 1000 times and kept track of the proportion of flips that were heads as we went along. Although none of these sequences actually ended up with an exact value of .5, if we'd extended the experiment for an infinite number of coin flips they would have.

6.2.2 The Bayesian view

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

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

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

6.2.3 What's the difference? And who is right?

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

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

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

In any case, whilst I personally prefer the Bayesian view, the majority of statistical analyses are based on the frequentist approach. My reasoning is pragmatic. The goal of this book is to cover roughly the same territory as a typical undergraduate stats class in psychology, and if you want to understand the statistical tools used by most psychologists you'll need a good grasp of frequentist methods. I promise you that this isn't wasted effort. Even if you end up wanting to switch to the Bayesian perspective, you really should read through at least one book on the "orthodox" frequentist view. Besides, I won't completely ignore the Bayesian perspective. Every now and then I'll add some commentary from a Bayesian point of view.

6.3 Basic probability theory

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

6.3.1 Introducing probability distributions

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

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

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

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

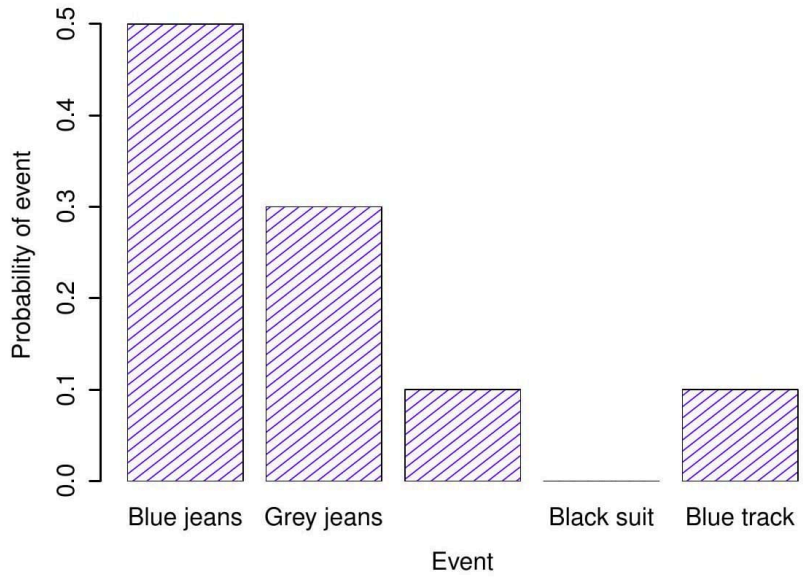


Figure 6.1.2: A visual depiction of the “tro users” probability distribution. There are five “elementary events”, corresponding to the five pairs of trousers that I own. Each event has some probability of occurring: this probability is a number between 0 to 1. The sum of these probabilities is 1.

Each of the events has a probability that lies between 0 and 1 , and if we add up the probability of all events they sum to 1. Awesome. We can even draw a nice bar graph to visualise this distribution, as shown in Figure 6.2. And, at this point, we've all achieved something. You've learned what a probability distribution is, and I've finally managed to find a way to create a graph that focuses entirely on my trousers. Everyone wins!

The only other thing that I need to point out is that probability theory allows you to talk about non elementary events as well as elementary ones. The easiest way to illustrate the concept is with an example. In the trousers example it's perfectly legitimate to refer to the probability that I wear jeans. In this scenario, the "Dan wears jeans" event is said to have happened as long as the elementary event that actually did occur is one of the appropriate ones. In this case "blue jeans", "black jeans" or "grey jeans". In mathematical terms we defined the "jeans" event E to correspond to the set of elementary events (X_1, X_2, X_3). If any of these elementary events occurs then E is also said to have occurred. Having decided to write down the definition of the E this way, it's pretty straightforward to state what the probability $P(E)$ is: we just add everything up. In this particular case

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

and, since the probabilities of blue, grey and black jeans respectively are .5, . 3 and .1, the probability that I wear jeans is equal to .9 .At this point you might be thinking that this is all terribly obvious and simple and you'd be right. All we've really done is wrap some basic mathematics around a few common sense intuitions. However, from these simple beginnings it's possible to construct some extremely powerful mathematical tools. I'm definitely not going to go into the details in this book, but what I will

do is list, in Table 6.1, some of the other rules that probabilities satisfy. These rules can be derived from the simple assumptions that I've outlined above, but since we don't actually use these rules for anything in this book I won't do so here.

6.4 The binomial distribution

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

Table 6.1: Some basic rules that probabilities must satisfy. You don't really need to know these rules in order to understand the analyses that we'll talk about later in the book, but they are important if you want to understand probability theory a bit more deeply.

English	Notation		Formula	
Not A	$P(A^c) = 1 - P(A)$	=	1-P(A)	
A or B	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	=	P(A)+P(B)-P(A nn B)	
A and B	$P(A \cap B) = P(A \mid B)P(B)$	=	P(Aâˆ£B)P(B)	

6.4.1 Introducing the binomial

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

Table 6.2: Formulas for the binomial and normal distributions. We don't really use these formulas for anything in this book, but they're pretty important for more advanced work, so I thought it might be best to put them here in a table, where they can't get in the way of the text. In the equation for the binomial, $X!$ is the factorial function (i.e., multiply all whole numbers from 1 to X), and for the normal distribution "exp" refers to the exponential function. If these equations don't make a lot of sense to you, don't worry too much about them.

Binomial	Normal
$P(X \mid \theta, N) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{N-X}$	$p(X \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$

As usual, we'll want to introduce some names and some notation. We'll let N denote the number of dice rolls in our experiment, which is often referred to as the size parameter of our binomial distribution. We'll also use θ to refer to the the probability that a single die comes up skulls, a quantity that is usually called the success probability of the binomial. [3] Finally, we'll use X to refer to the results of our experiment, namely the number of skulls I get when I roll the dice. Since the actual value of X is due to chance we refer to it as a random variable. In any case, now that we have all this terminology and notation we can use it to state the problem a little more precisely. The quantity that we want to calculate is the probability that $X = 4$ given that we know that $\theta = .167$ and $N = 20$. The general "form" of the thing I'm interested in calculating could be written as

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

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

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

Yeah, yeah. I know what you're thinking: notation, notation, notation. Really, who cares? Very few readers of this book are here for the notation, so I should probably move on and talk about how to use the binomial distribution. I've included the formula for the binomial distribution in Table 6.2, since some readers may want to play with it themselves, but since most people probably don't care that much and because we don't need the formula in this book, I won't talk about it in any detail. Instead, I just want to show you what the binomial distribution looks like.

To that end, Figure 6.3 plots the binomial probabilities for all possible values of X for our dice rolling experiment, from $X = 0$ (no skulls) all the way up to $X = 20$ (all skulls). Note that this is basically a bar chart, and is no different to the "trousers probability" plot I drew in Figure 6.2. On the horizontal axis we have all the possible events, and on the vertical axis we can read off the probability of each of those events. So, the probability of rolling 4 skulls out of 20 is about 0.20 (the actual answer is 0.20220036, as we'll see in a moment). In other words, you'd expect that to happen about 20% of the times you repeated this experiment.

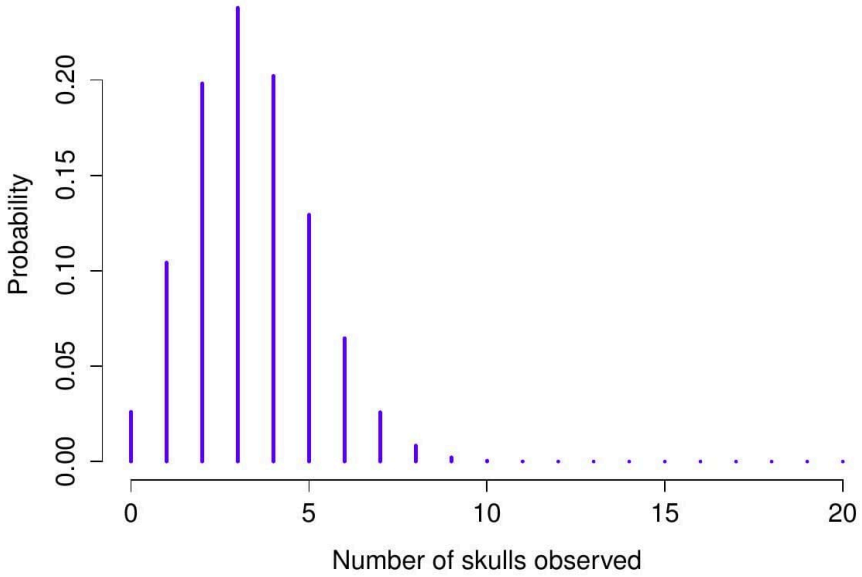


Figure 6.1.3: The binomial distribution with size parameter of $N = 20$ and an underlying success probability of $\theta = 1/6$. Each vertical bar depicts the probability of one specific outcome (i.e., one possible value of X). Because this is a probability distribution, each of the probabilities must be a number between 0 and 1 , and the heights of the bars must sum to 1 as well.

To give you a feel for how the binomial distribution changes when we alter the values of θ and N, let's suppose that instead of rolling dice I'm actually flipping coins. This time around, my experiment involves flipping a fair coin repeatedly and the outcome that I'm interested in is the number of heads that I observe. In this scenario, the success probability is now $\theta = 1/2$. Suppose I were to flip the coin $N = 20$ times. In this example, I've changed the success probability but kept the size of the experiment the same. What does this do to our binomial distribution? Well, as Figure 6.4a shows, the main effect of this is to shift the whole distribution, as you'd expect. Okay, what if we flipped a coin $N = 100$ times? Well, in that case we get Figure 6.4b. The distribution stays roughly in the middle but there's a bit more variability in the possible outcomes.

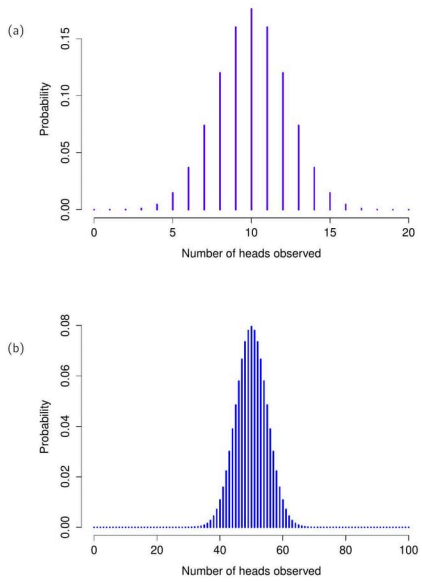


Figure 6.1.4: Two binomial distributions, involving a scenario in which I'm flipping a fair coin, so the underlying success probability is $\theta = 1/2$. In panel (a), we assume I'm flipping the coin $N = 20$ times. In panel (b) we assume that the coin is flipped $N = 100$ times.

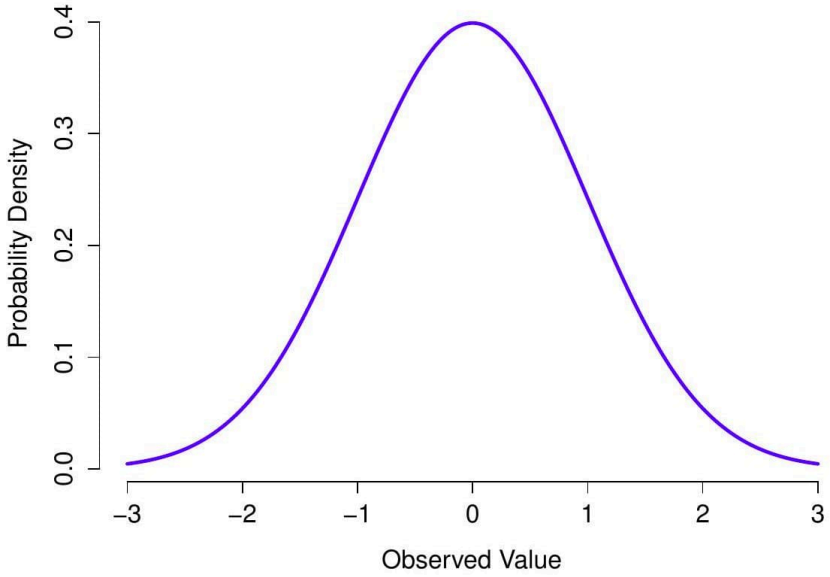


Figure 6.1.5: The normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The x-axis corresponds to the value of some variable, and the y-axis tells us something about how likely we are to observe that value. However, notice that the y-axis is labelled "Probability Density" and not "Probability". There is a subtle and somewhat frustrating characteristic of continuous distributions that makes the y axis behave a bit oddly: the height of the curve here isn't actually the probability of observing a particular x value. On the other hand, it is true that the heights of the curve tells you which x values are more likely (the higher ones!). (see Section 6.5.1 for all the annoying details)

The normal distribution

While the binomial distribution is conceptually the simplest distribution to understand, it's not the most important one. That particular honour goes to the normal distribution, also referred to as "the bell curve" or a "Gaussian distribution". A normal distribution is described using two parameters: the mean of the distribution μ and the standard deviation of the distribution σ.

The notation that we sometimes use to say that a variable X is normally distributed is as follows:

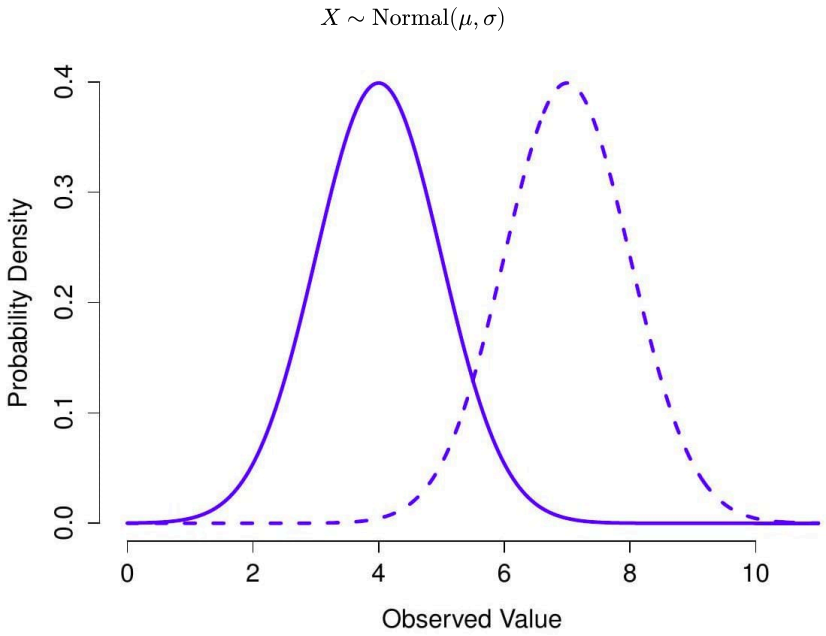


Figure 6.1.6: An illustration of what happens when you change the mean of a normal distribution. The solid line depicts a normal distribution with a mean of $\mu = 4$. The dashed line shows a normal distribution with a mean of $\mu = 7$. In both cases, the standard deviation is $\sigma = 1$. Not surprisingly, the two distributions have the same shape, but the dashed line is shifted to the right.

Of course, that's just notation. It doesn't tell us anything interesting about the normal distribution itself. As was the case with the binomial distribution, I have included the formula for the normal distribution in this book, because I think it's important enough that everyone who learns statistics should at least look at it, but since this is an introductory text I don't want to focus on it, so I've tucked it away in Table 6.2.

Instead of focusing on the maths, let's try to get a sense for what it means for a variable to be normally distributed. To that end, have a look at Figure 6.5 which plots a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. You can see where the name "bell curve" comes from; it looks a bit like a bell. Notice that, unlike the plots that I drew to illustrate the binomial distribution, the picture of the normal distribution in Figure 6.5 shows a smooth curve instead of "histogram-like" bars. This isn't an arbitrary choice, the normal distribution is continuous whereas the binomial is discrete. For instance, in the die rolling example from the last section it was possible to get 3 skulls or 4 skulls, but impossible to get 3.9 skulls. The figures that I drew in the previous section reflected this fact. In Figure 6.3, for instance, there's a bar located at $X = 3$ and another one at $X = 4$ but there's nothing in between. Continuous quantities don't have this constraint. For instance, suppose we're talking about the weather. The temperature on a pleasant Spring day could be 23 degrees, 24 degrees, 23.9 degrees, or anything in between since temperature is a continuous variable. And so a normal distribution might be quite appropriate for describing Spring temperatures. [4]

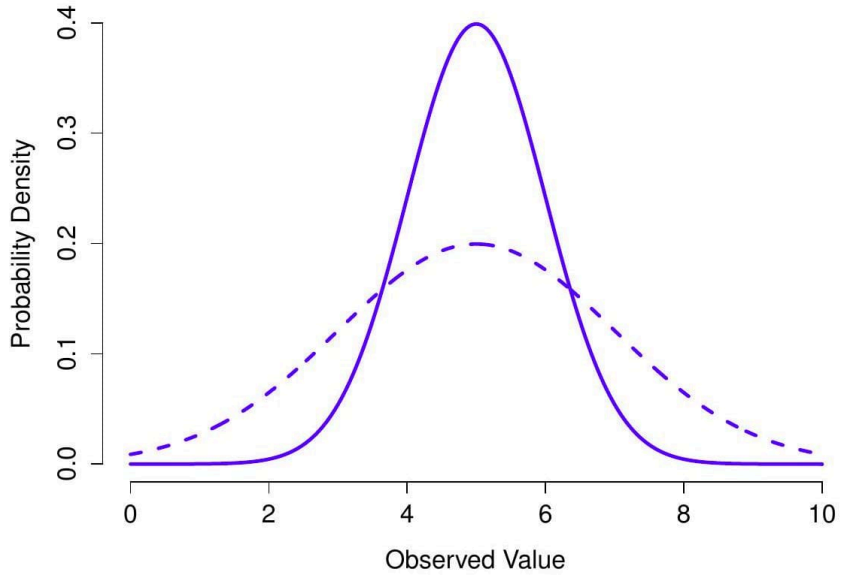


Figure $6.1.7$: Figure 6.7: An illustration of what happens when you change the the standard deviation of a normal distribution. Both distributions plotted in this figure have a mean of $\mu = 5$, but they have different standard deviations. The solid line plots a distribution with standard deviation $\sigma = 1$, and the dashed line shows a distribution with standard deviation $\sigma = 2$. As a consequence, both distributions are "centred" on the same spot, but the dashed line is wider than the solid one.

With this in mind, let's see if we can't get an intuition for how the normal distribution works. First, let's have a look at what happens when we play around with the parameters of the distribution. To that end, Figure 6.6 plots normal distributions that have different means but have the same standard deviation. As you might expect, all of these distributions have the same "width". The only difference between them is that they've been shifted to the left or to the right. In every other respect they're identical. In contrast, if we increase the standard deviation while keeping the mean constant, the peak of the distribution stays in the same place but the distribution gets wider, as you can see in Figure 6.7. Notice, though, that when we widen the distribution the height of the peak shrinks. This has to happen, in the same way that the heights of the bars that we used to draw a discrete binomial distribution have to sum to 1, the total area under the curve for the normal distribution must equal 1. Before moving on, I want to point out one important characteristic of the normal distribution. Irrespective of what the actual mean and standard deviation are, 68.3\% of the area falls within 1 standard deviation of the mean. Similarly, 95.4\% of the distribution falls within 2 standard deviations of the mean, and 99.7\% of the distribution is within 3 standard deviations. This idea is illustrated in Figure 6.8.

6.5.1 Probability density

There's something I've been trying to hide throughout my discussion of the normal distribution, something that some introductory textbooks omit completely. They might be right to do so. This "thing" that I'm hiding is weird and counter-intuitive even by the admittedly distorted standards that apply in statistics. Fortunately, it's not something that you need to understand at a deep level in order to do basic statistics. Rather, it's something that starts to become important later on when you move beyond the basics. So, if it doesn't make complete sense, don't worry too much, but try to make sure that you follow the gist of it.

Throughout my discussion of the normal distribution there's been one or two things that don't quite make sense. Perhaps you noticed that the y-axis in these figures is labelled "Probability Density" rather than probability. Maybe you noticed that I used $p(X)$ instead of $P(X)$ when giving the formula for the normal distribution.

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

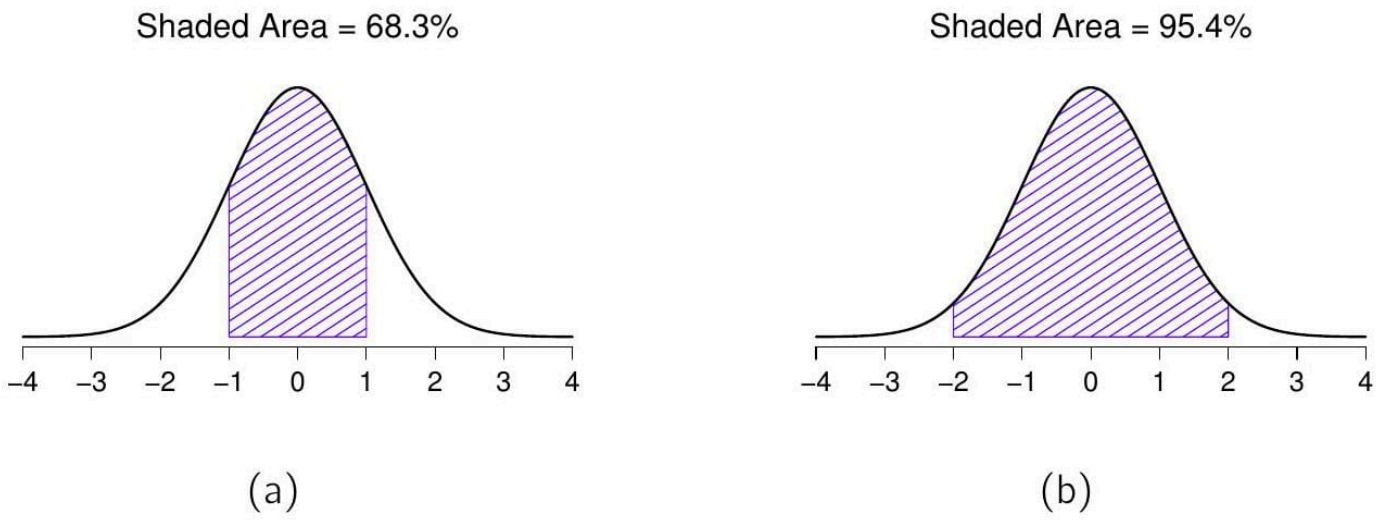


Figure 6.1.8: Figure 6.8: The area under the curve tells you the probability that an observation falls within a particular range. The solid lines plot normal distributions with mean $\mu = 0$ and standard deviation $\sigma = 1$. The shaded areas illustrate "areas under the curve" for two important cases. In panel a, we can see that there is a 68.3\% chance that an observation will fall within one standard deviation of the mean. In panel b, we see that there is a 95.4\% chance that an observation will fall within two standard deviations of the mean.

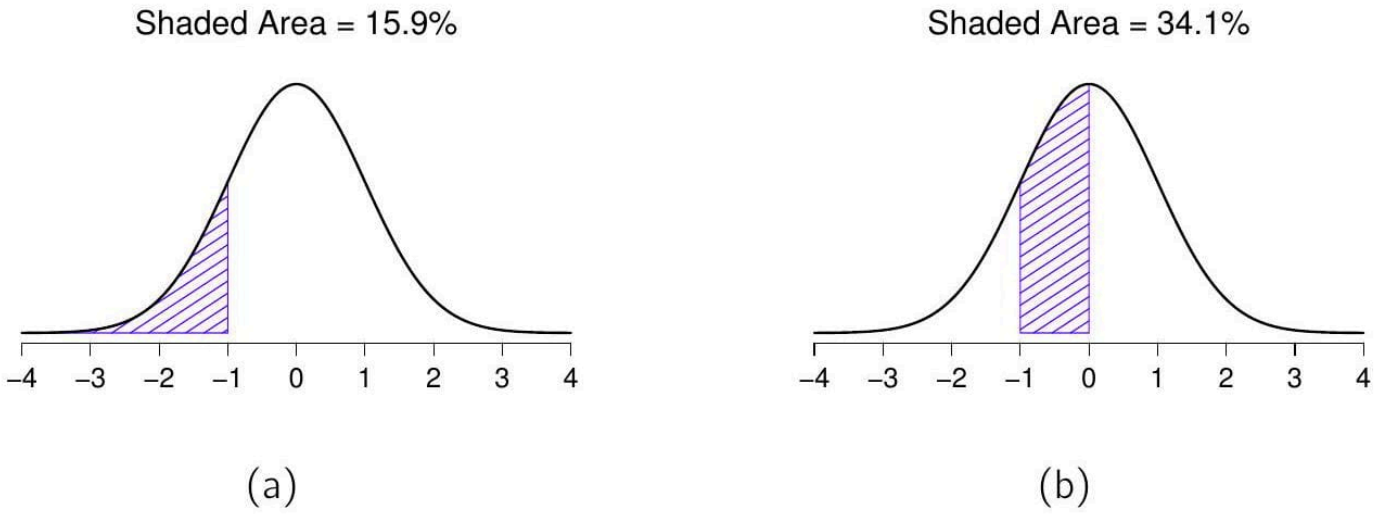


Figure 6.1.9: Two more examples of the "area under the curve idea". There is a 15.9% chance that an observation is one standard deviation below the mean or smaller (panel a), and a 34.1% chance that the observation is somewhere between one standard deviation below the mean and the mean (panel b). Notice that if you add these two numbers together you get $15.9\% + 34.1\% = 50\%$. For normally distributed data, there is a 50\% chance that an observation falls below the mean. And of course that also implies that there is a 50% chance that it falls above the mean.

Now think about what this implies when we talk about probabilities. Suppose that tomorrow's maximum temperature is sampled from a normal distribution with mean 23 and standard deviation 1. What's the probability that the temperature will be exactly 23 degrees? The answer is "zero", or possibly "a number so close to zero that it might as well be zero". Why is this? It's like trying to throw a dart at an infinitely small dart board. No matter how good your aim, you'll never hit it. In real life you'll never get a value of exactly 23. It'll always be something like 23.1 or 22.99998 or suchlike. In other words, it's completely meaningless to talk about the probability that the temperature is exactly 23 degrees. However, in everyday language if I told you that it was 23 degrees outside and it turned out to be 22.9998 degrees you probably wouldn't call me a liar. Because in everyday language "23 degrees" usually means something like "somewhere between 22.5 and 23.5 degrees". And while it doesn't feel very meaningful to ask about the probability that the temperature is exactly 23 degrees, it does seem sensible to ask about the probability that the temperature lies between 22.5 and 23.5, or between 20 and 30 , or any other range of temperatures.

The point of this discussion is to make clear that when we're talking about continuous distributions it's not meaningful to talk about the probability of a specific value. However, what we can talk about is the probability that the value lies within a particular

range of values. To find out the probability associated with a particular range what you need to do is calculate the "area under the curve". We've seen this concept already, in Figure 6.8 the shaded areas shown depict genuine probabilities (e.g., in Figure 6.8 a it shows the probability of observing a value that falls within 1 standard deviation of the mean).

Okay, so that explains part of the story. I've explained a little bit about how continuous probability distributions should be interpreted (i.e., area under the curve is the key thing). But what does the formula for $p(x)$ that I described earlier actually mean? Obviously, $p(x)$ doesn't describe a probability, but what is it? The name for this quantity $p(x)$ is a probability density, and in terms of the plots we've been drawing it corresponds to the height of the curve. The densities themselves aren't meaningful in and of themselves, but they're "rigged" to ensure that the area under the curve is always interpretable as genuine probabilities. To be honest, that's about as much as you really need to know for now. [5]

Other useful distributions

The normal distribution is the distribution that statistics makes most use of (for reasons to be discussed shortly), and the binomial distribution is a very useful one for lots of purposes. But the world of statistics is filled with probability distributions, some of which we'll run into in passing. In particular, the three that will appear in this book are the t distribution, the χ^2 distribution and the F distribution.

- The t distribution is a continuous distribution that looks very similar to a normal distribution. The "tails" of the t distribution are "heavier" (i.e., extend further outwards) than the tails of the normal distribution). That's the important difference between the two. This distribution tends to arise in situations where you think that the data actually follow a normal distribution, but you don't know the mean or standard deviation.
- The χ^2 distribution is another distribution that turns up in lots of different places. The situation in which we'll see it is when doing categorical data analysis, but it's one of those things that actually pops up all over the place. When you dig into the maths (and who doesn't love doing that?), it turns out that the main reason why the χ^2 distribution turns up all over the place is that if you have a set of independent standard normal variables, square their values and then add them up (a procedure referred to as taking a "sum of squares"), this sum has a χ^2 distribution. You'd be amazed how often this fact turns out to be useful.
- The F distribution looks a bit like a χ^2 distribution, and it arises whenever you need to compare two χ^2 distributions to one another. Admittedly, this doesn't exactly sound like something that any sane person would want to do, but it turns out to be very important in real world data analysis. Remember when I said that χ^2 turns out to be the key distribution when we're taking a "sum of squares"? Well, what that means is if you want to compare two different "sums of squares", you're probably talking about something that has an F distribution.

  Okay, time to wrap this section up. We've seen three new distributions: χ^2, t and F. They're all continuous distributions, and they're all closely related to the normal distribution. The main thing for our purposes is that you grasp the basic idea that these distributions are all deeply related to one another, and to the normal distribution. Later on in this book we're going to run into data that are normally distributed, or at least assumed to be normally distributed. What I want you to understand right now is that, if you make the assumption that your data are normally distributed, you shouldn't be surprised to see χ^2, t and F distributions popping up all over the place when you start trying to do your data analysis.

Summary

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

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

As you'd expect, my coverage is by no means exhaustive. Probability theory is a large branch of mathematics in its own right, entirely separate from its application to statistics and data analysis. As such, there are thousands of books written on the subject and universities generally o!er multiple classes devoted entirely to probability theory. Even the "simpler" task of documenting

standard probability distributions is a big topic. I've described five standard probability distributions in this chapter, but sitting on my bookshelf I have a 45-chapter book called "Statistical Distributions" (Evans, Hastings, and Peacock 2011) that lists a lot more than that. Fortunately for you, very little of this is necessary. You're unlikely to need to know dozens of statistical distributions when you go out and do real world data analysis, and you definitely won't need them for this book, but it never hurts to know that there's other possibilities out there.

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

[1] This doesn't mean that frequentists can't make hypothetical statements, of course. It's just that if you want to make a statement about probability then it must be possible to redescribe that statement in terms of a sequence of potentially observable events, together with the relative frequencies of different outcomes that appear within that sequence.

[2] The term "binomial" comes in because it is related to the coefficients you get when expanding the general "binomial" expression $(a+b)^n$.

[3] Note that the term "success" is pretty arbitrary and doesn't actually imply that the outcome is something to be desired. If θ referred to the probability that any one passenger gets injured in a bus crash I'd still call it the success probability, but that doesn't mean I want people to get hurt in bus crashes!

[4] In practice, the normal distribution is so handy that people tend to use it even when the variable isn't actually continuous. As long as there are enough categories (e.g., Likert scale responses to a questionnaire), it's pretty standard practice to use the normal distribution as an approximation. This works out much better in practice than you'd think.

[5] For those readers who know a little calculus, I'll give a slightly more precise explanation. In the same way that probabilities are non-negative numbers that must sum to 1 , probability densities are non-negative numbers that must integrate to 1 (where the integral is taken across all possible values of X). To calculate the probability that X falls between a and b we calculate the definite integral of the density function over the corresponding range, $\int_a^b p(x)dx$. If you don't remember or never learned calculus, don't worry about this. It's not needed for this book.

This page titled 6.1: Introduction to probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 6: Introduction to probability by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

7: SAMPLING AND ESTIMATION

7.1: Estimating unknown quantities from a sample

This page titled 7: Sampling and Estimation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

7.1: Estimating unknown quantities from a sample

As we've learned, 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: first, we focus on sampling theory, then we make use of sampling theory to discuss how statisticians think about estimation.

7.1 Samples, populations and sampling

All learning requires you to make assumptions. Accepting that this is true, our first task is 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. We might have run experiment with some number of participants, a polling company might have phoned some number of people to ask questions about voting intentions, and so on. In this way
the data set available to us is finite and incomplete. We can't possibly get every person in the world to do our experiment, for example a polling company doesn't have the time or the money to ring up every voter in the country. In our earlier discussion of descriptive statistics this sample was the only thing we were interested in. Our only goal was to find ways of describing, summarising and graphing that sample. This is about to change.

7.1.1 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.

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

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

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

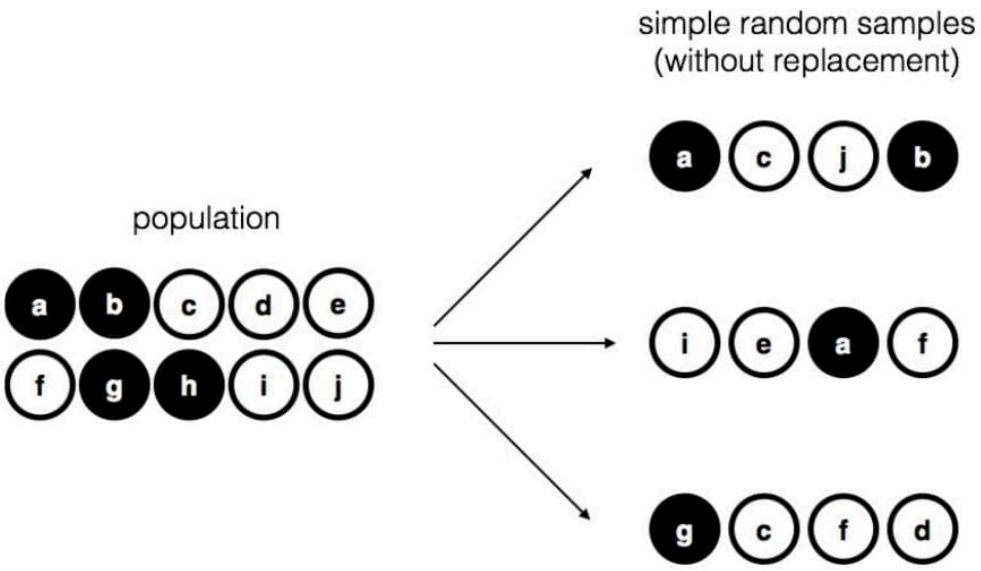


Figure 7.1: Simple random sampling without replacement from a finite population

7.1.2 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 7.1. 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 7.1. 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 time we refer to as a random process. [1] 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.

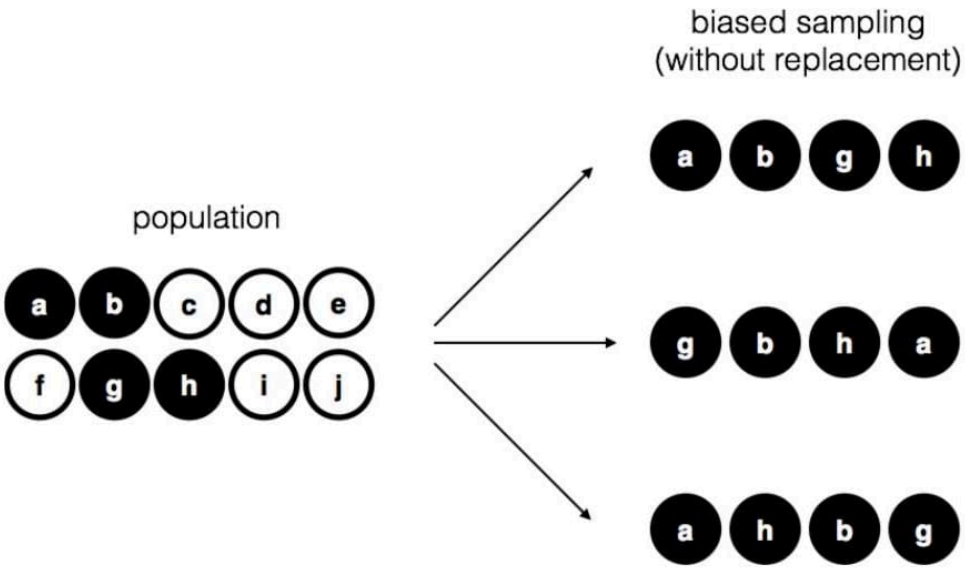


Figure 7.2: Biased sampling without replacement from a finite population

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 7.2. 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 in Figure 7.3.

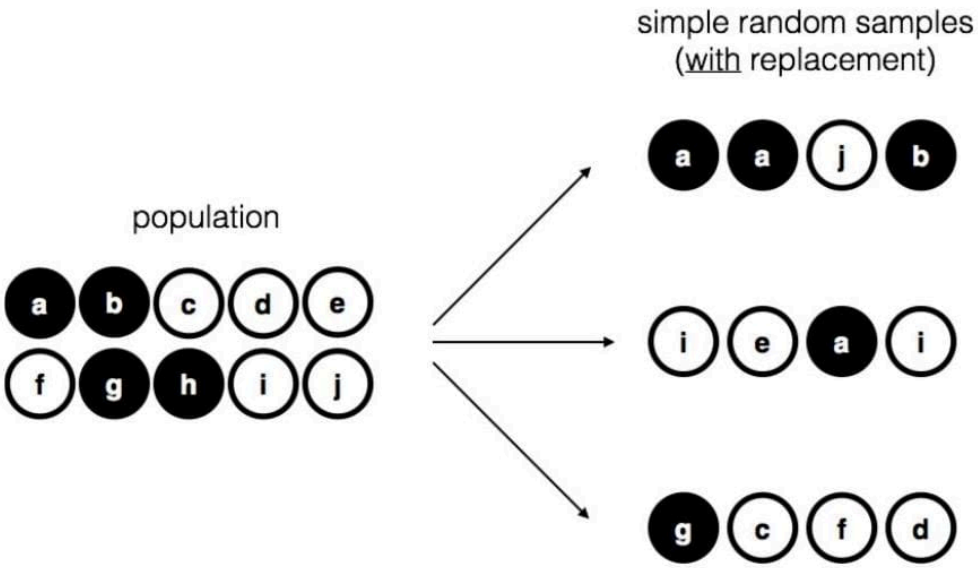


Figure 7.3: Simple random sampling with replacement from a finite population

In my experience, most psychology experiments tend to be sampling without replacement, because the same person is not allowed to participate in the experiment twice. However, most statistical theory is based on the assumption that the data arise from a simple random sample with replacement. In real life this very rarely matters. If the sample is a small fraction of the population, the difference between sampling with and without replacement is usually 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.

7.1.3 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. When I run experiments I'd consider it a minor miracle if my participants turned out to be a random sampling of the undergraduate psychology students at Adelaide university, even though this is by far the narrowest population that I might want to generalise to. A thorough discussion of other types of sampling schemes is beyond the scope of this book, but to give you a sense of what's out there I'll list a few of the more important ones.

- Stratified sampling. Suppose your population is (or can be) divided into several different subpopulations, 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 [2] strata (schizophrenic and not-schizophrenic) and then sample an equal number of people from each group. If you selected people randomly you would get so few schizophrenic people in the sample that your study would be useless. This specific kind of of stratified sampling is referred to as oversampling because it makes a deliberate attempt to over-represent rare groups.

- Snowball sampling is a technique that is especially useful when sampling from a "hidden" or hard to access population and is especially common in social sciences. For instance, suppose the researchers want to conduct an opinion poll among transgender people. The research team might only have contact details for a few trans folks, so the survey starts by asking them to participate (stage 1). At the end of the survey the participants are asked to provide contact details for other people who might want to participate. In stage 2 those new contacts are surveyed. The process continues until the researchers have sufficient data. The big advantage to snowball sampling is that it gets you data in situations that might otherwise be impossible to get any. On the statistical side, the main disadvantage is that the sample is highly non-random, and non-random in ways that are difficult to address. On the real life side, the disadvantage is that the procedure can be unethical if not handled well, because hidden populations are often hidden for a reason. I chose transgender people as an example here to highlight this 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 psychology are studies that rely on undergraduate psychology students. These samples are generally non-random in two respects. First, reliance on undergraduate psychology 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 psychology 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.

7.1.4 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 7.1 and 7.2. 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 psychologically relevant phenomenon of interest. Suppose I'm doing a study looking at working memory capacity. In study 1, I actually have the ability to sample randomly from all human beings currently alive, with one exception: I can only sample people born on a Monday. In study 2, I am able to sample randomly from the Australian population. I want to generalise my results to the population of all living humans. Which study is better? The answer, obviously, is study 1. Why? Because we have no reason to think that being "born on a Monday" has any interesting relationship to working memory capacity. In contrast, I can think of several reasons why "being Australian" might matter. Australia is a wealthy, industrialised country with a very well-developed education system. People growing up in that system will have had life experiences much more similar to the experiences of the people who designed the tests for working memory capacity. This shared experience might easily translate into similar beliefs about how to "take a test", a shared assumption about how psychological experimentation works, and so on. These things might actually matter. For instance, "test taking" style might have taught the Australian participants how to direct their attention exclusively on fairly abstract test materials much more than people who haven't grown up in a similar environment. This could therefore lead to a misleading picture of what working memory capacity is.

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., psychology lecturers sample psychology students because that's the least expensive way to collect data, and our coffers aren't exactly overflowing with gold), but if so you should at least spend some time thinking about what the dangers of this practice might be. 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.

7.1.5 Population parameters and sample statistics

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

On the one hand, they are interested in real world data and real science in the same way that scientists are. On the other hand, they also operate in the realm of pure abstraction in the way that mathematicians do. As a consequence, statistical theory tends to be a bit abstract in how a population is defined. In much the same way that psychological researchers operationalise our abstract theoretical ideas in terms of concrete measurements, statisticians operationalise the concept of a "population" in terms of mathematical objects that they know how to work with. They're called probability distributions.

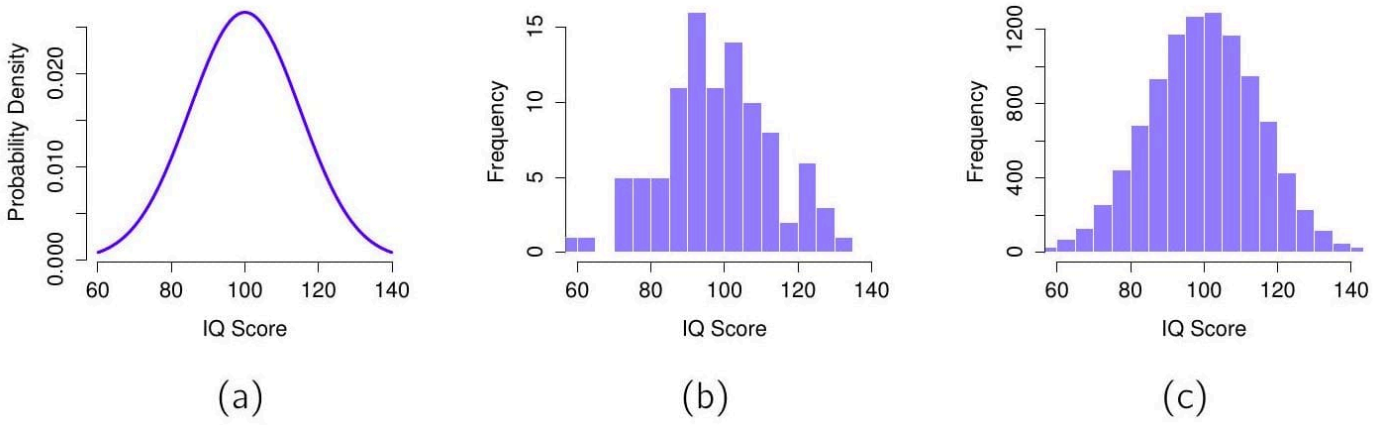


Figure 7.4: The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and panel c we have a sample of 10,000 observations.

The idea is quite simple. Let's say we're talking about IQ scores. To a psychologist the population of interest is a group of actual

humans who have IQ scores. A statistician "simplifies" this by operationally defining the population as the probability distribution depicted in Figure 7.4a. 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 μ is 100 and the population standard deviation σ 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 7.4b. 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 7.4a. 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 and how to work out how confident you are in your estimates but before we get to that there's a few more ideas in sampling theory that you need to know about.

7.2 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 $\mathrm{sd} = 15$. 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 7.5

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 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.

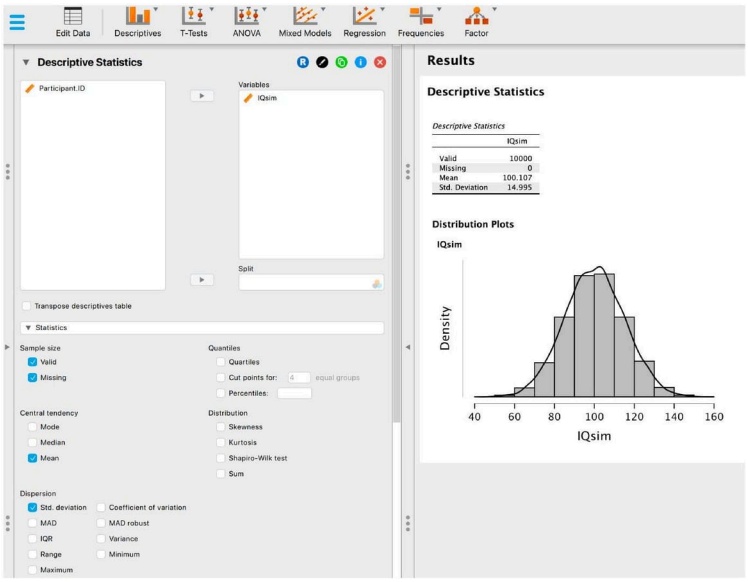


Figure 7.5: A random sample drawn from a normal distribution using JASP

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 ($\bar{X} \rightarrow \mu$). [3]

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.
7.3

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)

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

7.3.1 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

7.6). These are the five numbers that JASP generated for me (yours will be different!). I rounded to the nearest whole number for convenience:

```
01 124 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:

```
01 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 7.1, and as you can see the sample mean varies from one replication to the next.

Table 7.1: 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

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 7.1, so my data set starts out like this:

96.0, 107.0, 101.6, 103.8, 104.4, ...

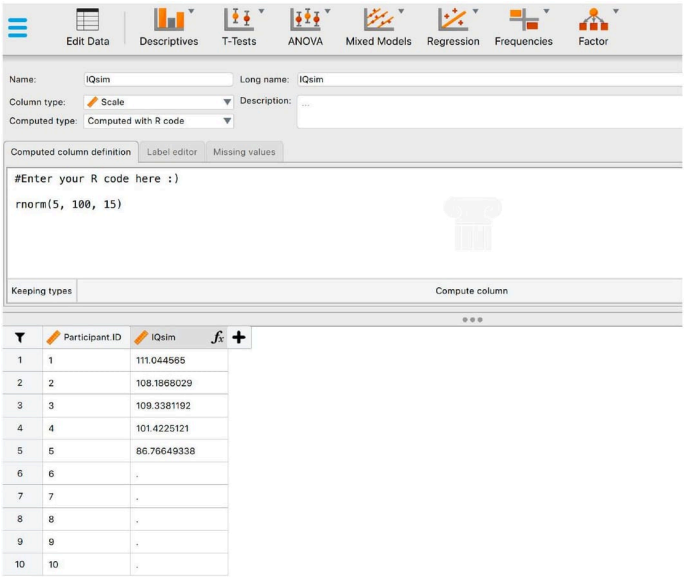


Figure 7.6: Using JASP to draw a random sample of 5 from a normal distribution with $\mu = 100$ and $\sigma = 15$.

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 7.7. As this picture illustrates, 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 7.7 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.

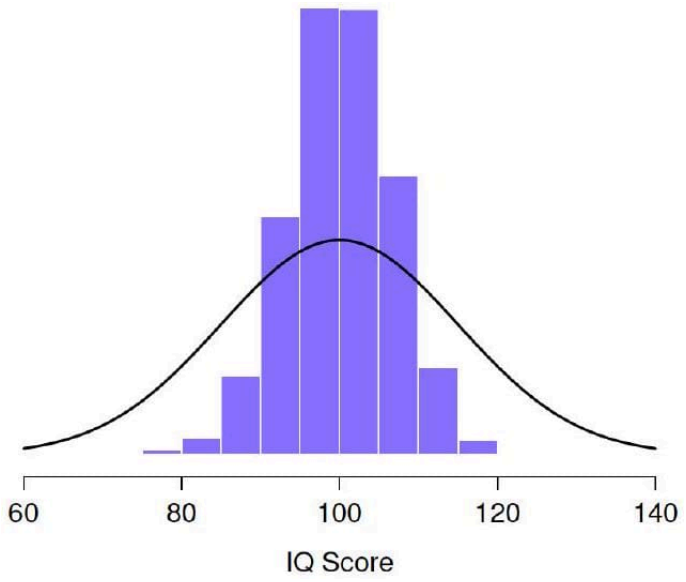


Figure 7.7: The sampling distribution of the mean for the "five IQ scores experiment". If you sample 5 people at random and calculate their average IQ you'll almost certainly get a number between 80 and 120 , even though there are quite a lot of individuals who have IQs above 120 or below 80 . For comparison, the black line plots the population distribution of IQ scores.

7.3.2 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, ...

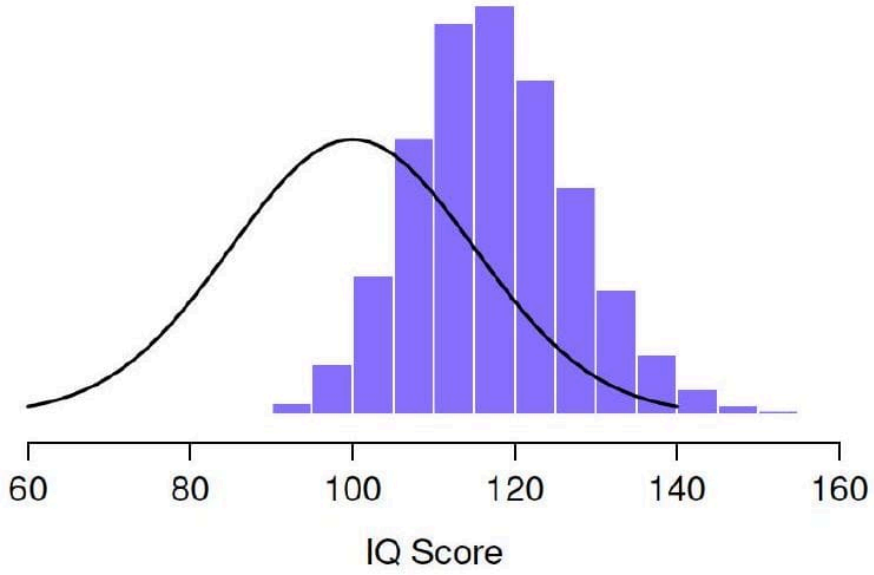


Figure 7.8: The sampling distribution of the maximum for the "five IQ scores experiment". If you sample 5 people at random and select the one with the highest IQ score you'll probably see someone with an IQ between 100 and 140.

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 7.8. 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.

7.3.3 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 7.9, 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.

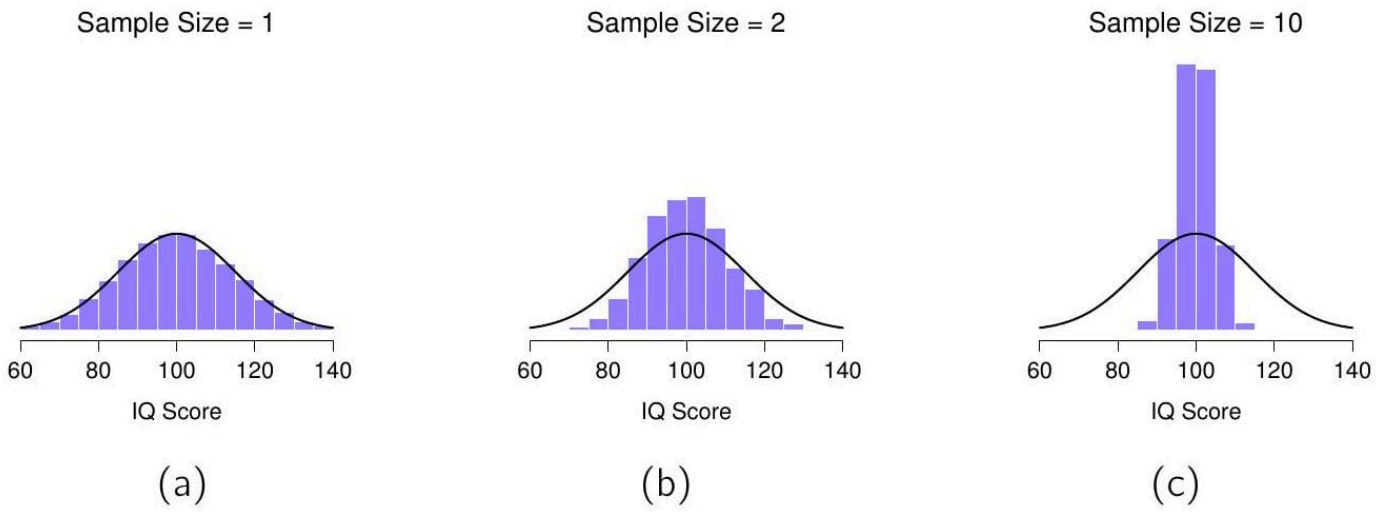


Figure 7.9: An illustration of the how sampling distribution of the mean depends on sample size. In each panel I generated 10,000 samples of IQ data and calculated the mean IQ observed within each of these data sets. The histograms in these plots show the distribution of these means (i.e., the sampling 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.

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: when observations are independent and drawn from the same population with a finite mean and variance, the sampling distribution of the mean becomes approximately normal as the sample size increases. 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 7.10. 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 7.10b plots the histogram of these sample means (i.e., the sampling distribution of the mean for $N = 2$). This time, the histogram produces a n-shaped distribution. It's still not normal, but it's a lot closer to the black line than the population distribution in Figure 7.10a. When I increase the sample size to $N = 4$, the sampling distribution of the mean is very close to normal (Figure 7.10c), and by the time we reach a sample size of $N = 8$ it's almost perfectly normal. In other words, under these conditions, the sampling distribution of the mean becomes approximately normal when the sample is sufficiently large. How large it needs to be depends on the population distribution.

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

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

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

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

Because we divide the population standard deviation σ by the square root of the sample size N, the SEM gets smaller as the

sample size increases. It also tells us that the shape of the sampling distribution becomes normal. [4]

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.

7.4 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).

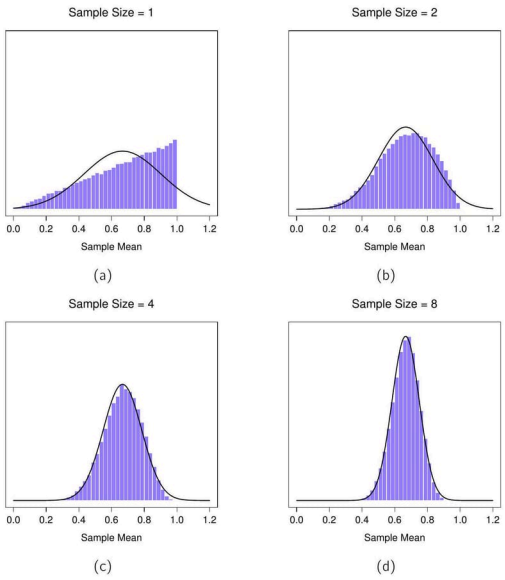


Figure 7.10: A demonstration of the central limit theorem. In panel a, we have a non-normal population distribution, and panels b-d show the sampling distribution of the mean for samples of size 2,4 and 8 for data drawn from the distribution in panel a. As you can see, even though the original population 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.

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, suppose you wanted to measure the effect of low level lead poisoning on cognitive functioning in Port Pirie, a South Australian industrial town with a lead smelter. Perhaps you decide that you want to compare IQ scores among people in Port Pirie to a comparable sample in Whyalla, a South Australian industrial town with a steel refinery. [5] 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 South Australian industrial towns. We're going to have to estimate the population parameters from a sample of data. So how do we do this?

7.4.1 Estimating the population mean

Suppose we go to Port Pirie 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 poulation 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 different notation to refer to them. For instance, if the true population mean is denoted μ, then we would use $\hat{\mu}$ to refer to our estimate of the population mean. In contrast, the sample mean is denoted $\bar{X}$ or sometimes m. However, in simple random samples the estimate of the population mean is identical to the sample mean. If I observe a sample mean of $\bar{X} = 98.5$ then my estimate of the population mean is also $\hat{\mu} = 98.5$. To help keep the notation clear, here's a handy table:

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

7.4.2 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 identical to the corresponding sample statistic (i.e. $\bar{X}$). However, that's not always true. To see this, let's have a think about how to construct an estimate of the population standard deviation, which we'll denote $\hat{\sigma}$. What shall we use as our estimate in this case? Your first thought might be that we could do the same thing we did when estimating the mean, and just use the sample statistic as our estimate. That's almost the right thing to do, but not quite.

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

20

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

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

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

20, 22

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

guess we can possibly make is the sample mean. If forced to guess we'd probably guess that the population mean cromulence is 21. What about the standard deviation? This is a little more complicated. The sample standard deviation is only based on two observations, and if you're at all like me you probably have the intuition that, with only two observations we haven't given the population "enough of a chance" to reveal its true variability to us. It's not just that we suspect that the estimate is wrong, after all with only two observations we expect it to be wrong to some degree. The worry is that the error is systematic. 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 7.11. 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 7.9b 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 7.12. 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. [6]

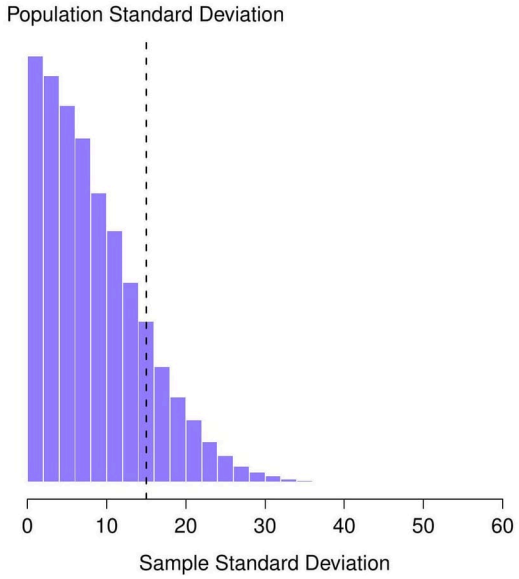


Figure 7.11: The sampling distribution of the sample standard deviation for a "two IQ scores" experiment. The true population standard deviation is 15 (dashed line), but as you can see from the histogram the vast majority of experiments will produce a much smaller sample standard deviation than this. 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.

The plot on the right is quite different: on average, the sample standard deviation s is smaller than the population standard deviation σ. It is a biased estimator. In other words, if we want to make a "best guess" $\hat{\sigma}$ about the value of the population standard deviation σ we should make sure our guess is a little bit larger than the sample standard deviation s.

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

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

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

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

This is an unbiased estimator of the population variance σ. Moreover, this finally answers the question - 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:

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

and when we use JASP's built in standard deviation function, what it's doing is calculating $\hat{\sigma}$, not s. [a]

[a] Okay, I'm hiding something else here. In a bizarre and counter-intuitive twist, since $\hat{\sigma}^2$ is an unbiased estimator of σ^2, you'd assume that taking the square root would be fine and $\hat{\sigma}$ would be an unbiased estimator of σ. 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 σ^2, but when you take the square root, it turns out that $\hat{\sigma}$ is a biased estimator of the population standard deviation σ. 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.This is a common convention. In this book, however, we use “sample standard deviation” for the formula where we divide by N. These aren't the same thing, either conceptually or numerically. One is a property of the sample, the other is an estimated characteristic of the population. However, in almost every real life application what we actually care about is the estimate of the population parameter, and so people always report $\hat{\sigma}$ rather than s. 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.

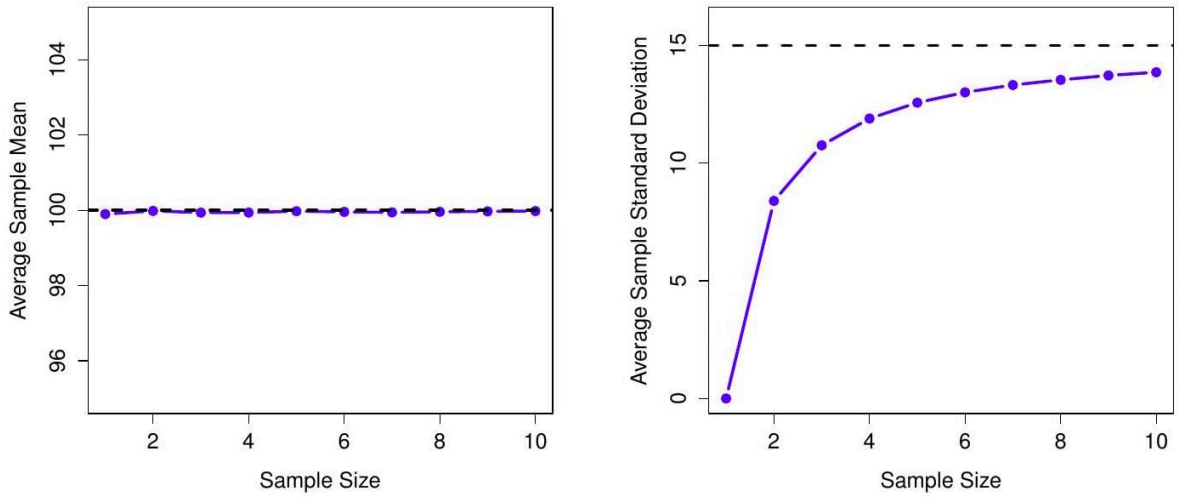


Figure 7.12: An illustration of the fact that the sample mean is an unbiased estimator of the population mean (panel a), but the sample standard deviation is a biased estimator of the population standard deviation (panel b). For the figure I generated 10,000 simulated data sets with 1 observation each, 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.

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
σ	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
σ^2	Population variance	Almost never known for sure
$\hat{\sigma}^2$	Estimate of the population variance	Yes, but not the same as the sample variance

7.5 Estimating a confidence interval

Statistics means never having to say you're certain

- Unknown origin [7]

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 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 μ and the standard deviation is σ. I've just finished running my study that has N participants, and the mean IQ among those participants is $\bar{X}$. We know from our discussion of the central limit theorem that the sampling distribution of the mean is approximately normal. We also know from our discussion of the normal distribution that there is a 95% chance that a normally-distributed quantity will fall within 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:

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

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:

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

What this is telling us is that the range of values has a 95% probability of containing the population mean μ. We refer to this range as a 95% confidence interval, denoted Cl_{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:

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

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.

7.5.1 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 σ. Yet, in Section 7.4 I stressed the fact that we don't actually know the true population parameters. Because we don't know the true value of σ we have to use an estimate of the population standard deviation $\hat{\sigma}$ instead. This is pretty straightforward to do, but this has the consequence that we need to use the 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 μ 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.

7.5.2 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 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 7.13, 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.

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.

7.5.3 Calculating confidence intervals in JASP

JASP can calculate confidence intervals for the mean directly. Under Descriptives, move your variable into the Variables box, expand Statistics, and select Confidence interval for mean. Use a width of 95% and the T model method. You can also calculate an approximate interval by hand using the mean and its standard error:

```
Mean - (1.96 * S.E. Mean), and the upper 95% confidence interval as:
Mean + (1.96 * 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:

$$\text{Lower } 95\%\text{CI} = 100.107 - (1.96 * 0.150) = 99.813$$
$$\text{Upper } 95\%\text{CI} = 100.107 + (1.96 * 0.150) = 100.401$$

So, in our simulated large sample data with N = 10,000, the mean IQ score is 100.107 with an approximate 95% CI from 99.813 to 100.401.

JASP can also display the mean and its 95% confidence interval using an interval plot under Descriptives.

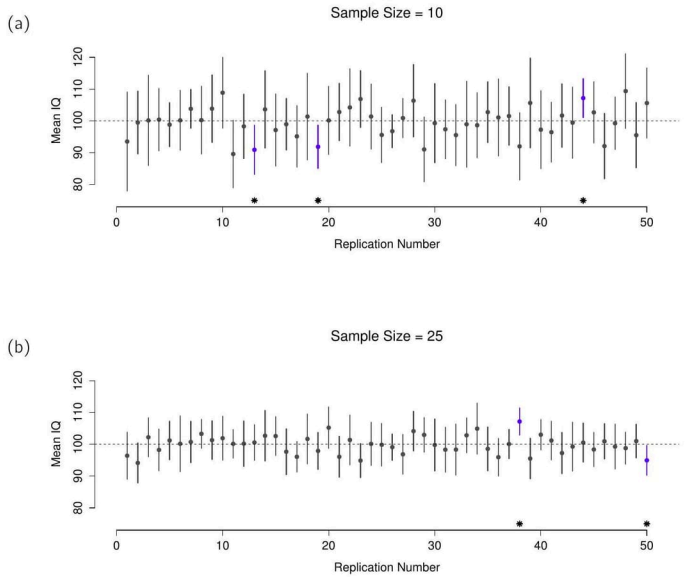


Figure 7.13: 95% confidence intervals. The top (panel a) shows 50 simulated replications of an experiment in which we measure the IQs of 10 people. The dot marks the location of the sample mean and the line shows the 95% confidence interval. In total 47 of the 50 confidence intervals do contain the true 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.

analysis. That's pretty cool, so we'll show you how to do that later on.

7.6 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
- -Statistical theory of sampling: the law of large numbers, sampling distributions and the central limit theorem
- -Estimating means and standard deviations
- -Estimating a confidence interval

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

---

1. [1] 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.
2. [2] 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 psychology text so please forgive me a few simplifications here and there.
3. [3] 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.
4. [4] 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.
5. [5] 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 in Whyalla to Port Pirie and assume that any differences are due to lead poisoning. Even if it were true that the only differences between the two towns corresponded to the different refineries (and it isn't, not by a long shot), you need to account for the fact that people already believe that lead pollution causes cognitive deficits. This means that there are different demand effects for the Port Pirie sample than for the Whyalla sample. 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 Port Pirie 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. Other people in Port Pirie might be more motivated to do well because they don't want their home town to look bad. The motivational effects that would apply in Whyalla are likely to be weaker, because people don't have any concept of "iron ore poisoning" in the same way that they have a concept for "lead poisoning". Psychology is hard.
6. [6] 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.
7. [7] 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.

This page titled 7.1: Estimating unknown quantities from a sample is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 7: Estimating unknown quantities from a sample by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

8: HYPOTHESIS TESTING

8.1: Hypothesis testing

This page titled 8: Hypothesis Testing is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

8.1: 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 [1]

Estimation 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. [2] Afterwards, I'll spend a bit of time talking about the various dogmas, rules and heresies that surround the theory of hypothesis testing.

8.1 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 psychological research, the search for extrasensory perception (ESP). [3]

Let's suppose that this glorious day has come. My first study is a simple one in which I seek to test whether clairvoyance exists. Each participant sits down at a table and is shown a card by an experimenter. The card is black on one side and white on the other. The experimenter takes the card away and places it on a table in an adjacent room. The card is placed black side up or white side up completely at random, with the randomisation occurring only after the experimenter has left the room with the participant. A second experimenter comes in and asks the participant which side of the card is now facing upwards. It's purely a one-shot experiment. Each person sees only one card and gives only one answer, and at no stage is the participant actually in contact with someone who knows the right answer. My data set, therefore, is very simple. I have asked the question of N people and some number X of these people have given the correct response. To make things concrete, let's suppose that I have tested $N = 100$ people and $X = 62$ of these got the answer right. A surprisingly large number, sure, but is it large enough for me to feel safe in claiming I've found evidence for ESP? This is the situation where hypothesis testing comes in useful. However, before we talk about how to test hypotheses, we need to be clear about what we mean by hypotheses.

8.1.1 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 ESP study my overall scientific goal is to demonstrate that clairvoyance exists. In this situation I have a clear research goal: I am hoping to discover evidence for ESP. 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 clairvoyance exists. 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. If you are a psychologist then your research hypotheses are fundamentally about psychological constructs. Any of the following would count as research hypotheses:

- Listening to music reduces your ability to pay attention to other things. This is a claim about the causal relationship between two psychologically meaningful concepts (listening to music and paying attention to things), so it's a perfectly reasonable research hypothesis.

- Intelligence is related to personality. Like the last one, this is a relational claim about two psychological constructs (intelligence and personality), but the claim is weaker: correlational not causal.
- Intelligence is speed of information processing. This hypothesis has a quite different character. It's not actually a relational claim at all. It's an ontological claim about the fundamental character of intelligence (and I'm pretty sure it's wrong). It's worth expanding on this one actually. It's usually easier to think about how to construct experiments to test research hypotheses of the form "does X affect Y ?" than it is to address claims like "what is X ?" And in practice what usually happens is that

you find ways of testing relational claims that follow from your ontological ones. For instance, if I believe that intelligence is speed of information processing in the brain, my experiments will often involve looking for relationships between measures of intelligence and measures of speed. As a consequence most everyday research questions do tend to be relational in nature, but they're almost always motivated by deeper ontological questions about the state of nature.

Notice that in practice, my research hypotheses could overlap a lot. My ultimate goal in the ESP experiment might be to test an ontological claim like "ESP exists", but I might operationally restrict myself to a narrower hypothesis like "Some people can 'see' objects in a clairvoyant fashion". That said, there are some things that really don't count as proper research hypotheses in any meaningful sense:

- Love is a battlefield. 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 people in my experiment will say "yes" than "no". This one fails as a research hypothesis because it's a claim about the data set, not about the psychology (unless of course your actual research question is whether people have some kind of "yes" bias!). 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 ESP study my research hypothesis is that some people are able to see through walls 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 ("correct"), the true-but-unknown probability with which the participants in my experiment answer the question correctly. Let's use the Greek letter θ (theta) to refer to this probability. Here are four different statistical hypotheses:

- If ESP doesn't exist 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 ESP does exist and participants can see the card. If that's true people will perform better than chance and the statistical hypothesis is that $\theta > 0.5$.
- A third possibility is that ESP does exist, but the colours 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 ESP exists but I have no idea whether people are seeing the right colour 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 (a claim about psychology), and this then corresponds to a statistical hypothesis (a claim about the data generating population). In my ESP example these might be:

$$\begin{array}{ll} \text{Dani's research hypothesis:} & \text{"ESP exists"} \\ \text{Dani's statistical hypothesis:} & \theta \neq 0.5 \end{array}$$

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 ESP study was conducted in a situation where the participant can actually see the card reflected in a window. 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

"ESP exists".

8.1.2 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 ESP example, the null hypothesis is that $\theta = 0.5$, since that's what we'd expect if ESP didn't exist. My hope, of course, is that ESP 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 θ 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 determine whether the data provide sufficient evidence to reject the null hypothesis.

The best way to think about it, in my experience, is to imagine that a hypothesis test is a criminal trial [4], 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.

8.2 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. [5] 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 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 α, 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 α if the type I error rate is no larger than α.

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 β. 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)	α (type I error rate)
H_0 is false	β (type II error rate)	$1 - \beta$ (power of the test)

A "powerful" hypothesis test is one that has a small value of β, while still keeping α fixed at some (small) desired level. By convention, scientists make use of three different α levels: .05, .01 and .001. Notice the asymmetry here; the tests are designed to ensure that the α level is kept small but there's no corresponding guarantee regarding β. 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.

8.3 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 ESP 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 ESP doesn't exist 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 question 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 is 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. 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.

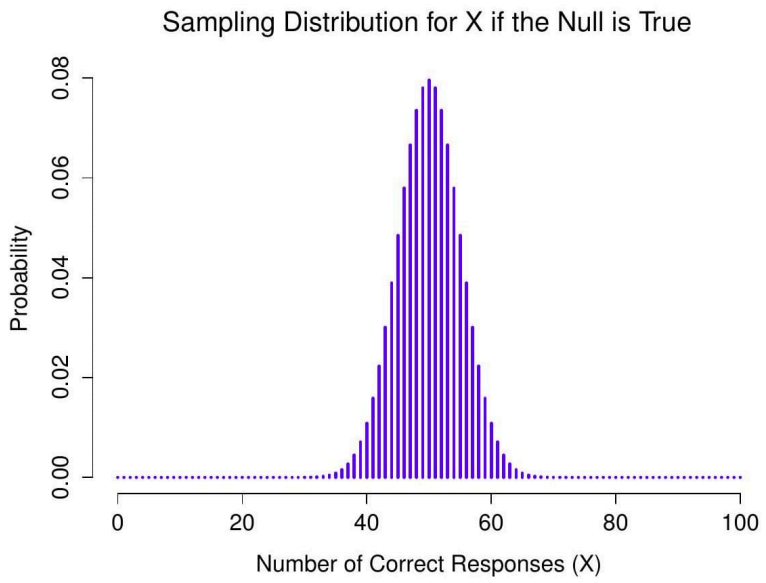


Figure 8.1: The sampling distribution for our test statistic X when the null hypothesis is true. For our ESP scenario this is a binomial distribution. Not surprisingly, since the null hypothesis says that the probability of a correct response is $\theta = .5$, the sampling distribution says that the most likely value is 50 (out of 100) correct responses. Most of the probability mass lies between 40 and 60 .

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 ESP example provides us with one of the easiest cases. Our population parameter θ 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, and that's exactly what the binomial distribution describes! So, to use mathematical notation, 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 8.1. 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.

8.4 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 ESP 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?

8.4.1 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 no more than 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 α level of 0.2 . If we want $\alpha = .05$, the critical region is only allowed to cover up to 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 8.2. Using an approximation to a 5% significance level, our critical regions correspond to X ≤ 40 and X ≥ 60. These cutoffs actually give a Type I error rate of about 5.7%, as explained in the footnote. 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.

Critical Regions for a Two-Sided Test

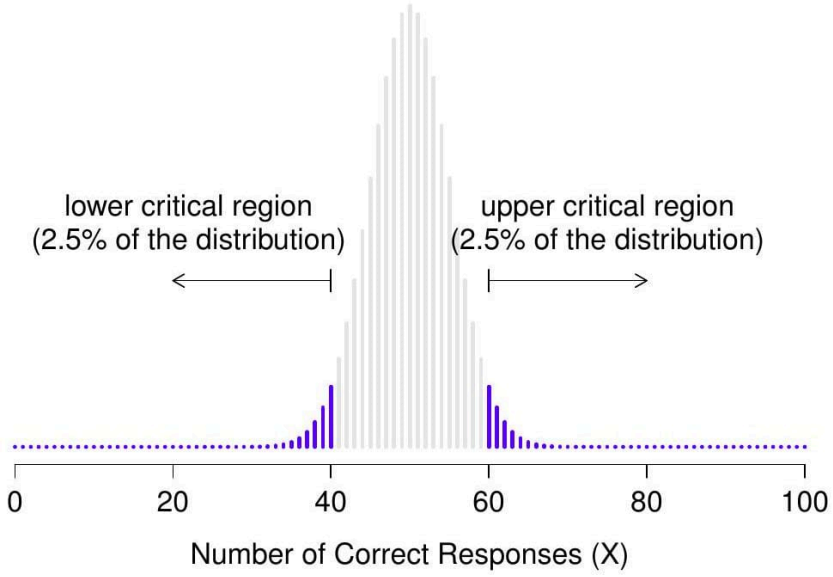


Figure 8.2: The critical region associated with the hypothesis test for the ESP study, for a hypothesis test with a significance level of $\alpha = .05$. The plot shows the sampling distribution of X under the null hypothesis (i.e., same as Figure 8.1). The grey bars correspond to those values of X for 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 α level of .05, we need to ensure that each tail encompasses approximately 2.8% of the sampling distribution.

At this point, our hypothesis test is essentially complete:

1. we choose an α level (e.g., $\alpha = .05$);
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. 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. calculate the critical region that produces an appropriate α 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.

8.4.2 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 [7]

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.

8.4.3 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,

$$\begin{array}{ll} H_0: & \theta = .5 \\ H_1: & \theta \neq .5 \end{array} \tag{8.1.1}$$

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 ESP 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. In the approximate example illustrated in Figure 8.2, each tail contains about 2.8% of the null distribution, giving an actual Type I error rate of about 5.7%.

However, that's not the only possibility. I might only be willing to believe in ESP if it produces better than chance performance. If so, then my alternative hypothesis would only cover the possibility that $\theta > .5$, and as a consequence the null hypothesis now becomes $\theta \leq .5$.

$$\begin{array}{ll} H_0: & \theta \leq .5 \\ H_1: & \theta > .5 \end{array} \tag{8.1.2}$$

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 8.3.

8.5 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.

8.5.1 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 ESP 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 my 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.

Critical Region for a One-Sided Test

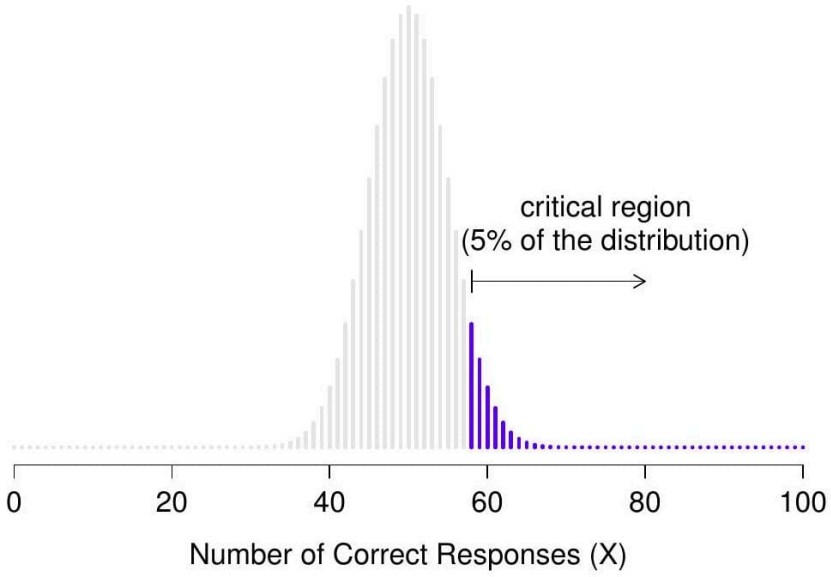


Figure 8.3: The critical region for a one sided test. In this case, the alternative hypothesis is that $\theta > .5$ so we would only reject the null hypothesis for large values of X. As a consequence, the critical region only covers the upper tail of the sampling distribution, specifically the upper 5% of the distribution. Contrast this to the two-sided version in Figure 8.2.

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 α in each case. When we do that for my original ESP data what we'd get is something like this

Value of α	0.05	0.04	0.03	0.02	0.01
Reject the null?	Yes	Yes	Yes	No	No

When we test the ESP data ($X = 62$ successes out of $N = 100$ observations), using α levels of . 03 and above, we'd always find ourselves rejecting the null hypothesis. For α 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 α that would allow us to reject the null hypothesis for this data. This is the p value. As it turns out the ESP data has $p = .021$. In short,
p is defined to be the smallest Type I error rate (α) 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 α 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 ESP study I obtained X = 62 and as a consequence I've ended up with *p* = .021. So the smallest significance level at which I could reject the null is approximately 2.1%. In contrast, suppose my experiment had yielded X = 97. What happens to my *p* value now? This time it's shrunk to approximately $p = 2.63 \times 10^{-25}$, which is a tiny, tiny number. For this second case I would be able to reject the null hypothesis even at an extremely small significance level—about 1 in 3.8×10^{24}.

8.5.2 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, β). 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. The *p*-value is the probability, assuming the null hypothesis is true, of obtaining a test statistic at least as extreme as the one we actually observed. A small *p*-value indicates that the observed result would be unusual under the null hypothesis

8.5.3 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.

8.6 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, 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., $p < .05$).

8.6.1 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 α 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 ESP 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 ESP exists, 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 α 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 α value in advance and then only report whether the test was significant or not. It's the only way to keep ourselves honest.

8.6.2 Two proposed solutions

In practice, it's pretty rare for a researcher to specify a single α 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 8.1. 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 α 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 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 8.1. 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.

Table 8.1: 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

8.7 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 $\mathrm{p} = .5$, [9] and using data in which $\mathrm{x} = 62$ of $\mathrm{n} = 100$ people made the correct response, available in the binomialtest.jasp data file, we get the results shown in Figure 8.4.

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 $\alpha = .05$, 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.

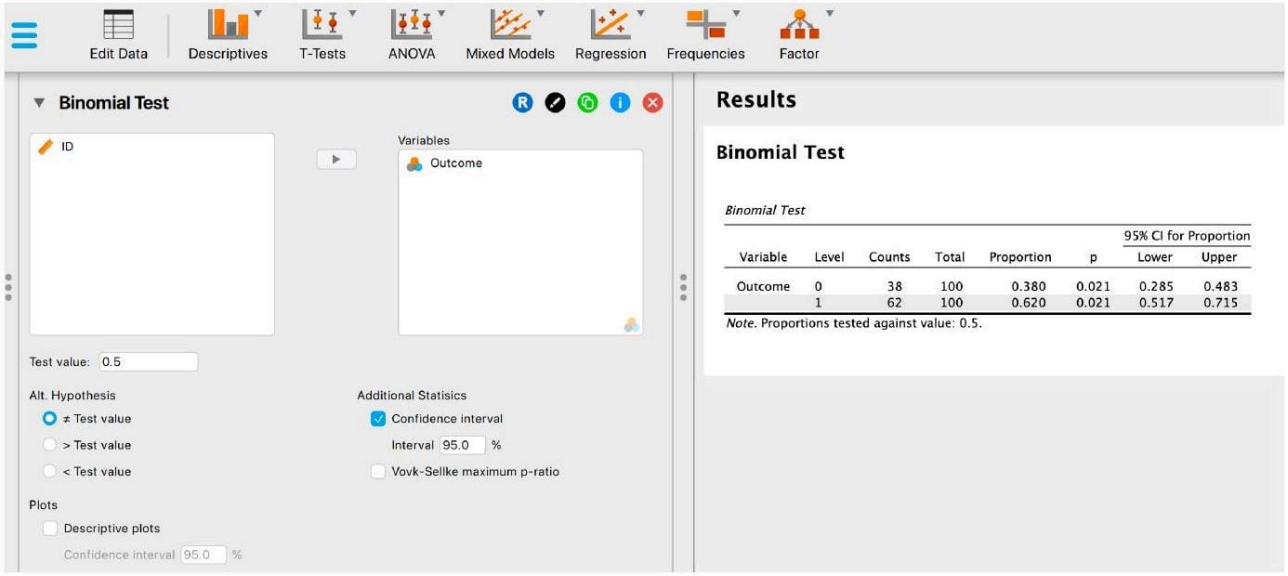


Figure 8.4: Binomial test analysis and results in JASP

8.8 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 β, 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.

8.8.1 The power function

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 β 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 my ESP study the alternative hypothesis actually corresponds to lots of possible values of θ. In fact, the alternative hypothesis corresponds to every value of θ except 0.5. Let's suppose that the true probability of someone choosing the correct response is 55%(i.e., $\theta = .55$).

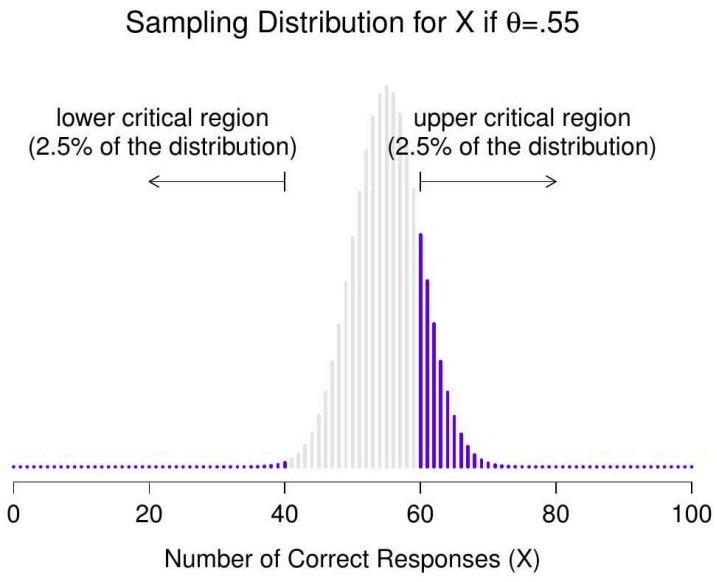


Figure 8.5: Sampling distribution under the alternative hypothesis for a population parameter value of $\theta = 0.55$. A reasonable proportion of the distribution lies in the rejection region.

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 8.5. 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 θ is actually 0.7. What happens to the sampling distribution when this occurs? The answer, shown in Figure 8.6, 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.

What all this means is that the power of a test (i.e., $1 - \beta$) depends on the true value of θ. To illustrate this, I've calculated the expected probability of rejecting the null hypothesis for all values of θ, and plotted it in Figure 8.7. 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 θ. As you can see, when the true value of θ is very close to 0.5 , the power of the test drops very sharply, but when it is further away, the power is large.

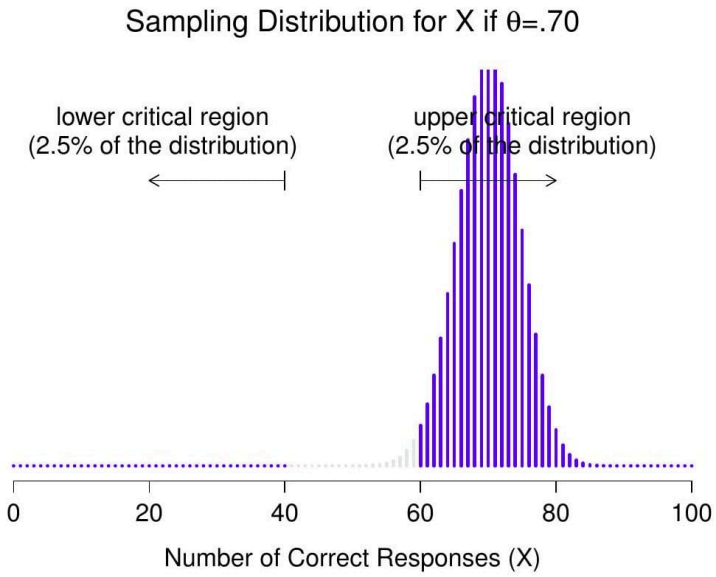


Figure 8.6: Sampling distribution under the alternative hypothesis for a population parameter value of $\theta = 0.70$. Almost all of the distribution lies in the rejection region.

8.8.2 Effect size

> Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned with mice when there are tigers abroad - George Box (Box 1976, p. 792)

The plot shown in Figure 8.7 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?

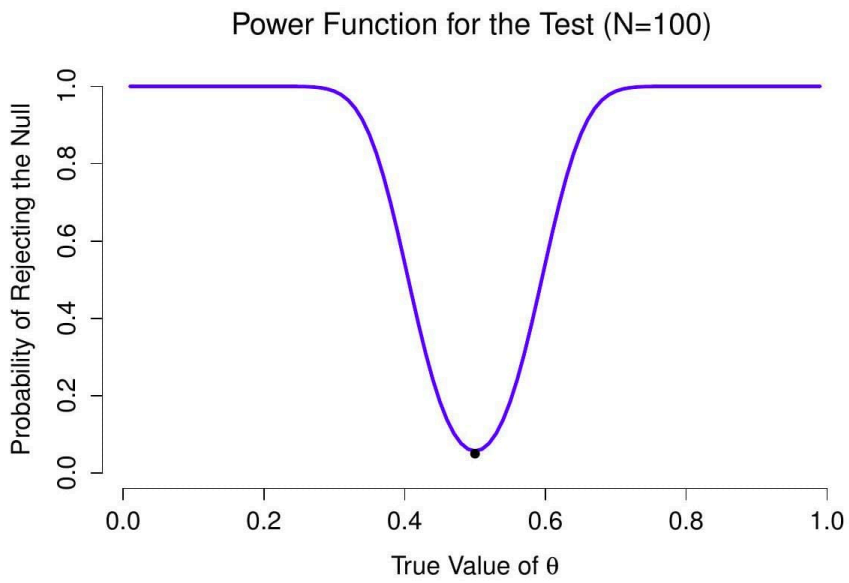


Figure 8.7: The probability that we will reject the null hypothesis, plotted as a function of the true value of θ. Obviously, the test is more powerful (greater chance of correct rejection) if the true value of θ is very different from the value that the null hypothesis specifies (i.e., $\theta = .5$). Notice that when θ 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.

In our ESP example, if we let $\theta_0 = 0.5$ denote the value assumed by the null hypothesis and let θ 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 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 θ 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 ESP study we might still care since any demonstration of real psychic powers would actually be pretty cool [10], 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 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 assesses the evidence against the null hypothesis, whereas the effect size describes the magnitude of the effect. Its practical importance depends on the context.

Table 8.2: Statistical significance and effect size describe different aspects of a result. A non-significant result does not establish that there is no effect, and practical importance depends on the context.

Result	Big effect size	Small effect size
Significant result	Evidence against the null; estimated effect is large	Evidence against the null; estimated effect is small
Non-significant result	Insufficient evidence against the null; estimated effect is large	Insufficient evidence against the null; estimated effect is small

8.8.3 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 ESP 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 ESP abilities somehow then the true value of θ will go up [11] 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, ESP really does exist 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 ESP experiment with 10 participants and 7 of them correctly guessed the colour of the hidden card 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 8.8, 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 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. [12] 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.

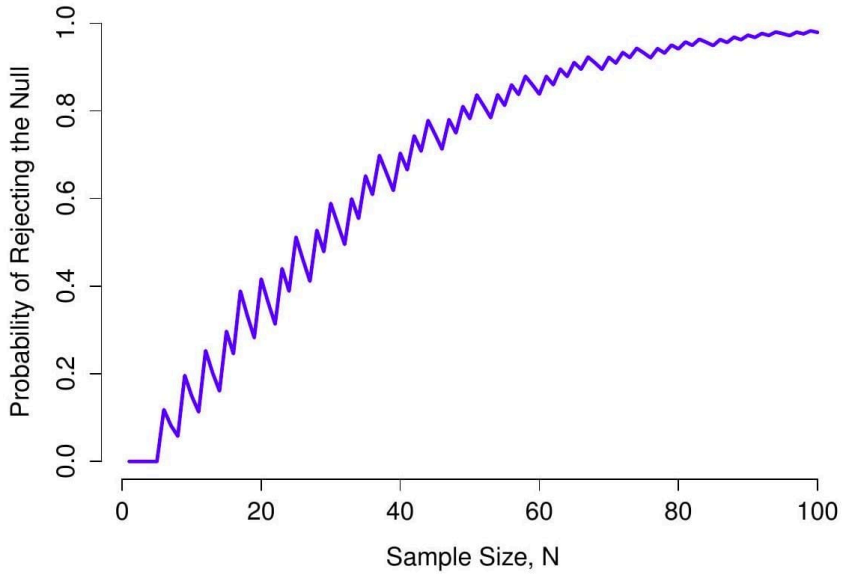


Figure 8.8: The power of our test plotted as a function of the sample size N. In this case, the true value of θ is 0.7 but the null hypothesis is that $\theta = 0.5$. Overall, larger N means greater power. (The small zig-zags in this function occur because of some odd interactions between θ, α and the fact that the binomial distribution is discrete, it doesn't matter for any serious purpose).

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.

8.9 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.

8.9.1 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 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 [13] define the p value in terms of exreme data (Fisher), but we still have α 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.

8.9.2 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 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. For now what I want to point out to you is the p value is a terrible approximation 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!

8.9.3 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 ESP study and I've decided to analyse the data separately for the male participants and the female participants. Of the male participants, 33 out of 50 guessed the colour of the card 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, [14] 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 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.

8.10 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
- -Type 1 and Type 2 errors
- -Test statistics and sampling distributions
- -Hypothesis testing as a decision making process
- -*p*-values as "soft" decisions
- -Writing up the results of a hypothesis test
- -Running the hypothesis test in practice
- -Effect size and power
- -A few issues to consider regarding hypothesis testing

---

1. [1] The quote comes from Wittgenstein's (1922) text, Tractatus Logico-Philosphicus.
   [2] 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.
2. [3] My apologies to anyone who actually believes in this stuff, but on my reading of the literature on ESP it's just not reasonable to think this is real. To be fair, though, some of the studies are rigorously designed, so it's actually an interesting area for thinking about psychological research design. And of course it's a free country so you can spend your own time and effort proving me wrong if you like, but I wouldn't think that's a terribly practical use of your intellect.
3. [4] 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.
4. [5] 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.
5. [6] 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 $\alpha = .05$.
6. [7] 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.
7. [8] That's $p = .00000000000000000000000136$ for folks that don't like scientific notation!
8. [9] 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 θ value.
9. [10] 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.
10. [11] Notice that the true population parameter θ doesn't necessarily correspond to an immutable fact of nature. In this context θ 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 ESP actually exists!

[12] 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.

11. [13] 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.
12. [14] In this case, the Pearson chi-square test of independence

This page titled 8.1: Hypothesis testing is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 8: Hypothesis testing by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

9: CHI-SQUARE TESTS

9.1: Categorical data analysis
9.2: Chi-squared Test

This page titled 9: Chi-Square Tests is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

9.1: Categorical data analysis

Now that we've covered the basic theory behind hypothesis testing it's time to start looking at specific tests that are commonly used in psychology. So where should we start? Not every textbook agrees on where to start, but I'm going to start with " χ^2 tests" (pronounced "chi-square" [1]) and "t-tests". Both of these tools are very frequently used in scientific practice, and whilst they're not as powerful as "regression" and "analysis of variance" they're much easier to understand.

Categorical data include nominal and ordinal data. Categorical data analysis refers to a collection of tools that you can use to analyse counts across categories. There are a lot of different tools that can be used for categorical data analysis, and this chapter covers only a few of the more common ones.

9.1 The χ^2 (chi-square) goodness-of-fit test

The χ^2 goodness-of-fit test is one of the oldest hypothesis tests around. It was invented by Karl Pearson around the turn of the century (Pearson 1900), with some corrections made later by Sir Ronald Fisher (Fisher 1922a). It tests whether an observed frequency distribution of a nominal variable matches an expected frequency distribution. For example, suppose a group of patients has been undergoing an experimental treatment and have had their health assessed to see whether their condition has improved, stayed the same or worsened. A goodness-of-fit test could be used to determine whether the numbers in each category - improved, no change, worsened - match the numbers that would be expected given the standard treatment option, or differ significantly from those numbers, providing evidence that the distribution of outcomes differs from what would be expected under the standard treatment. Let's think about this some more, with some psychology.

9.1.1 The cards data

Over the years there have been many studies showing that humans find it difficult to simulate randomness. Try as we might to "act" random, we think in terms of patterns and structure and so, when asked to "do something at random", what people actually do is anything but random. As a consequence, the study of human randomness (or non-randomness, as the case may be) opens up a lot of deep psychological questions about how we think about the world. With this in mind, let's consider a very simple study. Suppose I asked people to imagine a shuffled deck of cards, and mentally pick one card from this imaginary deck "at random". After they've chosen one card I ask them to mentally select a second one. For both choices what we're going to look at is the suit (hearts, clubs, spades or diamonds) that people chose. After asking, say, $N = 200$ people to do this, I'd like to look at the data and figure out whether or not the cards that people pretended to select were really random. The data are contained in the randomness.csv file in which, when you open it up in JASP, you will see three variables. These are: an id variable that assigns a unique identifier to each participant, and the two variables choice_1 and choice_2 that indicate the card suits that people chose.

For the moment, let's just focus on the first choice that people made. We'll use the Frequency tables option under 'Descriptives' - 'Descriptive Statistics' to count the number of times that we observed people choosing each suit. This is what we get:

```
clubs diamonds hearts spades
    35 51 64 50
```

That little frequency table is quite helpful. Looking at it, there's a bit of a hint that people might be more likely to select hearts than clubs, but it's not completely obvious just from looking at it whether that's really true, or if this is just due to chance. So we'll probably have to do some kind of statistical analysis to find out, which is what I'm going to talk about in the next section.

Excellent. From this point on, we'll treat this table as the data that we're looking to analyse. However, since I'm going to have to talk about this data in mathematical terms, it might be a good idea to be clear about what the notation is. In mathematical notation, we shorten the human-readable word "observed" to the letter O, and we use subscripts to denote the position of the observation. So the second observation in our table is written as O_2 in maths. The relationship between the English descriptions and the mathematical symbols are illustrated below:

label	index, i	math. symbol	the value
clubs,	1	O_1	35
diamonds, ♢	2	O_2	51
hearts,	3	O_3	64
spades,	4	O_4	50

It's also worth noting that mathematicians prefer to talk about general rather than specific things, so you'll also see the notation O_i, which refers to the number of observations that fall within the i-th category (where i could be 1, 2, 3 or 4). Finally, if we want to refer to the set of all observed frequencies, statisticians group all observed values into a vector [2], which I'll refer to using boldface type as $\boldsymbol{O}$.

$$\boldsymbol{O} = (O_1, O_2, O_3, O_4)$$

Again, this is nothing new or interesting. It's just notation. If I say that $\boldsymbol{O} = (35, 51, 64, 50)$ all I'm doing is describing the table of observed frequencies (i.e., observed), but I'm referring to it using mathematical notation.

9.1.2 The null hypothesis and the alternative hypothesis

As the last section indicated, our research hypothesis is that "people don't choose cards randomly". What we're going to want to do now is translate this into some statistical hypotheses and then construct a statistical test of those hypotheses. The test that I'm going to describe to you is Pearson's χ^2 (chi-square) goodness-of-fit test, and as is so often the case we have to begin by carefully constructing our null hypothesis. In this case, it's pretty easy. First, let's state the null hypothesis in words:

$$H_0: \quad \text{All four suits are chosen with equal probability}$$

Now, because this is statistics, we have to be able to say the same thing in a mathematical way. To do this, let's use the notation P_j to refer to the true probability that the j-th suit is chosen. If the null hypothesis is true, then each of the four suits has a 25% chance of being selected. In other words, our null hypothesis claims that $P_1 = .25, P_2 = .25, P_3 = .25$ and finally that $P_4 = .25$. However, in the same way that we can group our observed frequencies into a vector $\boldsymbol{O}$ that summarises the entire data set, we can use $\boldsymbol{P}$ to refer to the probabilities that correspond to our null hypothesis. So if I let the vector $\boldsymbol{P} = (P_1, P_2, P_3, P_4)$ refer to the collection of probabilities that describe our null hypothesis, then we have:

$$H_0: \quad \boldsymbol{P} = (.25, .25, .25, .25)$$

In this particular instance, our null hypothesis corresponds to a vector of probabilities $\boldsymbol{P}$ in which all of the probabilities are equal to one another. But this doesn't have to be the case. For instance, if the experimental task was for people to imagine they were drawing from a deck that had twice as many clubs as any other suit, then the null hypothesis would correspond to something like $\boldsymbol{P} = (.4, .2, .2, .2)$. As long as the probabilities are all positive numbers, and they all sum to 1, then it's a perfectly legitimate choice for the null hypothesis. However, the most common use of the goodness-of-fit test is to test a null hypothesis that all of the categories are equally likely, so we'll stick to that for our example.

What about our alternative hypothesis, H_1 ? All we're really interested in is demonstrating that the probabilities involved aren't all identical (that is, people's choices weren't completely random). As a consequence, the "human friendly" versions of our hypotheses look like this:

H_0 : All four suits are chosen with equal probability
H_1 : At least one of the suit-choice probabilities isn't 0.25

and the "mathematician friendly" version is:

$$\begin{aligned} H_0: &\quad \boldsymbol{P} = (.25, .25, .25, .25) \\ H_1: &\quad \boldsymbol{P} \neq (.25, .25, .25, .25) \end{aligned}$$

9.1.3 The "goodness-of-fit" test statistic

At this point, we have our observed frequencies $\boldsymbol{O}$ and a collection of probabilities $\boldsymbol{P}$ corresponding to the null hypothesis that we

want to test. What we now want to do is construct a test of the null hypothesis. As always, if we want to test H_0 against H_1, we're going to need a test statistic. The basic trick that a goodness-of-fit test uses is to construct a test statistic that measures how "close" the data are to the null hypothesis. If the data don't resemble what you'd "expect" to see if the null hypothesis were true, then it probably isn't true. Okay, if the null hypothesis were true, what would we expect to see? Or, to use the correct terminology, what are the expected frequencies. There are $N = 200$ observations, and (if the null is true) the probability of any one of them choosing a heart is $P_3 = .25$, so I guess we're expecting $200 \times .25 = 50$ hearts, right? Or, more specifically, if we let E_i refer to "the number of category i responses that we're expecting if the null is true", then

$$E_i = N \times P_i$$

This is pretty easy to calculate. If there are 200 observations that can fall into four categories, and we think that all four categories are equally likely, then on average we'd expect to see 50 observations in each category, right?

Now, how do we translate this into a test statistic? Clearly, what we want to do is compare the expected number of observations in each category (E_i) with the observed number of observations in that category (O_i). And on the basis of this comparison we ought to be able to come up with a good test statistic. To start with, let's calculate the difference between what the null hypothesis expected us to find and what we actually did find. That is, we calculate the "observed minus expected" difference score, $O_i - E_i$. This is illustrated in the following table.

		clubs	diamonds	hearts	spades
expected frequency	E_i	50	50	50	50
observed frequency	O_i	35	51	64	50
difference score	$O_i - E_i$	-15	1	14	0

So, based on our calculations, it's clear that people chose more hearts and fewer clubs than the null hypothesis predicted. However, a moment's thought suggests that these raw differences aren't
quite what we're looking for. Intuitively, it feels like it's just as bad when the null hypothesis predicts too few observations (which is what happened with hearts) as it is when it predicts too many (which is what happened with clubs). So it's a bit weird that we have a negative number for clubs and a positive number for hearts. One easy way to fix this is to square everything [3], so that we now calculate the squared differences, $(O_i - E_i)^2$. As before, we can do this by hand:

		clubs	diamonds	hearts	spades
expected frequency	E_i	50	50	50	50
observed frequency	O_i	35	51	64	50
difference score	$O_i - E_i$	-15	1	14	0
squared difference	$(O_i - E_i)^2$	225	1	196	0

Now we're making progress. What we've got now is a collection of numbers that are big whenever the null hypothesis makes a bad prediction (clubs and hearts), but are small whenever it makes a good one (diamonds and spades). Next, for some technical reasons that I'll explain in a moment, let's also divide all these numbers by the expected frequency E_i, so we're actually calculating the scaled squared difference, $\frac{(E_i - O_i)^2}{E_i}$. Since $E_i = 50$ for all categories in our example, it's not a very interesting calculation, but let's do it anyway:

		clubs	diamonds	hearts	spades
expected frequency	E_i	50	50	50	50
observed frequency	O_i	35	51	64	50
difference score	$O_i - E_i$	-15	1	14	0
squared difference	$(O_i - E_i)^2$	225	1	196	0
scaled sq. diff.	$(O_i - E_i)^2 / E_i$	4.50	0.02	3.92	0.00

In effect, what we've got here are four different "error" scores, each one telling us how big a "mistake" the null hypothesis made

when we tried to use it to predict our observed frequencies. So, in order to convert this into a useful test statistic, one thing we could do is just add these numbers up. The result is called the goodness-of-fit statistic, conventionally referred to either as χ^2 (chi-square) or GOF. We can calculate it as $4.50 + 0.02 + 3.92 + 0.00 = 8.44$.

If we let k refer to the total number of categories (i.e., $k = 4$ for our cards data), then the χ^2 statistic is given by:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Intuitively, it's clear that if χ^2 is small, then the observed data O_i are very close to what the null hypothesis predicted E_i, so we're going to need a large χ^2 statistic in order to reject the null.

As we've seen from our calculations, in our cards data set we've got a value of $\chi^2 = 8.44$. So now the question becomes is this a big enough value to reject the null?

9.1.4 The sampling distribution of the GOF statistic

To determine whether or not a particular value of χ^2 is large enough to justify rejecting the null hypothesis, we're going to need to figure out what the sampling distribution for χ^2 would be if the null hypothesis were true. So that's what I'm going to do in this section. I'll show you in a fair amount of detail how this sampling distribution is constructed, and then, in the next section, use it to build up a hypothesis test. If you want to cut to the chase and are willing to take it on faith that the sampling distribution is a χ^2 (chi-square) distribution with $k - 1$ degrees of freedom, you can skip the rest of this section. However, if you want to understand why the goodness-of-fit test works the way it does, read on.

Okay, let's suppose that the null hypothesis is actually true. If so, then the true probability that an observation falls in the i-th category is P_i. After all, that's pretty much the definition of our null hypothesis. Let's think about what this actually means. This is kind of like saying that "nature" makes the decision about whether or not the observation ends up in category i by flipping a weighted coin (i.e., one where the probability of getting a head is P_j). And therefore we can think of our observed frequency O_i by imagining that nature flipped N of these coins (one for each observation in the data set), and exactly O_i of them came up heads. Obviously, this is a pretty weird way to think about the experiment. But what it does (I hope) is remind you that we've actually seen this scenario before. It's exactly the same set up that gave rise to the binomial distribution. In other words, if the null hypothesis is true, then it follows that our observed frequencies were generated by sampling from a binomial distribution:

$$O_i \sim \text{Binomial}\,(P_i, N)$$

When the expected frequencies are sufficiently large, the binomial distribution for each category count is approximately normal. This helps explain why a normal approximation is useful, but the category counts are not independent: they must add up to the total sample size.

For the goodness-of-fit statistic, the result is that its sampling distribution is approximately chi-square with k – 1 degrees of freedom, where k is the number of categories. This assumes that the expected probabilities were specified in advance rather than estimated from the same data. The fixed total imposes one constraint, leaving k – 1 degrees of freedom. That’s the topic of the next section.

9.1.5 Degrees of freedom

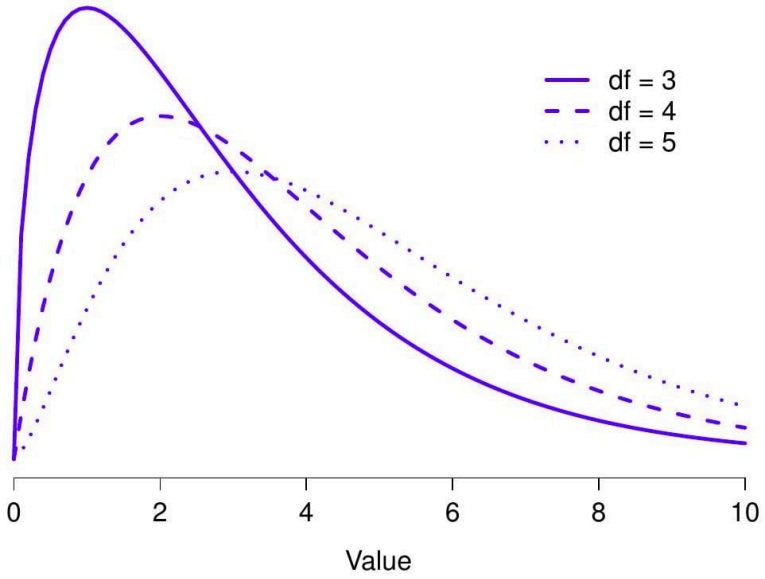


Figure 9.1: χ^2 (chi-square) distributions with different values for the "degrees of freedom".

When I introduced the chi-square distribution, I was a bit vague about what "degrees of freedom" actually means. Obviously, it matters. Looking at Figure 9.1, you can see that if we change the degrees of freedom then the chi-square distribution changes shape quite substantially. But what exactly is it? Again, when I introduced the distribution and explained its relationship to the normal distribution, I did offer an answer: it's the number of "normally distributed variables" that I'm squaring and adding together. But, for most people, that's kind of abstract and not entirely helpful. What we really need to do is try to understand degrees of freedom in terms of our data. So here goes.

The basic idea behind degrees of freedom is quite simple. You calculate it by counting up the number of distinct "quantities" that are used to describe your data and then subtracting off all of the "constraints" that those data must satisfy. [5] This is a bit vague, so let's use our cards data as a concrete example. We describe our data using four numbers, O_1, O_2, O_3 and O_4 corresponding to the observed frequencies of the four different categories (hearts, clubs, diamonds, spades). These four numbers are the random outcomes of our experiment. But my experiment actually has a fixed constraint built into it: the sample size N. [6] That is, if we know how many people chose hearts, how many chose diamonds and how many chose clubs, then we'd be able to figure out exactly how many chose spades. In other words, although our data are described using four numbers, they only actually correspond to $4 - 1 = 3$ degrees of freedom. A slightly different way of thinking about it is to notice that there are four probabilities that we're interested in (again, corresponding to the four different categories), but these probabilities must sum to one, which imposes a constraint. Therefore the degrees of freedom is $4 - 1 = 3$. Regardless of whether you want to think about it in terms of the observed frequencies or in terms of the probabilities, the answer is the same. In general, when running the χ^2 (chi-square) goodness-of-fit test for an experiment involving k groups, then the degrees of freedom will be $k - 1$.

9.1.6 Testing the null hypothesis

The final step in the process of constructing our hypothesis test is to figure out what the rejection region is. That is, what values of χ^2 would lead us to reject the null hypothesis. As we saw earlier, large values of χ^2 imply that the null hypothesis has done a poor job of predicting the data from our experiment, whereas small values of χ^2 imply that it's actually done pretty well. Therefore, a pretty sensible strategy would be to say there is some critical value such that if χ^2 is bigger than the critical value we reject the null, but if χ^2 is smaller than this value we retain the null. In other words, the chi-square goodness-of-fit test is always a one-sided test. Right, so all we have to do is figure out what this critical value is. And it's pretty straightforward. If we want our test to have significance level of $\alpha = .05$ (that is, we are willing to tolerate a Type I error rate of 5%), then we have to choose our critical value so that there is only a 5% chance that χ^2 could get to be that big if the null hypothesis is true. This is illustrated in Figure 9.2.

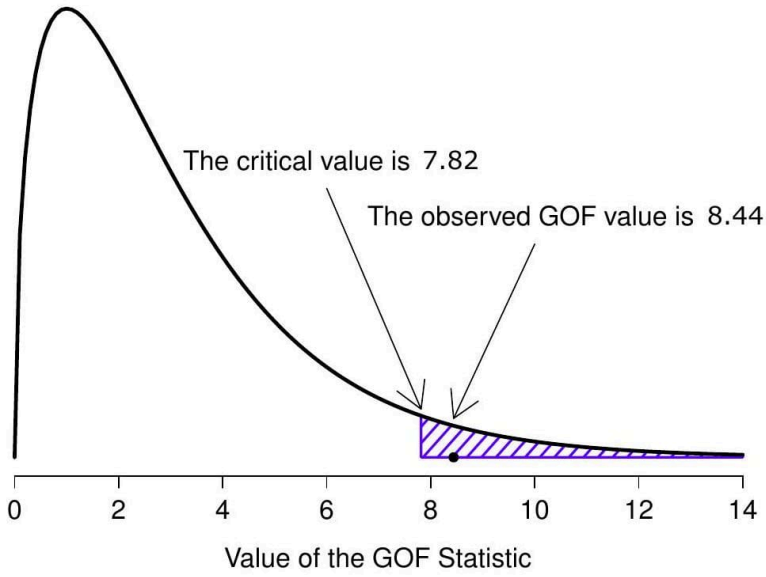


Figure 9.2: Illustration of how the hypothesis testing works for the χ^2 (chi-square) goodness-of-fit test.

Ah but, I hear you ask, how do I find the critical value of a chi-square distribution with $k-1$ degrees of freedom? Many many years ago when I first took a psychology statistics class we used to look up these critical values in a book of critical value tables, like the one in Figure 9.3. Looking at this Figure, we can see that the critical value for a χ^2 distribution with 3 degrees of freedom, and $\mathrm{p}=0.05$ is 7.815.

So, if our calculated χ^2 statistic is bigger than the critical value of 7.815 , then we can reject the null hypothesis (remember that the null hypothesis, H_0, is that all four suits are chosen with equal probability). Since we actually already calculated that before (i.e., $\chi^2=8.44$) we can reject the null hypothesis. And that's it, basically. You now know "Pearson's χ^2 test for the goodness-of-fit". Lucky you.

9.1.7 Doing the test in JASP

Not surprisingly, JASP provides an analysis that will do these calculations for you. From the main 'Analyses' toolbar select 'Frequencies' - 'Multinomial Test'. Then in the analysis window that appears move the variable you want to analyse (choice_1) across into the 'Factor' box. Also, click on the 'Descriptives' check box so that you see the expected counts in the results table. When you have done all this, you should see the analysis results in JASP as in Figure 9.4. No surprise then that JASP provides the same expected counts and statistics that we calculated by hand above, with a χ^2 value of 8.44 with 3 d.f. and $p=0.038$. Note that we don't need to look up a critical p -value threshold value any more, as JASP gives us the actual p-value of the calculated χ^2 for 3 d.f.

	Probability								
Degrees of Freedom	0.95	0.90	0.70	0.50	0.30	0.10	0.05	0.01	0.001
1	0.004	0.016	0.148	0.455	1.074	2.706	3.841	6.635	10.828
2	0.103	0.211	0.713	1.386	2.408	4.605	5.991	9.210	13.816
3	0.352	0.584	1.424	2.366	3.665	6.251	7.815	11.345	16.266
4	0.711	1.064	2.195	3.357	4.878	7.779	9.488	13.277	18.467
5	1.145	1.610	3.000	4.351	6.064	9.236	11.070	15.086	20.515
6	1.635	2.204	3.828	5.348	7.231	10.645	12.592	16.812	22.458
7	2.167	2.833	4.671	6.346	8.383	12.017	14.067	18.475	24.322
8	2.733	3.490	5.527	7.344	9.524	13.362	15.507	20.090	26.124

	Probability								
9	3.325	4.168	6.393	8.343	10.656	14.684	16.919	21.666	27.877
10	3.940	4.865	7.267	9.342	11.781	15.987	18.307	23.209	29.588
	Non-significant						Significant		

Figure 9.3: Table of critical values for the chi-square distribution

9.1.8 Specifying a different null hypothesis

At this point you might be wondering what to do if you want to run a goodness-of-fit test but your null hypothesis is not that all categories are equally likely. For instance, let's suppose that someone had made the theoretical prediction that people should choose red cards 60% of the time, and black cards 40% of the time (I've no idea why you'd predict that), but had no other preferences. If that were the case, the null hypothesis would be to expect 30% of the choices to be hearts, 30% to be diamonds, 20% to be spades and 20% to be clubs. In other words we would expect hearts and diamonds to each appear 60 times (30% of 200 is 60), and spades and clubs to each appear 40 times (20% of 200 is 40). This seems like a silly theory to me, but nonetheless, it's pretty easy to test this explicitly specified null hypothesis with the data in our JASP analysis. In the analysis window (see Figure 9.4) you can click the radio button for 'Custom expected proportions (χ^2 test)'. When you do this, there are options for entering different expected counts for the variable you have selected, in our case this is choice_1. Change the counts to reflect the new null hypothesis, as in Figure 9.5, and see how the results change.

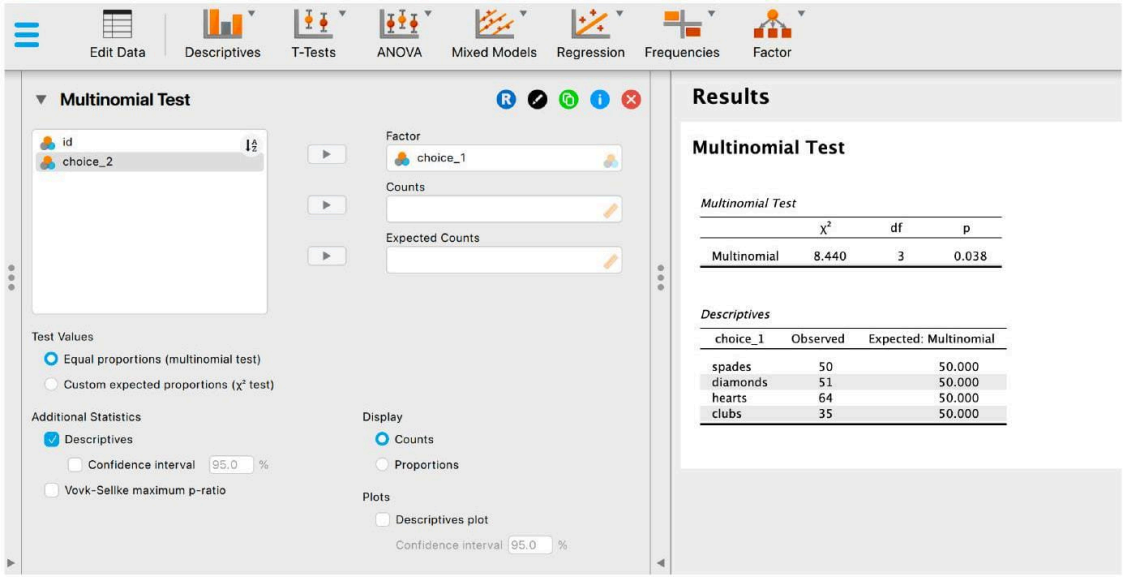


Figure 9.4: A χ^2 goodness-of-fit test in JASP, with table showing both observed and expected frequencies.

The expected counts are now:

		♠	♢	♡	♣
expected frequency	E_i	40	60	60	40

and the χ^2 statistic is $4.742, 3$ d.f., $p = 0.192$. Now, the results of our updated hypotheses and the expected frequencies are different from what they were last time. As a consequence our χ^2 test statistic is different, and our p-value is different too. Annoyingly, the p-value is .192 , so we can't reject the null hypothesis. Sadly, despite the fact that the null hypothesis corresponds to a very silly theory, these data don't provide enough evidence against it.

9.1.9 How to report the results of the test

So now you know how the test works, and you know how to do the test using a wonderful JASP-flavoured magic computing box. The next thing you need to know is how to write up the results. After all, there's no point in designing and running an experiment and then analysing the data if you don't tell anyone about it! So let's now talk about what you need to do when reporting your analysis. Let's stick with our card-suits example. If I wanted to write this result up for a paper or other report, then the conventional way to report this would be to write something like this:

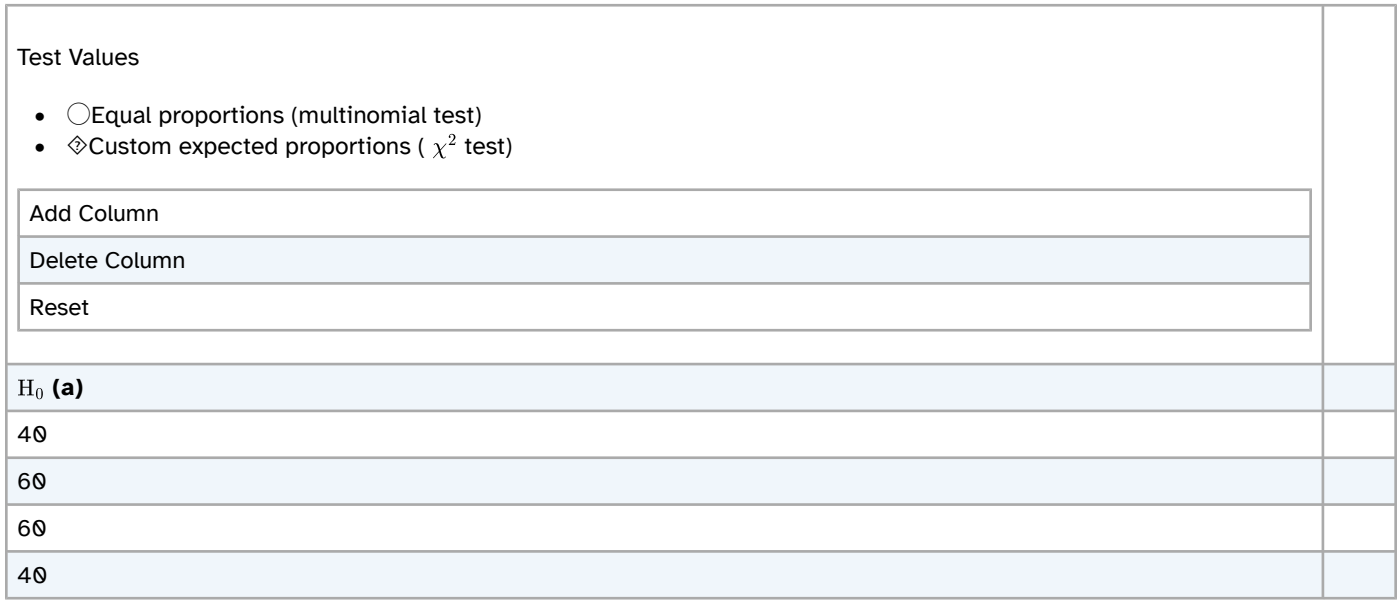


Figure 9.5: Changing the expected proportions in the χ^2 goodness-of-fit test in JASP

Of the 200 participants in the experiment, 64 selected hearts for their first choice, 51 selected diamonds, 50 selected spades, and 35 selected clubs. A chi-square goodness-of-fit test was conducted to test whether the choice probabilities were identical for all four suits. The results were significant $(\chi^2(3) = 8.44, p < .05)$, suggesting that people did not select suits purely at random.

There are a few things that you should note about this description:

- -The statistical test is preceded by the descriptive statistics. That is, I told the reader something about what the data look like before going on to do the test. In general, this is good practice. Always remember that your reader doesn't know your data anywhere near as well as you do. So, unless you describe it to them properly, the statistical tests won't make any sense to them and they'll get frustrated and cry.
- -The description tells you what the null hypothesis being tested is. To be honest, writers don't always do this but it's often a good idea in those situations where some ambiguity exists, or when you can't rely on your readership being intimately familiar with the statistical tools that you're using. Quite often the reader might not know (or remember) all the details of the test that your using, so it's a kind of politeness to "remind" them! As far as the goodness-of-fit test goes, you can usually rely on a scientific audience knowing how it works (since it's covered in most intro stats classes). However, it's still a good idea to be explicit about stating the null hypothesis (briefly!) because the null hypothesis can be different depending on what you're using the test for. For instance, in the cards example my null hypothesis was that all the four suit probabilities were identical (i.e., $P_1 = P_2 = P_3 = P_4 = 0.25$), but there's nothing special about that hypothesis. I could just as easily have tested the null hypothesis that $P_1 = 0.7$ and $P_2 = P_3 = P_4 = 0.1$ using a goodness-of-fit test. So it's helpful to the reader if you explain to them what your null hypothesis was. Also, notice that I described the null hypothesis in words, not in maths. That's perfectly acceptable. You can describe it in maths if you like, but since most readers find words easier to read than symbols, most writers tend to describe the null using words if they can.

- -A "stat block" is included. When reporting the results of the test itself, I didn't just say that the result was significant, I included a "stat block" (i.e., the dense mathematical-looking part in the parentheses) which reports all the "key" statistical information. For the chi-square goodness-of-fit test, the information that gets reported is the test statistic (that the goodness-of-fit statistic was 8.44), the information about the distribution used in the test (χ^2 with 3 degrees of freedom which is usually shortened to $\chi^2(3)$), and then the information about whether the result was significant (in this case $p < .05$

). The particular information that needs to go into the stat block is different for every test, and so each time I introduce a new test I'll show you what the stat block should look like. [7] However the general principle is that you should always provide enough information so that the reader could check the test results themselves if they really wanted to.
- -The results are interpreted. In addition to indicating that the result was significant, I provided an interpretation of the result (i.e., that people didn't choose randomly). This is also a kindness to the reader, because it tells them something about what they should believe about what's going on in your data. If you don't include something like this, it's really hard for your reader to understand what's going on. [8]

As with everything else, your overriding concern should be that you explain things to your reader. Always remember that the point of reporting your results is to communicate to another human being. I cannot tell you just how many times I've seen the results section of a report or a thesis or even a scientific article that is just gibberish, because the writer has focused solely on making sure they've included all the numbers and forgotten to actually communicate with the human reader.

9.1.10 A comment on statistical notation

Satan delights equally in statistics and in quoting scripture

If you've been reading very closely, and are as much of a mathematical pedant as I am, there is one thing about the way I wrote up the chi-square test in the last section that might be bugging you a little bit. There's something that feels a bit wrong with writing " $\chi^2(3) = 8.44''$, you might be thinking. After all, it's the goodness-of-fit statistic that is equal to 8.44 , so shouldn't I have written $X^2 = 8.44$ or maybe $\text{GOF} = 8.44$? This seems to be conflating the sampling distribution (i.e., χ^2 with $df = 3$) with the test statistic (i.e., X^2). Odds are you figured it was a typo, since χ and X look pretty similar. Oddly, it's not. Writing $\chi^2(3) = 8.44$ is essentially a highly condensed way of writing "the sampling distribution of the test statistic is $\chi^2(3)$, and the value of the test statistic is 8.44".

In one sense, this is kind of silly. There are lots of different test statistics out there that turn out to have a chi-square sampling distribution. The X^2 statistic that we've used for our goodness-of-fit test is only one of many (albeit one of the most commonly encountered ones). In a sensible, perfectly organised world we'd always have a separate name for the test statistic and the sampling distribution. That way, the stat block itself would tell you exactly what it was that the researcher had calculated. Sometimes this happens. For instance, the test statistic used in the Pearson goodness-of-fit test is written X^2, but there's a closely related test known as the G-test [9] (Sokal and Rohlf 1994), in which the test statistic is written as G. As it happens, the Pearson goodness-of-fit test and the G-test both test the same null hypothesis, and the sampling distribution is exactly the same (i.e., chi-square with $k - 1$ degrees of freedom). If I'd done a G-test for the cards data rather than a goodness-of-fit test, then I'd have ended up with a test statistic of $G = 8.65$, which is slightly different from the $X^2 = 8.44$ value that I got earlier and which produces a slightly smaller p-value of $p = .034$. Suppose that the convention was to report the test statistic, then the sampling distribution, and then the p-value. If that were true, then these two situations would produce different stat blocks: my original result would be written $X^2 = 8.44, \chi^2(3), p = .038$, whereas the new version using the G-test would be written as $G = 8.65, \chi^2(3), p = .034$. However, using the condensed reporting standard, the original result is written $\chi^2(3) = 8.44, p = .038$, and the new one is written $\chi^2(3) = 8.65, p = .034$, and so it's actually unclear which test I actually ran.

So why don't we live in a world in which the contents of the stat block uniquely specifies what tests were ran? The deep reason is that life is messy. We (as users of statistical tools) want it to be nice and neat and organised. We want it to be designed, as if it were a product, but that's not how life works. Statistics is an intellectual discipline just as much as any other one, and as such it's a massively distributed, partly-collaborative and partly-competitive project that no-one really understands completely. The things that you and I use as data analysis tools weren't created by an Act of the Gods of Statistics. They were invented by lots of different people, published as papers in academic journals, implemented, corrected and modified by lots of other people, and then explained to students in textbooks by someone else. As a consequence, there's a lot of test statistics that don't even have names, and as a consequence they're just given the same name as the corresponding sampling distribution. As we'll see later, any test statistic that follows a χ^2 distribution is commonly called a "chi-square statistic", anything that follows a t-distribution is called a " t-statistic", and so on. But, as the χ^2 versus G example illustrates, two different things with the same sampling distribution are still, well, different.

As a consequence, it's sometimes a good idea to be clear about what the actual test was that you ran, especially if you're doing something unusual. If you just say "chi-square test" it's not actually clear what test you're talking about. Although, since the two most common chi-square tests are the goodness-of-fit test and the independence test (Section 9.2), most readers with stats training can probably guess. Nevertheless, it's something to be aware of.

9.2 The χ^2 test of independence (or association)

GUARDBOT 1:	Halt!
GUARDBOT 2:	Be you robot or human?
LEELA:	Robot...we be.
FRY:	Uh, yup! Just two robots out roboting it up! Eh?
GUARDBOT 1:	Administer the test.
GUARDBOT 2:	Which of the following would you most prefer? A: A puppy, B: A pretty flower from your sweetie, or C : A large properly-formatted data file?
GUARDBOT 1:	Choose!

- Futurama, "Fear of a Bot Planet"

The other day I was watching an animated documentary examining the quaint customs of the natives of the planet Chapek 9. Apparently, in order to gain access to their capital city a visitor must prove that they're a robot, not a human. In order to determine whether or not a visitor is human, the natives ask whether the visitor prefers puppies, flowers, or large, properly formatted data files. "Pretty clever," I thought to myself "but what if humans and robots have the same preferences? That probably wouldn't be a very good test then, would it?" As it happens, I got my hands on the testing data that the civil authorities of Chapek 9 used to check this. It turns out that what they did was very simple. They found a bunch of robots and a bunch of humans and asked them what they preferred. I saved their data in a file called chapek9.csv, which we can now load into JASP. As well as the ID variable that identifies individual people, there are two nominal text variables, species and choice. In total there are 180 entries in the data set, one for each person (counting both robots and humans as "people") who was asked to make a choice. Specifically, there are 93 humans and 87 robots, and overwhelmingly the preferred choice is the data file. You can check this yourself by asking JASP for Frequency Tables, under the 'Descriptives' - 'Descriptive Statistics' button. However, this summary does not address the question we're interested in. To do that, we need a more detailed description of the data. What we want to do is look at the choices broken down by species. That is, we need to cross-tabulate the data. In JASP we do this using the 'Frequencies' - 'Contingency Tables' button, moving species into the 'Columns' box and choice into the 'Rows' box. This procedure should produce a table similar to this:

	Robot	Human	Total
Puppy	13	15	28
Flower	30	13	43
Data	44	65	109
Total	87	93	180

From this, it's quite clear that the vast majority of the humans chose the data file, whereas the robots tended to be a lot more even in their preferences. Leaving aside the question of why the humans might be more likely to choose the data file for the moment (which does seem quite odd, admittedly), our first order of business is to determine if the discrepancy between human choices and robot choices in the data set is statistically significant.

9.2.1 Constructing our hypothesis test

How do we analyse this data? Specifically, since my research hypothesis is that "humans and robots answer the question in different ways", how can I construct a test of the null hypothesis that "humans and robots answer the question the same way"? As before, we begin by establishing some notation to describe the data:

	Robot	Human	Total
Puppy	O_{11}	O_{12}	R_1
Flower	O_{21}	O_{22}	R_2
Data	O_{31}	O_{32}	R_3
Total	C_1	C_2	N

In this notation we say that O_{ij} is a count (observed frequency) of the number of respondents that are of species j (robots or human) who gave answer i (puppy, flower or data) when asked to make a choice. The total number of observations is written N, as usual. Finally, I've used R_i to denote the row totals (e.g., R_1 is the total number of people who chose the puppy), and C_j to denote the column totals (e.g., C_1 is the total number of robots). [10]

So now let's think about what the null hypothesis says. If robots and humans are responding in the same way to the question, it means that the probability that "a robot says puppy" is the same as the probability that "a human says puppy", and so on for the other two possibilities. So, if we use P_{ij} to denote "the probability that a member of species j gives response i " then our null hypothesis is that:

H_0 : All of the following are true:

$$P_{11} = P_{12}(\text{ same probability of saying "puppy" }),$$
$$P_{21} = P_{22}(\text{ same probability of saying "flower" }), \text{ and}$$
$$P_{31} = P_{32}(\text{ same probability of saying "data" }).$$

And actually, since the null hypothesis is claiming that the true choice probabilities don't depend on the species of the person making the choice, we can let P_i refer to this probability, e.g., P_1 is the true probability of choosing the puppy.

Next, in much the same way that we did with the goodness-of-fit test, what we need to do is calculate the expected frequencies. That is, for each of the observed counts O_{ij}, we need to figure out what the null hypothesis would tell us to expect. Let's denote this expected frequency by E_{ij}. This time, it's a little bit trickier. If there are a total of C_j people that belong to species j, and the true probability of anyone (regardless of species) choosing option i is P_i, then the expected frequency is just:

$$E_{ij} = C_j \times P_i$$

Now, this is all very well and good, but we have a problem. Unlike the situation we had with the goodness-of-fit test, the null hypothesis doesn't actually specify a particular value for P_i. It's something we have to estimate from the data! Fortunately, this is pretty easy to do. If 28 out of 180 people selected the puppies, then a natural estimate for the probability of choosing puppies is 28/180, which is approximately .16. If we phrase this in mathematical terms, what we're saying is that our estimate for the probability of choosing option i is just the row total divided by the total sample size:

$$\hat{P}_i = \frac{R_i}{N}$$

Therefore, our expected frequency can be written as the product (i.e. multiplication) of the row total and the column total, divided by the total number of observations: [11]

$$E_{ij} = \frac{R_i \times C_j}{N}$$

Now that we've figured out how to calculate the expected frequencies, it's straightforward to define a test statistic, following the exact same strategy that we used in the goodness-of-fit test. In fact, it's pretty much the same statistic.

For a contingency table with r rows and c columns, the equation that defines our X^2 statistic is

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

The only difference is that I have to include two summation signs (i.e., Σ) to indicate that we're summing over both rows and columns.

As before, large values of X^2 indicate that the null hypothesis provides a poor description of the data, whereas small values of X^2

suggest that it does a good job of accounting for the data. Therefore, just like last time, we want to reject the null hypothesis if X^2 is too large.

Not surprisingly, this statistic is χ^2 distributed. All we need to do is figure out how many degrees of freedom are involved, which actually isn't too hard. As I mentioned before, you can (usually) think of the degrees of freedom as being equal to the number of data points that you're analysing, minus the number of constraints. A contingency table with r rows and c columns contains a total of $r \times c$ observed frequencies, so that's the total number of observations. What about the constraints? Here, it's slightly trickier. The answer is always the same

$$df = (r-1)(c-1)$$

but the explanation for why the degrees of freedom takes this value is different depending on the experimental design. For the sake of argument, let's suppose that we had honestly intended to survey exactly 87 robots and 93 humans (column totals fixed by the experimenter), but left the row totals free to vary (row totals are random variables). Let's think about the constraints that apply here. Well, since we deliberately fixed the column totals by Act of Experimenter, we have c constraints right there. But, there's actually more to it than that. Remember how our null hypothesis had some free parameters (i.e., we had to estimate the P_i values)? Those matter too. I won't explain why in this book, but every free parameter in the null hypothesis is rather like an additional constraint. So, how many of those are there? Well, since these probabilities have to sum to 1 , there's only $r-1$ of these. So our total degrees of freedom is:

$$\begin{aligned} df &= (\text{ number of observations }) - (\text{ number of constraints }) \\ &= (rc) - (c + (r-1)) \\ &= rc - c - r + 1 \\ &= (r-1)(c-1) \end{aligned}$$

Alternatively, suppose that the only thing that the experimenter fixed was the total sample size N. That is, we quizzed the first 180 people that we saw and it just turned out that 87 were robots and 93 were humans. This time around our reasoning would be slightly different, but would still lead us to the same answer. Our null hypothesis still has $r-1$ free parameters corresponding to the choice
probabilities, but it now also has $c-1$ free parameters corresponding to the species probabilities, because we'd also have to estimate the probability that a randomly sampled person turns out to be a robot. [12] Finally, since we did actually fix the total number of observations N, that's one more constraint. So, now we have rc observations, and $(c-1)+(r-1)+1$ constraints. What does that give?

$$\begin{aligned} df &= (\text{number of observations}) - (\text{number of constraints}) \\ &= rc - ((c-1) + (r-1) + 1) \\ &= rc - c - r + 1 \\ &= (r-1)(c-1) \end{aligned}$$

Amazing.

9.2.2 Doing the test in JASP

Okay, now that we know how the test works let's have a look at how it's done in JASP. As tempting as it is to lead you through the tedious calculations so that you're forced to learn it the long way, I figure there's no point. I already showed you how to do it the long way for the goodness-of-fit test in the last section, and since the test of independence isn't conceptually any different, you won't learn anything new by doing it the long way. So instead I'll go straight to showing you the easy way. After you have run the test in JASP ('Frequencies' - 'Contingency Tables'), all you have to do is look underneath the contingency table in the JASP results window and there is the χ^2 statistic for you. This shows a χ^2 statistic value of 10.722, with 2 d.f. and p-value $= 0.005$.

That was easy, wasn't it! You can also ask JASP to show you the expected counts - just click on the check box for 'Counts' - 'Expected' in the 'Cells' options and the expected counts will appear in the contingency table. And whilst you are doing that, an effect size measure would be helpful. We'll choose Cramer's V, and you can specify this from a check box in the 'Statistics' options, and it gives a value for Cramer's V of 0.244 . We will talk about this some more in just a moment.

This output gives us enough information to write up the result:

> Pearson's χ^2 revealed a significant association between species and choice ($\chi^2(2) = 10.7$, $p < .01$). Robots appeared to be more likely to say that they prefer flowers, but the humans were more likely to say they prefer data.

Notice that, once again, I provided a little bit of interpretation to help the human reader understand what's going on with the data. Later on in my discussion section I'd provide a bit more context. To illustrate the difference, here's what I'd probably say later on:

> The fact that humans appeared to have a stronger preference for raw data files than robots is somewhat counter-intuitive. However, in context it makes some sense, as the civil authority on Chapek 9 has an unfortunate tendency to kill and dissect humans when they are identified. As such it seems most likely that the human participants did not

respond honestly to the question, so as to avoid potentially undesirable consequences. This should be considered to be a substantial methodological weakness.

This could be classified as a rather extreme example of a reactivity effect, I suppose. Obviously, in this case the problem is severe enough that the study is more or less worthless as a tool for understanding the difference preferences among humans and robots. However, I hope this illustrates the difference between getting a statistically significant result (our null hypothesis is rejected in favour of the alternative), and finding something of scientific value (the data tell us nothing of interest about our research hypothesis due to a big methodological flaw).

9.2.3 Postscript

I later found out the data were made up, and I'd been watching cartoons instead of doing work.

9.3 The continuity correction

An optional adjustment for chi-square tests with 1 degree of freedom is the continuity correction, sometimes called the Yates correction. Remember what I pointed out earlier: the χ^2 test is based on an approximation, specifically on the assumption that the binomial distribution starts to look like a normal distribution for large N. One problem with this is that it often doesn't quite work, especially when you've only got 1 degree of freedom (e.g., when you're doing a test of independence on a 2 × 2 contingency table). The main reason for this is that the true sampling distribution for the X^2 statistic is actually discrete (because you're dealing with categorical data!) but the χ^2 distribution is continuous. This can introduce systematic problems. Specifically, when N is small and when $df = 1$, the goodness-of-fit statistic tends to be "too big", meaning that you actually have a bigger α value than you think (or, equivalently, the p values are a bit too small).

Yates (1934) suggested a simple fix, in which you redefine the goodness-of-fit statistic as:

$$\chi^2 = \sum_i \frac{(|E_i - O_i| - 0.5)^2}{E_i}$$

Basically, he just subtracts off 0.5 everywhere.
As far as I can tell from reading Yates' paper, the correction is basically a hack. It's not derived from any principled theory. Rather, it's based on an examination of the behaviour of the test, and observing that the corrected version seems to work better. You can specify this correction in JASP from a check box in the 'Statistics' options, where it is called ' χ^2 continuity correction'.

9.4 Effect size

As we discussed earlier, it's becoming commonplace to ask researchers to report some measure of effect size. So, let's suppose that you've run your chi-square test, which turns out to be significant. So you now know that there is some association between your variables (independence test) or some deviation from the specified probabilities (goodness-of-fit test). Now you want to report a measure of effect size. That is, given that there is an association or deviation, how strong is it?

There are several different measures that you can choose to report, and several different tools that you can use to calculate them. I won't discuss all of them but will instead focus on the most commonly reported measures of effect size.

By default, the two measures that people tend to report most frequently are the ϕ statistic and the somewhat superior version, known as Cramér's V.

Mathematically, they're very simple. To calculate the ϕ statistic, you just divide your X^2 value by the sample size, and take the square root:

$$\phi = \sqrt{\frac{X^2}{N}}$$

The idea here is that the ϕ statistic is supposed to range between 0 (no association at all) and 1 (perfect association), but it doesn't always do this when your contingency table is bigger than 2 × 2, which is a total pain. For bigger tables it's actually possible to obtain $\phi > 1$, which is pretty unsatisfactory. So, to correct for this, people usually prefer to report the V statistic proposed by Cramer (1946). It's a pretty simple adjustment to ϕ. If you've got a contingency table with r rows and c columns, then define $k = \min(r, c)$ to be the smaller of the two values. If so, then Cramér's V statistic is

$$V = \sqrt{\frac{X^2}{N(k-1)}}$$

And you're done. This seems to be a fairly popular measure, presumably because it's easy to calculate, and it gives answers that aren't completely silly. With Cramer's V, you know that the value really does range from 0 (no association at all) to 1 (perfect association).

9.5 Assumptions of the test(s)

All statistical tests make assumptions, and it's usually a good idea to check that those assumptions are met. For the chi-square tests discussed so far in this chapter, the assumptions are:

- -Expected frequencies are sufficiently large. Remember how in the previous section we saw that the χ^2 sampling distribution emerges because the binomial distribution is pretty similar to a normal distribution? Well, this approximation is only reliable when the number of observations is sufficiently large. What that means in practice is that all of the expected frequencies need to be reasonably big. How big is reasonably big? Opinions differ, but the default assumption seems to be that you generally would like to see all your expected frequencies larger than about 5, though for larger tables you would probably be okay if at least 80% of the the expected frequencies are above 5 and none of them are below 1. However, from what I've been able to discover (e.g., Cochran 1954) these seem to have been proposed as rough guidelines, not hard and fast rules, and they seem to be somewhat conservative (Larntz 1978).
- -Data are independent of one another. One somewhat hidden assumption of the chi-square test is that you have to genuinely believe that the observations are independent. Here's what I mean. Suppose I'm interested in proportion of babies born at a particular hospital that are boys. I walk around the maternity wards and observe 20 girls and only 10 boys. Seems like a pretty convincing difference, right? But later on, it turns out that I'd actually walked into the same ward 10 times and in fact I'd only seen 2 girls and 1 boy. Not as convincing, is it? My original 30 observations were massively non-independent, and were only in fact equivalent to 3 independent observations. Obviously this is an extreme (and extremely silly) example, but it illustrates the basic issue. Non-independence "stuffs things up". Sometimes it causes you to falsely reject the null, as the silly hospital example illustrates, but it can go the other way too. To give a slightly less stupid example, let's consider what would happen if I'd done the cards experiment slightly differently Instead of asking 200 people to try to imagine sampling one card at random, suppose I asked 50 people to select 4 cards. One possibility would be that everyone selects one heart, one club, one diamond and one spade (in keeping with the "representativeness heuristic"; Tversky & Kahneman 1974). This is highly non-random behaviour from people, but in this case I would get an observed frequency of 50 for all four suits. For this example the fact that the observations are non-independent (because the four cards that you pick will be related to each other) actually leads to the opposite effect, falsely retaining the null.

If you happen to find yourself in a situation where independence is violated, it may be possible to use the nonparametric tests, such as the McNemar test or the Cochran test. Similarly, if your expected cell counts are too small, check out the Fisher exact test.

9.6 Summary

The key ideas discussed in this chapter are:

- -The χ^2 (chi-square) goodness-of-fit test (Section 9.1) is used when you have a table of observed frequencies of different categories, and the null hypothesis gives you a set of "known" probabilities to compare them to.

- -The χ^2 (chi-square) test of independence (Section 9.2) is used when you have a contingency table (cross-tabulation) of two categorical variables. The null hypothesis is that there is no relationship or association between the variables.
- -Effect size for a contingency table can be measured in several ways (Section 9.4). In particular we noted the Cramér's V statistic.
- -Both versions of the Pearson test rely on two assumptions: that the expected frequencies are sufficiently large, and that the observations are independent (Section 9.5). Various nonparametric tests can be used for certain kinds of violations of independence or count assumptions.

If you're interested in learning more about categorical data analysis a good first choice would be Agresti (1996) which, as the title suggests, provides an Introduction to Categorical Data Analysis. If the introductory book isn't enough for you (or can't solve the problem you're working on) you could consider Agresti (2002), Categorical Data Analysis. The latter is a more advanced text, so it's probably not wise to jump straight from this book to that one.

---

1. [1] Also sometimes referred to as "chi-squared"
2. [2] A vector is a sequence of data elements of the same basic type
3. [3] One could also take the absolute value, but there are a lot of reasons not to do this...mostly, it comes down to calculus, which we'll not discuss here.
4. [4] If you rewrite the equation for the goodness-of-fit statistic as a sum over $k-1$ independent things you get the "proper" sampling distribution, which is chi-square with $k-1$ degrees of freedom. It's beyond the scope of an introductory book to show the maths in that much detail. All I wanted to do is give you a sense of why the goodness-of-fit statistic is associated with the chi-square distribution.

5. [5] I feel obliged to point out that this is an over-simplification. It works nicely for quite a few situations, but every now and then we'll come across degrees of freedom values that aren't whole numbers. Don't let this worry you too much; when you come across this just remind yourself that "degrees of freedom" is actually a bit of a messy concept, and that the nice simple story that I'm telling you here isn't the whole story. For an introductory class it's usually best to stick to the simple story, but I figure it's best to warn you to expect this simple story to fall apart. If I didn't give you this warning you might start getting confused when you see $df = 3.4$ or something, (incorrectly) thinking that you had misunderstood something that I've taught you rather than (correctly) realising that there's something that I haven't told you.
   [6] In practice, the sample size isn't always fixed. For example, we might run the experiment over a fixed period of time and the number of people participating depends on how many people show up. That doesn't matter for the current purposes.

6. [7] Well, sort of. The conventions for how statistics should be reported tend to differ somewhat from discipline to discipline. I've tended to stick with how things are done in psychology, since that's what I do. But the general principle of providing enough information to the reader to allow them to check your results is pretty universal, I think.
   [8] To some people, this advice might sound odd, or at least in conflict with the "usual" advice on how to write a technical report. Very typically, students are told that the "results" section of a report is for describing the data and reporting statistical analysis, and the "discussion" section is for providing interpretation. That's true as far as it goes, but I think people often interpret it way too literally. The way I usually approach it is to provide a quick and simple interpretation of the data in the results section, so that my reader understands what the data are telling us. Then, in the discussion, I try to tell a bigger story about how my results fit with the rest of the scientific literature. In short, don't let the "interpretation goes in the discussion" advice turn your results section into incomprehensible garbage. Being understood by your reader is much more important.

7. [9] Complicating matters, the G-test is a special case of a whole class of tests that are known as likelihood ratio tests. I don't cover LRTs in this book, but they are quite handy things to know about.
8. [10] A technical note. The way I've described the test pretends that the column totals are fixed (i.e., the researcher intended to survey 87 robots and 93 humans) and the row totals are random (i.e., it just turned out that 28 people chose the puppy). To use the terminology from my mathematical statistics textbook (Hogg, McKean, and Craig 2005), I should technically refer to this situation as a chi-square test of homogeneity and reserve the term chi-square test of independence for the situation where both the row and column totals are random outcomes of the experiment. In the initial drafts of this book that's exactly what I did. However, it turns out that these two tests are identical, and so I've collapsed them together.
9. [11] Technically, E_{ij} here is an estimate, so I should probably write it $\hat{E}_{ij}$. But since no-one else does, I won't either.
10. [12] A problem many of us worry about in real life.

This page titled 9.1: Categorical data analysis is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 9: Categorical data analysis by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

9.2: Chi-squared Test

The Chi-squared test is used **only with categorical variables**. This test assesses whether the categorical variables are related and is used to make inferences about the likelihood that there is a certain pattern within the data (e.g., is it more likely that women who live alone have cats when compared to men who live alone).

Creating Frequency Tables

JASP can analyse raw data directly, with one row per participant and a column for each categorical variable. You do not need to convert the data into frequency counts first. The steps below show an alternative approach using summarised counts.

How do I do that?

1. In the **Descriptives** menu move one variable to the 'Variables' box, and the other variable to the 'Split' box. Where you put each variable does not matter but it does determine how your frequency table looks. We recommend swapping the variables to see which table layout you prefer.

2. In the **Tables** menu, select 'Frequency tables'.

image

- Doing so will give you the **Frequency Table** for your variables:

  image

3. If you choose to use summarised counts, create a new dataset using the frequency data, with each row representing one combination of variables.

image

Conducting a Chi-squared Test

Your data should now be in a form suitable for conducting the chi-squared test.

How do I do that?

1. In the **Frequencies** menu select 'Contingency Tables' from the 'Classical' list.

image

A new window will open, like so:

image

3. Move one variable of interest to the Rows box and the other to the Columns box. If you are using summarized counts, also move the frequency variable to the Counts box. If each row represents one participant, leave Counts empty.

image

4. In the **Statistics** menu, select:

image

- χ^2 (chi-squared), χ^2 continuity correction, and Likelihood ratio. The last two are statistics with slight adjustments.

- **Phi and Cramer's V** – These are measures of association. Phi is commonly used for 2 × 2 tables, while Cramer's V can also be used for larger tables.
- **Odds ratio** – selecting this option will give you **Fisher's exact test** which is useful when expected cell counts are small.

5. In the 'Cells' menu select 'Expected' under **Counts**, and 'Row' and 'Total' under **Percentages**. This will then give you all the outputs you need to interpret the Chi-squared analysis.

image

This page titled 9.2: Chi-squared Test is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss & Stanislava Naneva via source content that was edited to the style and standards of the LibreTexts platform.

10: COMPARING MEANS - Z-TESTS AND T-TESTS

10.1: Comparing two means
10.2: Independent t-test and Mann-Whitney U
10.3: Paired t-test and Wilcoxon Signed Ranks
10.4: Worked Example- Calculating an Independent Samples t-Test

This page titled 10: Comparing Means - z-Tests and t-Tests is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

10.1: Comparing two means

In a previous chapter, we covered the situation when your outcome variable is nominal scale and your predictor variable is also nominal scale. Lots of real world situations have that character, and so you'll find that chi-square tests are quite widely used. However, you're much more likely to find yourself in a situation where your outcome variable is interval scale or higher, and what you're interested in is whether the average value of the outcome variable is higher in one group or another. For instance, a psychologist might want to know if anxiety levels are higher among parents than non-parents, or if working memory capacity is reduced by listening to music (relative to not listening to music). In a medical context we might want to know if a new drug increases or decreases blood pressure. An agricultural scientist might want to know whether adding phosphorus to Australian native plants will kill them. [1] In all these situations our outcome variable is a fairly continuous, interval or ratio scale variable, and our predictor is a binary "grouping" variable. In other words, we want to compare the means of the two groups.

The standard answer to the problem of comparing means is to use a t-test, of which there are several varieties depending on exactly what question you want to solve. As a consequence, the majority of this chapter focuses on different types of t-test: one sample t-tests, independent samples t-tests, and paired samples t-tests. We'll then talk about one sided tests and, after that, we'll talk a bit about Cohen's d, which is the standard measure of effect size for a t-test. The later sections of the chapter focus on the assumptions of the t-tests, and possible remedies if they are violated. However, before discussing any of these useful things, we'll start with a discussion of the z-test.

10.1 The one-sample z-test

In this section I'll describe one of the most useless tests in all of statistics: the z-test. Seriously - this test is almost never used in real life. Its only real purpose is that, when teaching statistics, it's a very convenient stepping stone along the way towards the t-test, which is probably the most (over)used tool in all statistics.

10.1.1 The inference problem that the test addresses

To introduce the idea behind the z-test, let's use a simple example. A friend of mine, Dr. Zeppo, grades his introductory statistics class on a curve. Let's suppose that the average grade in his class is 67.5 , and the standard deviation is 9.5 . Of his many hundreds of students, it turns out that 20 of them also take psychology classes. Out of curiosity, I find myself wondering if the psychology students tend to get the same grades as everyone else (i.e., mean 67.5) or do they tend to score higher or lower? He emails me the zeppo.csv file, which I use to look at the grades of those students in JASP (stored in the variable x):

```
50 60 60 64 66 66 67 69 70 74 76 76 77 79 79 79 81 82 82 89
```

Then I calculate the mean in 'Descriptives' - 'Descriptive Statistics'. The mean value is 72.3.
Hmm. It might be that the psychology students are scoring a bit higher than normal. That sample mean of $\bar{X} = 72.3$is a fair bit higher than the hypothesised population mean of $\mu = 67.5$but, on the other hand, a sample size of $N = 20$isn't all that big. Maybe it's pure chance.

To answer the question, it helps to be able to write down what it is that I think I know. Firstly, I know that the sample mean is $\bar{X} = 72.3$. If I'm willing to assume that the psychology students have the same standard deviation as the rest of the class then I can say that the population standard deviation is $\sigma = 9.5$. I'll also assume that since Dr Zeppo is grading to a curve, the psychology student grades are normally distributed.

Next, it helps to be clear about what I want to learn from the data. In this case my research hypothesis relates to the population mean $\mu$for the psychology student grades, which is unknown. Specifically, I want to know if $\mu = 67.5$or not. Given that this is what I know, can we devise a hypothesis test to solve our problem? The data, along with the hypothesised distribution from which they are thought to arise, are shown in Figure 10.1. Not entirely obvious what the right answer is, is it? For this, we are going to need some statistics.

10.1.2 Constructing the hypothesis test

The first step in constructing a hypothesis test is to be clear about what the null and alternative hypotheses are. This isn't too hard to do. Our null hypothesis, H_0, is that the true population mean $\mu$for psychology student grades is 67.5%, and our alternative hypothesis is that the population mean isn't 67.5%.

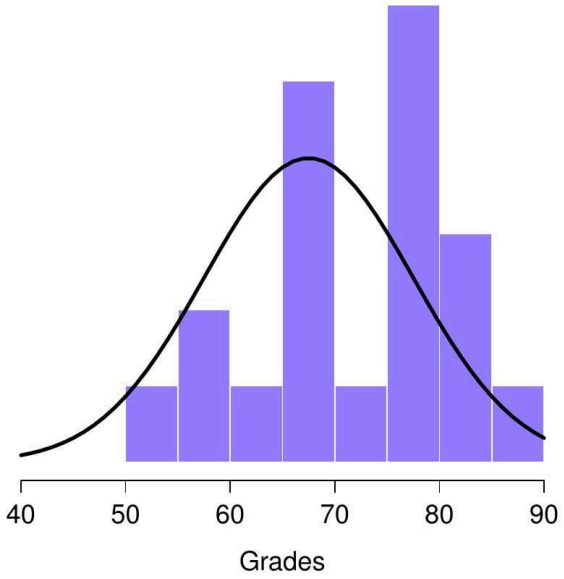


Figure 10.1: The theoretical distribution (solid line) from which the psychology student grades (bars) are supposed to have been generated.

If we write this in mathematical notation, these hypotheses become:

$$
\begin{aligned}
H_0: &\quad \mu = 67.5 \\
H_1: &\quad \mu \neq 67.5
\end{aligned}
$$

though to be honest this notation doesn't add much to our understanding of the problem, it's just a compact way of writing down what we're trying to learn from the data. The null hypotheses H_0 and the alternative hypothesis H_1 for our test are both illustrated in Figure 10.2. In addition to providing us with these hypotheses, the scenario outlined above provides us with a fair amount of background knowledge that might be useful. Specifically, there are two special pieces of information that we can add:

- 1. The psychology grades are normally distributed.
- 2. The true standard deviation of these scores $\sigma$is known to be 9.5.

For the moment, we'll act as if these are absolutely trustworthy facts. In real life, this kind of absolutely trustworthy background knowledge doesn't exist, and so if we want to rely on these facts we'll just
have make the assumption that these things are true. However, since these assumptions may or may not be warranted, we might need to check them. For now though, we'll keep things simple.

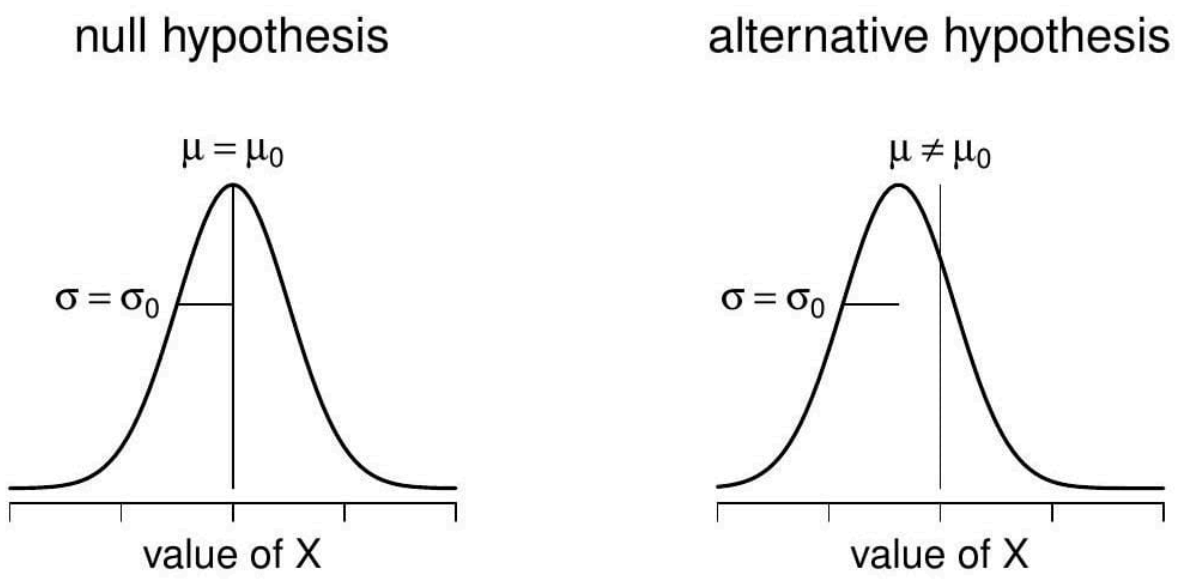


Figure 10.2: Graphical illustration of the null and alternative hypotheses assumed by the one sample z-test (the two sided version, that is). The null and alternative hypotheses both assume that the population distribution is normal, and additionally assumes that the population standard deviation is known (fixed at some value σ_0). The null hypothesis (left) is that the population mean μ is equal to some specified value μ_0. The alternative hypothesis is that the population mean differs from this value, $\mu \neq \mu_0$.

The next step is to figure out what we would be a good choice for a diagnostic test statistic, something that would help us discriminate between H_0 and H_1. Given that the hypotheses all refer to the population mean μ, you'd feel pretty confident that the sample mean $\bar{X}$ would be a pretty useful place to start. What we could do is look at the difference between the sample mean $\bar{X}$ and the value that the null hypothesis predicts for the population mean. In our example that would mean we calculate $\bar{X} - 67.5$. More generally, if we let μ_0 refer to the value that the null hypothesis claims is our population mean, then we'd want to calculate

$$\bar{X} - \mu_0$$

If this quantity equals or is very close to 0 , things are looking good for the null hypothesis. If this quantity is a long way away from 0, then it's looking less likely that the null hypothesis is worth retaining. But how far away from zero should it be for us to reject H_0 ?

To figure that out we need to be a bit more sneaky, and we'll need to rely on those two pieces of background knowledge that I wrote down previously; namely that the raw data are normally distributed and that we know the value of the population standard deviation σ. If the null hypothesis is actually true, and the true mean is μ_0, then these facts together mean that we know the complete population
distribution of the data: a normal distribution with mean μ_0 and standard deviation σ. Adopting the notation from, a statistician might write this as:

$$X \sim \text{Normal}\left(\mu_0, \sigma^2\right)$$

Okay, if that's true, then what can we say about the distribution of $\bar{X}$? Well, as we discussed earlier, the sampling distribution of the mean $\bar{X}$ is also normal, and has mean μ. But the standard deviation of this sampling distribution $\text{se}(\bar{X})$, which is called the standard error of the mean, is

$$\text{se}(\bar{X}) = \frac{\sigma}{\sqrt{N}}$$

In other words, if the null hypothesis is true then the sampling distribution of the mean can be written as follows:

$$\bar{X} \sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{N}\right) \tag{10.1.1}$$

Now comes the trick. What we can do is convert the sample mean $\bar{X}$ into a standard score. This is conventionally written as z, but for now I'm going to refer to it as $z_{\bar{X}}$. (The reason for using this expanded notation is to help you remember that we're calculating a

standardised version of a sample mean, not a standardised version of a single observation, which is what a z-score usually refers to). When we do so the z-score for our sample mean is

$$z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\text{se}(\bar{X})}$$

or, equivalently

$$z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}}$$

This z-score is our test statistic. Under the null hypothesis and the assumptions of this test, this statistic has a standard normal distribution:

$$z_{\bar{X}} \sim \text{Normal}(0, 1)$$

In other words, regardless of what scale the original data are on, the z-statistic itself always has the same interpretation: it's equal to the number of standard errors that separate the observed sample mean $\bar{X}$ from the population mean μ_0 predicted by the null hypothesis. Better yet, regardless of what the population parameters for the raw scores actually are, the 5% critical regions for the z-test are always the same, as illustrated in Figure 10.3. And what this meant, way back in the days where people did all their statistics by hand, is that someone could publish a table like this:

desired αlevel	Two-sided critical value	One-sided critical value
.10	1.644854	1.281552
.05	1.959964	1.644854
.01	2.575829	2.326348
.001	3.290527	3.090232

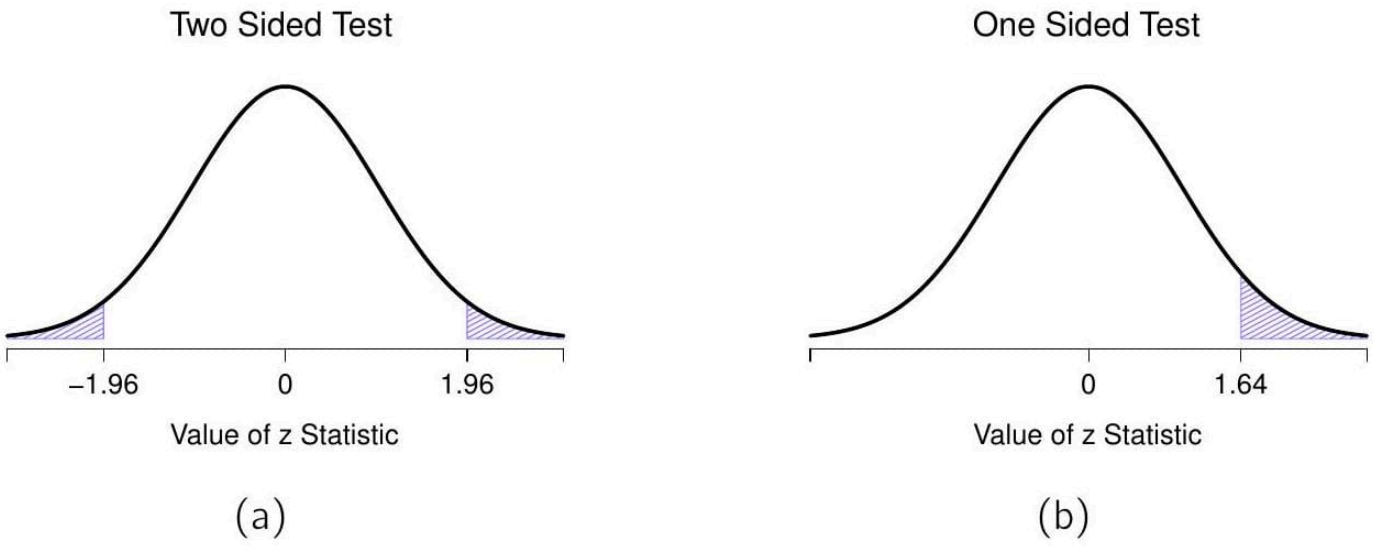


Figure 10.3: Rejection regions for the two-sided z-test (panel a) and the one-sided z-test (panel b).

This, in turn, meant that researchers could calculate their z-statistic by hand and then look up the critical value in a text book.

10.1.3 A worked example, by hand

Now, as I mentioned earlier, the z-test is almost never used in practice. It's so rarely used in real life that earlier versions of JASP

didn't even have a built in function for it [2]. However, the test is so incredibly simple that it's really easy to do one manually. Let's go back to the data from Dr Zeppo's class. Having loaded the grades data, the first thing I need to do is calculate the sample mean, which I've already done (72.3). We already have the known population standard deviation $(\sigma = 9.5)$, and the value of the population mean that the null hypothesis specifies $(\mu_0 = 67.5)$, and we know the sample size $(\mathrm{N} = 20)$.

Next, let's calculate the (true) standard error of the mean (easily done with a calculator):

$$\begin{aligned} \mathrm{se}(\bar{X}) &= \frac{\sigma}{\sqrt{N}} \\ &= \frac{9.5}{\sqrt{20}} \\ &= 2.124265 \end{aligned}$$

From this, we calculate our z-score:

$$\begin{aligned} z_{\bar{X}} &= \frac{\bar{X} - \mu_0}{\mathrm{se}(\bar{X})} \\ &= \frac{72.3 - 67.5}{2.124265} \\ &= 2.259606. \end{aligned} \tag{10.1.2}$$

At this point, we would traditionally look up the value 2.26 in our table of critical values. Our original hypothesis was two-sided (we didn't really have any theory about whether psych students would be better or worse at statistics than other students) so our hypothesis test is two-sided (or two-tailed) also. Looking at the little table that I showed earlier, we can see that 2.26 is bigger than the critical value of 1.96 that would be required to be significant at $\alpha = .05$, but smaller than the value of 2.58 that would be required to be significant at a level of $\alpha = .01$. Therefore, we can conclude that we have a significant effect, which we might write up by saying something like this:

> With a mean grade of 72.3 in the sample of psychology students, and assuming a true population standard deviation of 9.5, we can conclude that the psychology students have significantly different statistics scores to the class average ($z = 2.26, N = 20, p < .05$).

10.1.4 Assumptions of the z-test

As I've said before, all statistical tests make assumptions. Some tests make reasonable assumptions, while other tests do not. The test I've just described, the one sample z-test, makes three basic assumptions. These are:

- -Normality. As usually described, the z-test assumes that the true population distribution is normal. [3] This is often a pretty reasonable assumption, and it's also an assumption that we can check if we feel worried about it.
- -Independence. The second assumption of the test is that the observations in your data set are not correlated with each other, or related to each other in some funny way. This isn't as easy to check statistically, it relies a bit on good experimental design. An obvious (and silly) example of something that violates this assumption is a data set where you "copy" the same observation over and over again in your data file so that you end up with a massive "sample size", which consists of only one genuine observation. More realistically, you have to ask yourself if it's really plausible to imagine that each observation is a completely random sample from the population that you're interested in. In practice this assumption is never met, but we try our best to design studies that minimise the problems of correlated data.

- -Known standard deviation. The third assumption of the z-test is that the true standard deviation of the population is known to the researcher. This is just silly. In no real world data analysis problem do you know the standard deviation $\sigma$of some population but are completely ignorant about the mean μ. In other words, this assumption is always wrong.

In view of the stupidity of assuming that $\sigma$is known, let's see if we can live without it. This takes us out of the dreary domain of the z-test, and into the magical kingdom of the t-test!

10.2 The one-sample t-test

After some thought, I decided that it might not be safe to assume that the psychology student grades necessarily have the same standard deviation as the other students in Dr Zeppo's class. After all, if I'm entertaining the hypothesis that they don't have the

same mean, then why should I believe that they absolutely have the same standard deviation? In view of this, I should really stop assuming that I know the true value of σ. This violates the assumptions of my z-test, so in one sense I'm back to square one. However, it's not like I'm completely bereft of options. After all, I've still got my raw data, and those raw data give me an estimate of the population standard deviation, which is 9.52. In other words, while I can't say that I know that $\sigma = 9.5$, I can say that $\hat{\sigma} = 9.52$.

Okay, cool. The obvious thing that you might think to do is run a z-test, but using the estimated standard deviation of 9.52 instead of relying on my assumption that the true standard deviation is 9.5. And you probably wouldn't be surprised to hear that this would still give us a significant result. This approach is close, but it's not quite correct. Because we are now relying on an estimate of the population standard deviation we need to make some adjustment for the fact that we have some uncertainty about what the true population standard deviation actually is. Maybe our data are just a fluke ... maybe the true population standard deviation is 11, for instance. But if that were actually true, and we ran the z-test assuming $\sigma = 11$, then the result would end up being non-significant. That's a problem, and it's one we're going to have to address.

10.2.1 Introducing the t-test

This ambiguity is annoying, and it was resolved in 1908 by a guy called William Sealy Gosset (Student 1908), who was working as a chemist for the Guinness brewery at the time (see Box 1987). Because Guinness took a dim view of its employees publishing statistical analysis (apparently they felt it was a trade secret), he published the work under the pseudonym "A Student" and, to this day, the full name of the t-test is actually Student's t-test.

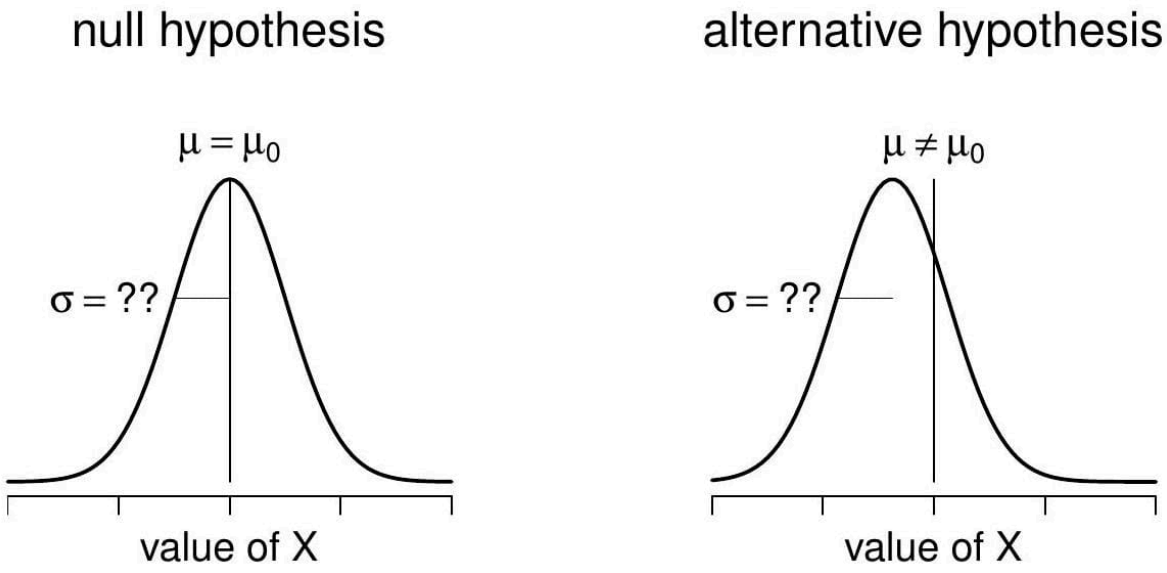


Figure 10.4: Graphical illustration of the null and alternative hypotheses assumed by the (two sided) one sample t-test. Note the similarity to the z-test (Figure 10.2). The null hypothesis is that the population mean $\mu$is equal to some specified value μ_0, and the alternative hypothesis is that it is not. Like the z-test, we assume that the data are normally distributed, but we do not assume that the population standard deviation $\sigma$is known in advance.

The key thing that Gosset figured out is how we should accommodate the fact that we aren't completely sure what the true standard deviation is. [4] The answer is that it subtly changes the sampling distribution. In the t-test our test statistic, now called a t-statistic, is calculated in exactly the same way I mentioned above. If our null hypothesis is that the true mean is μ, but our sample has mean $\bar{X}$ and our estimate of the population standard deviation is $\hat{\sigma}$, then our $t$statistic is:

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{N}}$$

The only thing that has changed in the equation is that instead of using the known true value σ, we use the estimate $\hat{\sigma}$. And if this estimate has been constructed from Nobservations, then the sampling distribution turns into a t-distribution with $N - 1$degrees of freedom (df). The $t$distribution is very similar to the normal distribution, but has "heavier" tails, as discussed earlier in and illustrated in Figure 10.5. Notice, though, that as df gets larger, the t-distribution starts to look identical to the standard normal

distribution. This is as it should be: if you have a sample size of $N = 70{,}000{,}000$ then your "estimate" of the standard deviation would be pretty much perfect, right?

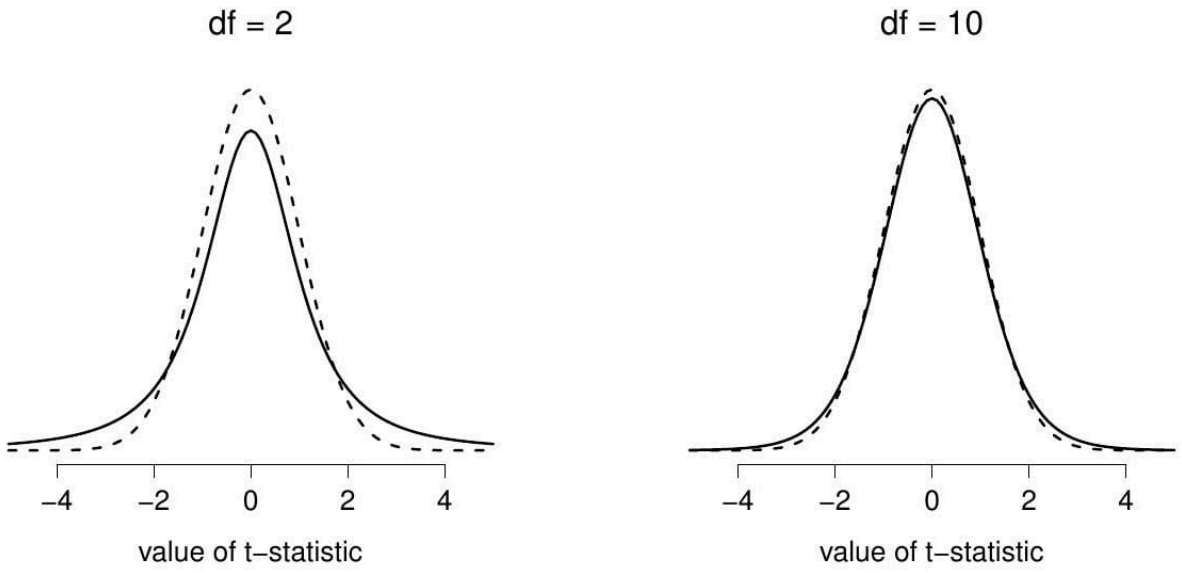


Figure 10.5: The $t$distribution with 2 degrees of freedom (left) and 10 degrees of freedom (right), with a standard normal distribution (i.e., mean 0 and std dev 1) plotted as dotted lines for comparison purposes. Notice that the $t$distribution has heavier tails (leptokurtic: higher kurtosis) than the normal distribution; this effect is quite exaggerated when the degrees of freedom are very small, but negligible for larger values. In other words, for large $df$the $t$distribution is essentially identical to a normal distribution.

So, you should expect that for large N, the t-test would behave exactly the same way as a z-test. And that's exactly what happens!

10.2.2 Doing the test in JASP

As you might expect, the mechanics of the t-test are almost identical to the mechanics of the z-test. So there's not much point in going through the tedious exercise of showing you how to do the calculations by hand. It's pretty much identical to the calculations that we did earlier, except that we use the estimated standard deviation and then we test our hypothesis using the $t$distribution rather than the normal distribution. And so instead of going through the calculations in tedious detail for a second time, I'll jump straight to showing you how t-tests are actually done. JASP comes with a dedicated analysis for t-tests that is very flexible (it can run lots of different kinds of t-tests). It's pretty straightforward to use; all you need to do is specify 'T-Tests' - 'One Sample T-Test', move the variable you are interested in (x) across into the 'Variables' box, and type in the mean value for the null hypothesis ('67.5') in the 'Test value' box. Easy enough. See Figure 10.6, which, amongst other things that we will get to in a moment, gives you a t -test statistic $= 2.25$, with 19 degrees of freedom and an associated p-value of 0.036 .

It is also easy to calculate a 95% confidence interval for our sample mean. If you select the 'Location parameter' and its associated 'Confidence interval' option under 'Additional Statistics', you'll see in the JASP output that the 'Mean difference' is 4.800 with 95% CI equal to [0.344, 9.256]. This simply means that we are 95% confident that the population mean minus the hypothesized value of 67.5 is between 0.344 and 9.256. If we add these "endpoints" to the hypothesized mean, we get a 95% CI of [67.5+0.344, 67.5+9.256], or said differently, [67.844, 76.756]. If this isn't clear, don't worry. We'll explain a bit more about this in the next section.

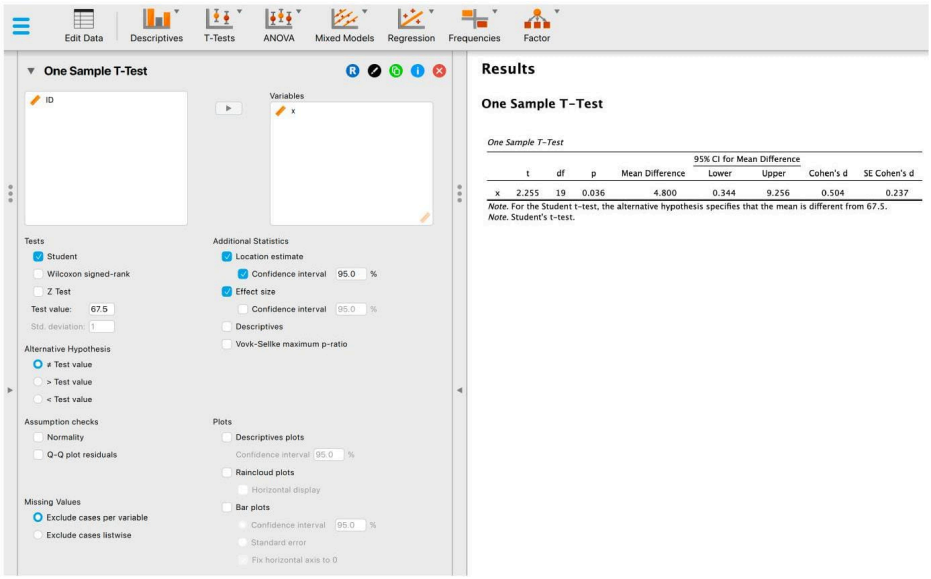


Figure 10.6: JASP does the one-sample t-test.

Now, what do we do with all this output? Well, since we're pretending that we actually care about my toy example, we're overjoyed to discover that the result is statistically significant (i.e. $p$value below .05). We could report the result by saying something like this:

With a mean grade of 72.3, the psychology students scored slightly higher than the average grade of $67.5(t(19) = 2.25, p < .05)/); the 95$is shorthand notation for a t-statistic that has 19 degrees of freedom. That said, it's often the case that people don't report the confidence interval, or do so using a much more compressed form than I've done here. For instance, it's not uncommon to see the confidence interval included as part of the stat block, like this:

$$t(19) = 2.25, p < .05, \text{CI}_{95} = [67.8, 76.8]$$

With that much jargon crammed into half a line, you know it must be really smart. [5]

10.2.3 Assumptions of the one sample t-test

Okay, so what assumptions does the one-sample t-test make? Well, since the t-test is basically a z-test with the assumption of known standard deviation removed, you shouldn't be surprised to see that it makes the same assumptions as the z-test, minus the one about the known standard deviation. That is

- -Normality. We're still assuming that the population distribution is normal [6], and as noted earlier, there are standard tools that you can use to check to see if this assumption is met, and other tests you can do in it's place if this assumption is violated.
- -Independence. Once again, we have to assume that the observations in our sample are generated independently of one another. See the earlier discussion about the z-test for specifics.

Overall, these two assumptions aren't terribly unreasonable, and as a consequence the one-sample t-test is pretty widely used in practice as a way of comparing a sample mean against a hypothesised population mean.

10.3 The independent samples t-test (Student test)

Although the one sample t-test has its uses, it's not the most typical example of a t-test [7]. A much more common situation arises when you've got two different groups of observations. In psychology, this tends to correspond to two different groups of participants, where each group corresponds to a different condition in your study. For each person in the study you measure some outcome variable

of interest, and the research question that you're asking is whether or not the two groups have the same population mean. This is the situation that the independent samples t-test is designed for.

10.3.1 The data

Suppose we have 33 students taking Dr Harpo's statistics lectures, and Dr Harpo doesn't grade to a curve. Actually, Dr Harpo's grading is a bit of a mystery, so we don't really know anything about what the average grade is for the class as a whole. There are two tutors for the class, Anastasia and Bernadette. There are $N_1 = 15$students in Anastasia's tutorials, and $N_2 = 18$in Bernadette's tutorials. The research question I'm interested in is whether Anastasia or Bernadette is a better tutor, or if it doesn't make much of a difference. Dr Harpo emails me the course grades, in the harpo.csv file. As usual, I'll load the file into JASP and have a look at what variables it contains - there are three variables, ID, grade and tutor. Not surprisingly, the grade variable contains each student's grade. The tutor variable indicates who each student's tutor was - either Anastasia or Bernadette.

We can calculate means and standard deviations, using the 'Descriptives' - 'Descriptive Statistics' analysis (being sure to split by tutor). Here's a nice little summary table:

	mean	std dev	N
Anastasia's students	74.53	9.00	15
Bernadette's students	69.06	5.77	18

To give you a more detailed sense of what's going on here, I've plotted histograms (not in JASP, but using R) showing the distribution of grades for both tutors (Figure 10.7), as well as a simpler plot showing the means and corresponding confidence intervals for both groups of students (Figure 10.8).

10.3.2 Introducing the test

The independent samples t-test comes in two different forms, Student's and Welch's. The original Student t-test, which is the one I'll describe in this section, is the simpler of the two but relies on much more restrictive assumptions than the Welch t-test. Assuming for the moment that you want to run a two-sided test, the goal is to determine whether two "independent samples" of data are drawn from populations with the same mean (the null hypothesis) or different means (the alternative hypothesis). When we say "independent" samples, what we really mean here is that there's no special relationship between observations in the two samples. This probably doesn't make a lot of sense right now, but it will be clearer when we come to talk about the paired samples t-test later on. For now, let's just point out that if we have an experimental design where participants are randomly allocated to one of two groups, and we want to compare the two groups' mean performance on some outcome measure, then an independent samples t-test (rather than a paired samples t-test) is what we're after.

Okay, so let's let μ_1 denote the true population mean for group 1 (e.g., Anastasia's students), and μ_2 will be the true population mean for group 2 (e.g., Bernadette's students), [8] and as usual we'll let $\bar{X}_1$ and $\bar{X}_2$ denote the observed sample means for both of these groups. Our null hypothesis states that the two population means are identical ($\mu_1 = \mu_2$) and the alternative to this is that they are not $(\mu_1 \neq \mu_2)$. Written in mathematical notation, this is:

$$
\begin{aligned}
H_0: &\quad \mu_1 = \mu_2 \\
H_1: &\quad \mu_1 \neq \mu_2
\end{aligned}
$$

To construct a hypothesis test that handles this scenario we start by noting that if the null hypothesis is true, then the difference between the population means is exactly zero, $\mu_1 - \mu_2 = 0$.

As a consequence, a diagnostic test statistic will be based on the difference between the two sample means. Because if the null hypothesis is true, then we'd expect $\bar{X}_1 - \bar{X}_2$ to be pretty close to zero. However, just like we saw with our one-sample tests (i.e., the one-sample z-test and the one-sample t-test) we have to be precise about exactly how close to zero this difference should be. And the solution to the problem is more or less the same one. We calculate a standard error estimate (SE), just like last time, and then divide the difference between means by this estimate. So our t-statistic will be of the form:

$$
t = \frac{\bar{X}_1 - \bar{X}_2}{\mathrm{SE}}
$$

We just need to figure out what this standard error estimate actually is. This is a bit trickier than was the case for either of the two tests we've looked at so far, so we need to go through it a lot more carefully to understand how it works.

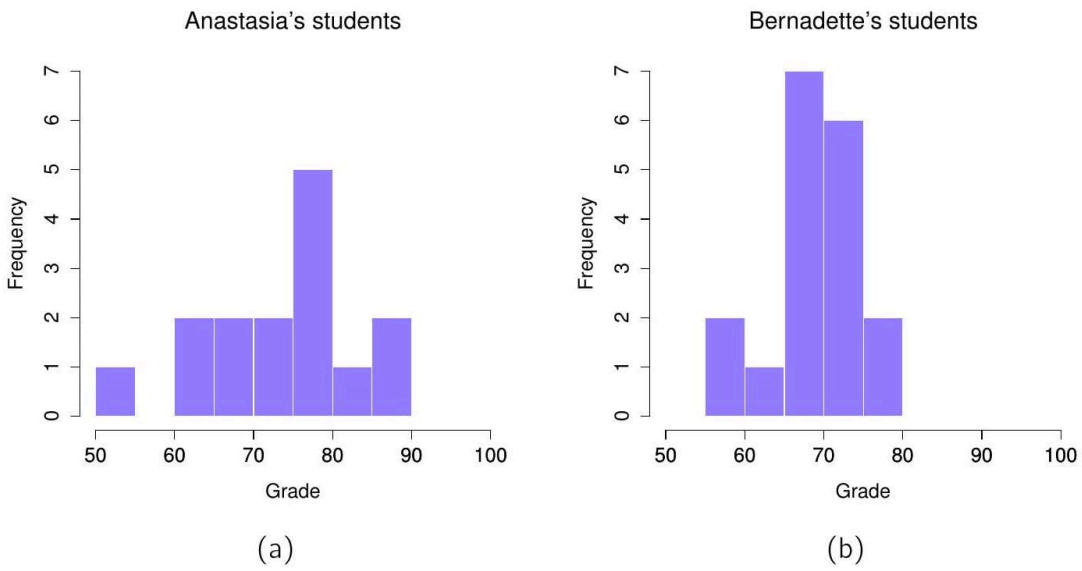


Figure 10.7: Histograms showing the distribution of grades for students in Anastasia's (panel a) and in Bernadette's (panel b) classes. Visually, these suggest that students in Anastasia's class may be getting slightly better grades on average, though they also seem a bit more variable.

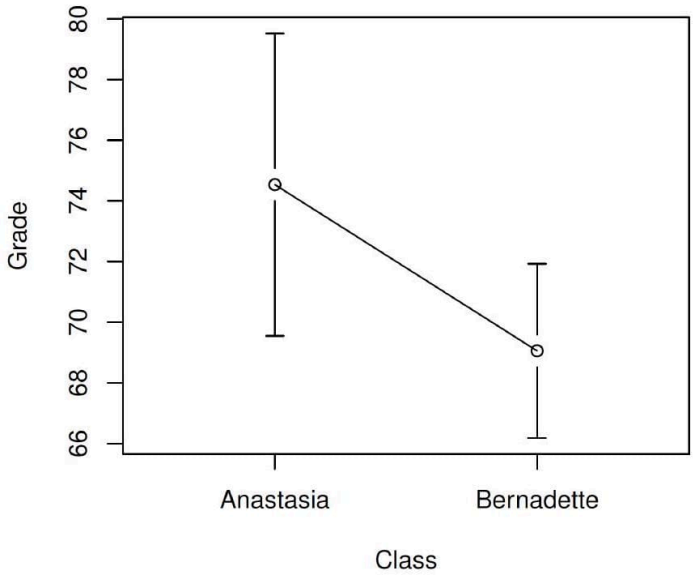


Figure 10.8: The plots show the mean grade for students in Anastasia's and Bernadette's tutorials. Error bars depict 95% confidence intervals around the mean. Visually, it does look like there's a real difference between the groups, though it's hard to say for sure.

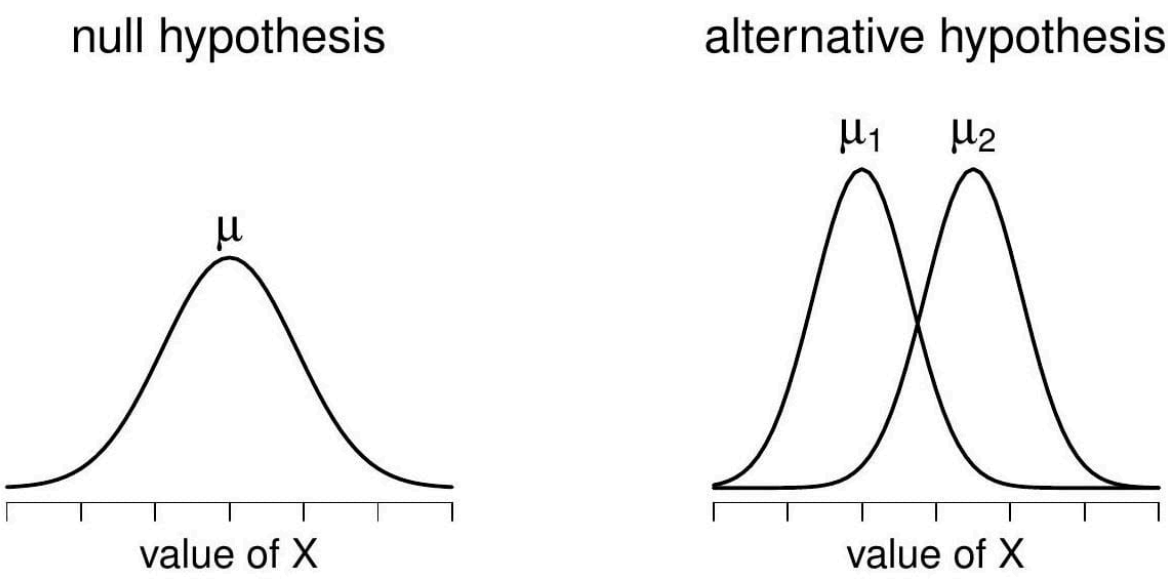


Figure 10.9: Graphical illustration of the null and alternative hypotheses assumed by the Student t-test. The null hypothesis assumes that both groups have the same mean μ, whereas the alternative assumes that they have different means μ_1 and μ_2. Notice that it is assumed that the population distributions are normal, and that, although the alternative hypothesis allows the group to have different means, it assumes they have the same standard deviation.

10.3.3 A "pooled estimate" of the standard deviation

In the original "Student t-test", we make the assumption that the two groups have the same population standard deviation. That is, regardless of whether the population means are the same, we assume that the population standard deviations are identical, $\sigma_1 = \sigma_2$. Since we're assuming that the two standard deviations are the same, we drop the subscripts and refer to both of them as σ. How should we estimate this? How should we construct a single estimate of a standard deviation when we have two samples? The answer is, basically, we average them. Well, sort of. Actually, what we do is take a weighed average of the variance estimates, which we use as our pooled estimate of the variance. The weight assigned to each sample is equal to the number of observations in that sample, minus 1.

Mathematically, we can write this as

$$w_1 = N_1 - 1$$
$$w_2 = N_2 - 1$$

Now that we've assigned weights to each sample we calculate the pooled estimate of the variance by taking the weighted average of the two variance estimates, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$

$$\hat{\sigma}_p^2 = \frac{w_1 \hat{\sigma}_1^2 + w_2 \hat{\sigma}_2^2}{w_1 + w_2}$$

Finally, we convert the pooled variance estimate to a pooled standard deviation estimate, by taking the square root.

$$\hat{\sigma}_p = \sqrt{\frac{w_1 \hat{\sigma}_1^2 + w_2 \hat{\sigma}_2^2}{w_1 + w_2}}$$

And if you mentally substitute $w_1 = N_1 - 1$and $w_2 = N_2 - 1$into this equation you get a very ugly looking formula. A very ugly formula that actually seems to be the "standard" way of describing the pooled standard deviation estimate. It's not my favourite way of thinking about pooled standard deviations, however. I prefer to think about it like this. Our data set actually corresponds to a set of Nobservations which are sorted into two groups. So let's use the notation X_{ik} to refer to the grade received by the i-th student in the k-th tutorial group. That is, X_{11} is the grade received by the first student in Anastasia's class, X_{21} is her second student, and so on. And we have two separate group means $\bar{X}_1$ and $\bar{X}_2$, which we could "generically" refer to using the notation $\bar{X}_k$, i.e., the mean grade for the k-th tutorial group. So far, so good. Now, since every single student falls into one of the two tutorials,

we can describe their deviation from the group mean as the difference

$$X_{ik} - \bar{X}_k$$

So why not just use these deviations (i.e., the extent to which each student's grade differs from the mean grade in their tutorial)? Remember, a variance is just the average of a bunch of squared deviations, so let's do that. Mathematically, we could write it like this

$$\frac{\sum_{ik} \left(X_{ik} - \bar{X}_k \right)^2}{N}$$

where the notation " $\sum_{ik}$ " is a lazy way of saying "calculate a sum by looking at all students in all tutorials", since each " ik" corresponds to one student. [a] But, as we saw in a previous chapter, calculating the variance by dividing by Nproduces a biased estimate of the population variance. And previously we needed to divide by $N-1$to fix this. However, as I mentioned at the time, the reason why this bias exists is because the variance estimate relies on the sample mean, and to the extent that the sample mean isn't equal to the population mean it can systematically bias our estimate of the variance. But this time we're relying on two sample means! Does this mean that we've got more bias? Yes, yes it does. And does this mean we now need to divide by $N-2$ instead of $N-1$, in order to calculate our pooled variance estimate? Why, yes

$$\hat{\sigma}_p^2 = \frac{\sum_{ik} \left(X_{ik} - \bar{X}_k \right)^2}{N-2}$$

Oh, and if you take the square root of this then you get $\hat{\sigma}_p$, the pooled standard deviation estimate. In other words, the pooled standard deviation calculation is nothing special. It's not terribly different to the regular standard deviation calculation.

10.3.4 Completing the test

Regardless of which way you want to think about it, we now have our pooled estimate of the
standard deviation. From now on, I'll drop the silly $p$subscript, and just refer to this estimate as $\hat{\sigma}$. Great. Let's now go back to thinking about the bloody hypothesis test, shall we? Our whole reason for calculating this pooled estimate was that we knew it would be helpful when calculating our standard error estimate. But standard error of what? In the one-sample t-test it was the standard error of the sample mean, $\mathrm{se}(\bar{X})$, and since $\mathrm{se}(\bar{X}) = \sigma/\sqrt{N}$ that's what the denominator of our t-statistic looked like. This time around, however, we have two sample means. And what we're interested in, specifically, is the the difference between the two $\bar{X}_1 - \bar{X}_2$. As a consequence, the standard error that we need to divide by is in fact the standard error of the difference between means.

As long as the two variables really do have the same standard deviation, then our estimate for the standard error is

$$\mathrm{se}\left(\bar{X}_1 - \bar{X}_2 \right) = \hat{\sigma} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

and our t-statistic is therefore

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\mathrm{se}\left(\bar{X}_1 - \bar{X}_2 \right)}$$

Just as we saw with our one-sample test, the sampling distribution of this t-statistic is a t-distribution (shocking, isn't it?) as long as the null hypothesis is true and all of the assumptions of the test are met. The degrees of freedom, however, is slightly different. As usual, we can think of the degrees of freedom to be equal to the number of data points minus the number of constraints. In this case, we have Nobservations (N_1 in sample 1, and N_2 in sample 2), and 2 constraints (the sample means). So the total degrees of freedom for this test are $N-2$.

10.3.5 Doing the test in JASP

Not surprisingly, you can run an independent samples t-test easily in JASP. The outcome variable for our test is the student grade, and the groups are defined in terms of the tutor for each class. So you probably won't be too surprised that all you have to do in JASP is go to the relevant analysis ('T-Tests' - 'Independent Samples T-Test') and move the grade variable across to the 'Variables'

box, and the tutor variable across into the 'Grouping Variable' box, as shown in Figure 10.10.

The output has a very familiar form. First, it tells you what test was run, and it tells you the name of the dependent variable that you used. It then reports the test results. Just like last time the test results consist of a t-statistic, the degrees of freedom, and the p-value. The final section reports two things: it gives you a confidence interval and an effect size. I'll talk about effect sizes later. The confidence interval, however, I should talk about now.

It's pretty important to be clear on what this confidence interval actually refers to. It is a confidence interval for the difference between the group means. In our example, Anastasia's students had an average grade of 74.533, and Bernadette's students had an average grade of 69.056, so the difference between the two sample means is 5.478. But of course the difference between population means might be bigger or smaller than this. The observed 95% confidence interval for the population mean difference is [0.197, 10.759]. If we repeated the study many times and calculated an interval each time, approximately 95% of those intervals would contain the true population mean difference.

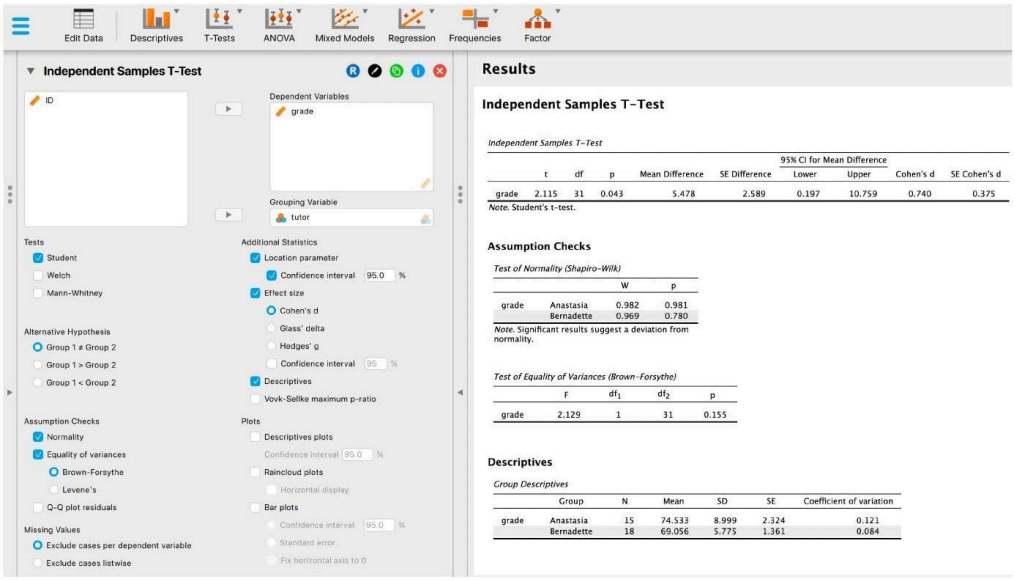


Figure 10.10: Independent t-test in JASP, with options checked for useful results

In any case, the difference between the two groups is significant (just barely), so we might write up the result using text like this:

> The mean grade in Anastasia's class was 74.5% (std dev = 9.0), whereas the mean in Bernadette's class was 69.1% (std dev = 5.8). A Student's independent samples t-test showed that this difference of approximately 5.5 percentage points was significant ($t(31) = 2.1, p < .05, Cl_{95} = [0.2, 10.8]$, $d = .74$), suggesting that a genuine difference in learning outcomes has occurred.

Notice that I've included the confidence interval and the effect size in the stat block. People don't always do this. At a bare minimum, you'd expect to see the t-statistic, the degrees of freedom and the $p$value. So you should include something like this at a minimum: $t(31) = 2.1, p < .05$. If statisticians had their way, everyone would also report the confidence interval and probably the effect size measure too, because they are useful things to know. But real life doesn't always work the way statisticians want it to so you should make a judgment based on whether you think it will help your readers and, if you're writing a scientific paper, the editorial standard for the journal in question. Some journals expect you to report effect sizes, others don't. Within some scientific communities it is standard practice to report confidence intervals, in others it is not. You'll need to figure out what your audience expects. But, just for the sake of clarity, if you're taking my class, my default position is that it's usually worth including both the effect size and the confidence interval.

10.3.6 Positive and negative $t$values

Before moving on to talk about the assumptions of the t-test, there's one additional point I want to make about the use of t-tests in practice. The first one relates to the sign of the t-statistic (that is, whether it is a positive number or a negative one). One very common worry that students have when they start running their first t-test is that they often end up with negative values for

the t-statistic and don't know how to interpret it. In fact, it's not at all uncommon for two people working independently to end up with results that are almost identical, except that one person has a negative $t$values and the other one has a positive $t$value. Assuming that you're running a two-sided test then the p-values will be identical. On closer inspection, the students will notice that the confidence intervals also have the opposite signs. This is perfectly okay. Whenever this happens, what you'll find is that the two versions of the results arise from slightly different ways of running the t-test. What's happening here is very simple. The t -statistic that we calculate here is always of the form

$$t = \frac{(\text{ mean 1}) - (\text{ mean 2})}{(\text{ SE })}$$

If "mean 1" is larger than "mean 2" the $t$statistic will be positive, whereas if "mean 2" is larger then the $t$statistic will be negative. Similarly, the confidence interval that JASP reports is the confidence interval for the difference "(mean 1) minus (mean 2)", which will be the reverse of what you'd get if you were calculating the confidence interval for the difference "(mean 2) minus (mean 1)".

Okay, that's pretty straightforward when you think about it, but now consider our t-test comparing Anastasia's class to Bernadette's class. Which one should we call "mean 1" and which one should we call "mean 2". It's arbitrary. However, you really do need to designate one of them as "mean 1" and the other one as "mean 2". Not surprisingly, the way that JASP handles this is also pretty arbitrary. In earlier versions of the book I used to try to explain it, but after a while I gave up, because it's not really all that important and to be honest I can never remember myself. Whenever I get a significant t-test result, and I want to figure out which mean is the larger one, I don't try to figure it out by looking at the t-statistic. Why would I bother doing that? It's foolish. It's easier just to look at the actual group means since the JASP output actually shows them!

Here's the important thing. Because it really doesn't matter what JASP shows you, I usually try to report the t-statistic in such a way that the numbers match up with the text. Suppose that what I want to write in my report is "Anastasia's class had higher grades than Bernadette's class". The phrasing here implies that Anastasia's group comes first, so it makes sense to report the t -statistic as if Anastasia's class corresponded to group 1. If so, I would write

Anastasia's class had higher grades than Bernadette's class $(t(31) = 2.1, p = .04)$.

(I wouldn't actually underline the word "higher" in real life, I'm just doing it to emphasise the point that "higher" corresponds to positive $t$values). On the other hand, suppose the phrasing I wanted to use has Bernadette's class listed first. If so, it makes more sense to treat her class as group 1, and if so, the write up looks like this

Bernadette's class had lower grades than Anastasia's class $(t(31) = -2.1, p = .04)$. Because I'm talking about one group having "lower" scores this time around, it is more sensible to use the negative form of the t-statistic. It just makes it read more cleanly.

One last thing: please note that you can't do this for other types of test statistics. It works for t-tests, but it wouldn't be meaningful for chi-square tests, F-tests or indeed for most of the tests I talk about in this book. So don't over-generalise this advice! I'm really just talking about t-tests here and nothing else!

10.3.7 Assumptions of the test

As always, our hypothesis test relies on some assumptions. So what are they? For the Student t-test there are three assumptions, some of which we saw previously in the context of the one sample t-test:

- -Normality. Like the one-sample t-test, it is assumed that the data are normally distributed. Specifically, we assume that both groups are normally distributed. In we'll discuss how to test for normality, and in we'll discuss possible solutions.
- -Independence. Once again, it is assumed that the observations are independently sampled. In the context of the Student test this has two aspects to it. Firstly, we assume that the observations within each sample are independent of one another (exactly the same as for the one-sample test). However, we also assume that there are no cross-sample dependencies. If, for instance, it turns out that you included some participants in both experimental conditions of your study (e.g., by accidentally allowing the same person to sign up to different conditions), then there are some cross sample dependencies that you'd need to take into account.
- -Homogeneity of variance (also called "homoscedasticity"). The third assumption is that the population standard deviation is the same in both groups. You can test this assumption using the Levene test. However, there's a very simple remedy for this assumption if you are worried, which I'll talk about in the next section.

10.4 The independent samples t-test (Welch test)

The biggest problem with using the Student test in practice is the third assumption listed in the previous section. It assumes that both groups have the same standard deviation. This is rarely true in real life. If two samples don't have the same means, why should we expect them to have the same standard deviation? There's really no reason to expect this assumption to be true. We'll talk a little bit about how you can check this assumption later on because it does crop up in a few different places, not just the t-test. But right now I'll talk about a different form of the t-test (Welch 1947) that does not rely on this assumption. A graphical illustration of what the Welch $t$test assumes about the data is shown in Figure 10.11, to provide a contrast with the Student test version in Figure 10.9. I'll admit it's a bit odd to talk about the cure before talking about the diagnosis, but as it happens the Welch test can be specified as one of the 'Independent Samples T-Test' options in JASP, so this is probably the best place to discuss it.

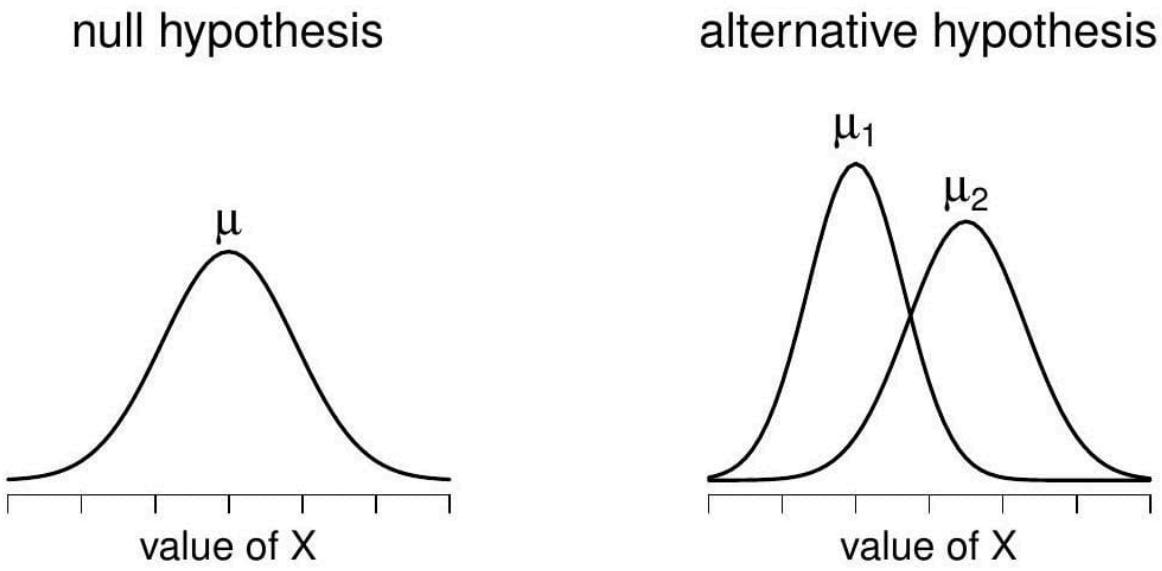


Figure 10.11: Graphical illustration of the null and alternative hypotheses assumed by the Welch t-test. Like the Student test (Figure 10.9) we assume that both samples are drawn from a normal population; but the alternative hypothesis no longer requires the two populations to have equal variance.

The Welch test is very similar to the Student test. For example, the t-statistic that we use in the Welch test is calculated in much the same way as it is for the Student test. That is, we take the difference between the sample means and then divide it by some estimate of the standard error of that difference

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\text{se}\left(\bar{X}_1 - \bar{X}_2\right)}$$

The main difference is that the standard error calculations are different. If the two populations have different standard deviations, then it's a complete nonsense to try to calculate a pooled standard deviation estimate, because you're averaging apples and oranges. [9]

But you can still estimate the standard error of the difference between sample means, it just ends up looking different

$$\text{se}\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$$

The reason why it's calculated this way is beyond the scope of this book. What matters for our purposes is that the t-statistic that comes out of the Welch t-test is actually somewhat different to the one that comes from the Student t-test.

The second difference between Welch and Student is that the degrees of freedom are calculated in a very different way. In the Welch test, the "degrees of freedom " doesn't have to be a whole number any more, and it doesn't correspond all that closely to the "number of data points minus the number of constraints" heuristic that I've been using up to this point.

The degrees of freedom are, in fact

$$\mathrm{df} = \frac{\left(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2\right)^2}{\left(\hat{\sigma}_1^2/N_1\right)^2/(N_1-1) + \left(\hat{\sigma}_2^2/N_2\right)^2/(N_2-1)}$$

which is all pretty straightforward and obvious, right? Well, perhaps not. It doesn't really matter for our purposes. What matters is that you'll see that the "df" value that pops out of a Welch test tends to be a little bit smaller than the one used for the Student test, and it doesn't have to be a whole number.

10.4.1 Doing the Welch test in JASP

If you tick the check box for the Welch test in the analysis we did above, then this is what it gives you (Figure 10.12):

Independent Samples T-Test

	Test	Statistic	df	p	Mean Difference	SE Difference	95% CI for Mean Difference		Cohen's d	SE Cohen's d
grade	Student	2.115	31.000	0.043	5.478	2.589	0.197	10.759	0.740	0.375
	Welch	2.034	23.025	0.054	5.478	2.693	-0.092	11.048	0.724	0.374

Figure 10.12: Results showing the Welch test alongside the default Student's t-test in JASP

The interpretation of this output should be fairly obvious. You read the output for the Welch's test in the same way that you would for the Student's test. You've got your descriptive statistics, the test results and some other information. So that's all pretty easy.

Except, except...our result isn't significant anymore. When we ran the Student test we did get a significant effect, but the Welch test on the same data set is not $(t(23.02) = 2.03, p = .054)$. What does this mean? Should we panic? Is the sky burning? Probably not. The fact that one test is significant and the other isn't doesn't itself mean very much, especially since I kind of rigged the data so that this would happen. As a general rule, it's not a good idea to go out of your way to try to interpret or explain the difference between a p-value of . 049 and a p-value of .051. If this sort of thing happens in real life, the difference in these p -values is almost certainly due to chance. What does matter is that you take a little bit of care in thinking about what test you use. The Student test and the Welch test have different strengths and weaknesses. If the two populations really do have equal variances, then the Student test is slightly more powerful (lower Type II error rate) than the Welch test. However, if they don't have the same variances, then the assumptions of the Student test are violated and you may not be able to trust it; you might end up with a higher Type I error rate. So it's a trade off. However, in real life I tend to prefer the Welch test, because almost no-one actually believes that the population variances are identical.

10.4.2 Assumptions of the test

The assumptions of the Welch test are very similar to those made by the Student t-test except that the Welch test does not assume homogeneity of variance. This leaves only the assumption of normality and the assumption of independence. The specifics of these assumptions are the same for the Welch test as for the Student test.

10.5 The paired-samples t-test

Regardless of whether we're talking about the Student test or the Welch test, an independent samples t-test is intended to be used in a situation where you have two samples that are, well, independent of one another. This situation arises naturally when participants are assigned randomly to one of two experimental conditions, but it provides a very poor approximation to other sorts of research designs. In particular, a repeated measures design, in which each participant is measured (with respect to the same outcome variable) in both experimental conditions, is not suited for analysis using independent samples t-tests. For example, we might be interested in whether listening to music reduces people's working memory capacity. To that end, we could measure each person's working memory capacity in two conditions: with music, and without music. In an experimental design such as this one, each participant appears in both groups. This requires us to approach the problem in a different way, by using the paired samples t-test.

10.5.1 The data

The data set that we'll use this time comes from Dr Chico's class. [10] In her class students take two major tests, one early in the semester and one later in the semester. To hear her tell it, she runs a very hard class, one that most students find very challenging. But she argues that by setting hard assessments students are encouraged to work harder. Her theory is that the first test is a bit of a "wake up call" for students. When they realise how hard her class really is, they'll work harder for the second test and get a better mark. Is she right? To test this, let's import the chico.csv file into JASP. The chico data set contains three variables: an id variable that identifies each student in the class, the grade_test1 variable that records the student grade for the first test, and the grade_test2 variable that has the grades for the second test.

If we look at the JASP spreadsheet it does seem like the class is a hard one (most grades are between 50% and 60%), but it does look like there's an improvement from the first test to the second one.

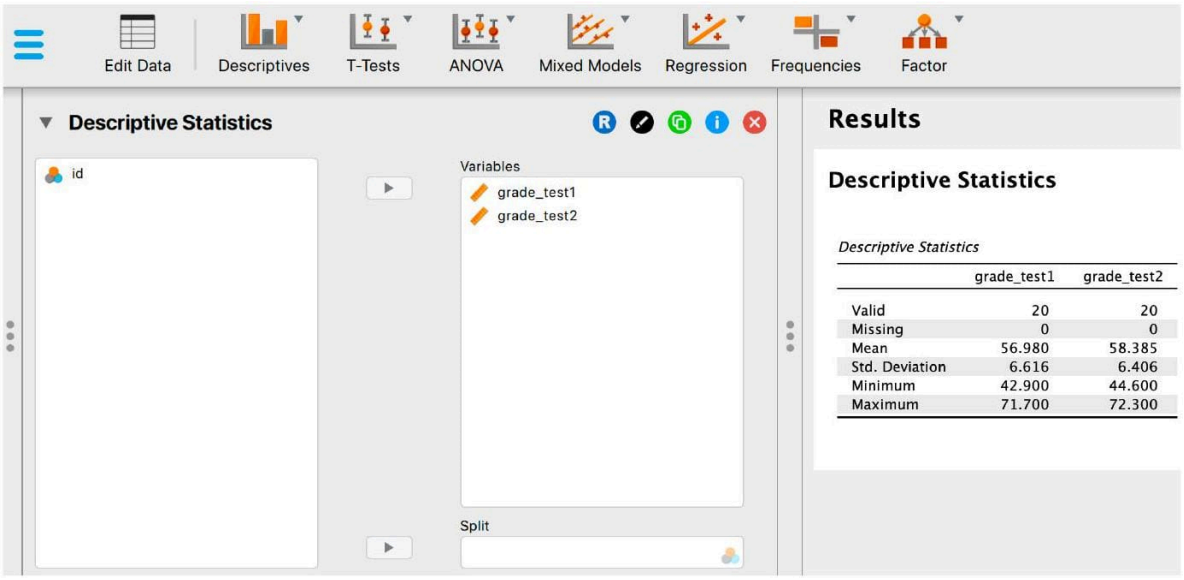


Figure 10.13: Descriptives for the two grade_test variables in the chico data set

If we take a quick look at the descriptive statistics, in Figure 10.13, we see that this impression seems to be supported. Across all 20 students the mean grade for the first test is 57%, but this rises to 58% for the second test. Although, given that the standard deviations are 6.6% and 6.4% respectively,

it's starting to feel like maybe the improvement is just illusory; maybe just random variation. This impression is reinforced when you see the means and confidence intervals plotted in Figure 10.14a. If we were to rely on this plot alone, looking at how wide those confidence intervals are, we'd be tempted to think that the apparent improvement in student performance is pure chance.

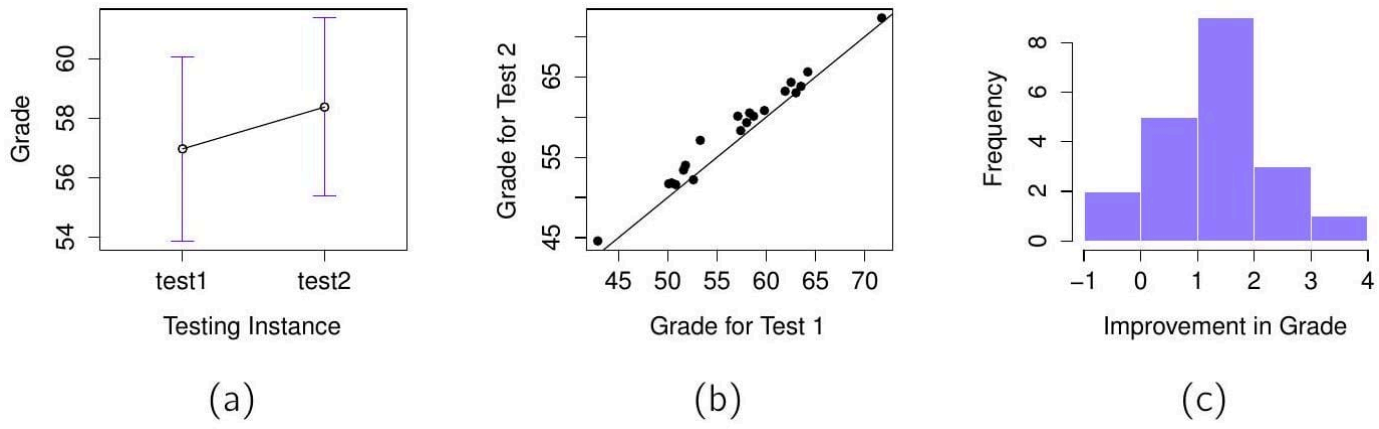


Figure 10.14: Mean grade for test 1 and test 2, with associated 95% confidence intervals (panel a). Scatterplot showing the individual grades for test 1 and test 2 (panel b). Histogram showing the improvement made by each student in Dr Chico's class (panel c). In panel c, notice that almost the entire distribution is above zero: the vast majority of students did improve their performance from the first test to the second one

Nevertheless, this impression is wrong. To see why, take a look at the scatterplot of the grades for test 1 against the grades for test 2, shown in Figure 10.14b. In this plot each dot corresponds to the two grades for a given student. If their grade for test 1 ($x$co-ordinate) equals their grade for test 2 ($y$co-ordinate), then the dot falls on the line. Points falling above the line are the students that performed better on the second test. Critically, almost all of the data points fall above the diagonal line: almost all of the students do seem to have improved their grade, if only by a small amount. This suggests that we should be looking at the improvement made by each student from one test to the next and treating that as our raw data. To do this, we'll need to create a new variable for the improvement that each student makes, and add it to the chico data set. The easiest way to do this is to compute a new variable. In JASP, click on the "+" at the right-most side of the data columns, name the variable improvement, and select the "R" button. After you click the 'Create column' button, you can enter the R code grade_test2 - grade_test1 (see Figure 10.15).

Once we have computed this new improvement variable we can draw a histogram showing the distribution of these improvement scores, shown in Figure 10.14c. When we look at the histogram, it's very clear that there is a real improvement here. The vast majority of the students scored higher on test 2 than on test 1 , reflected in the fact that almost the entire histogram is above zero.

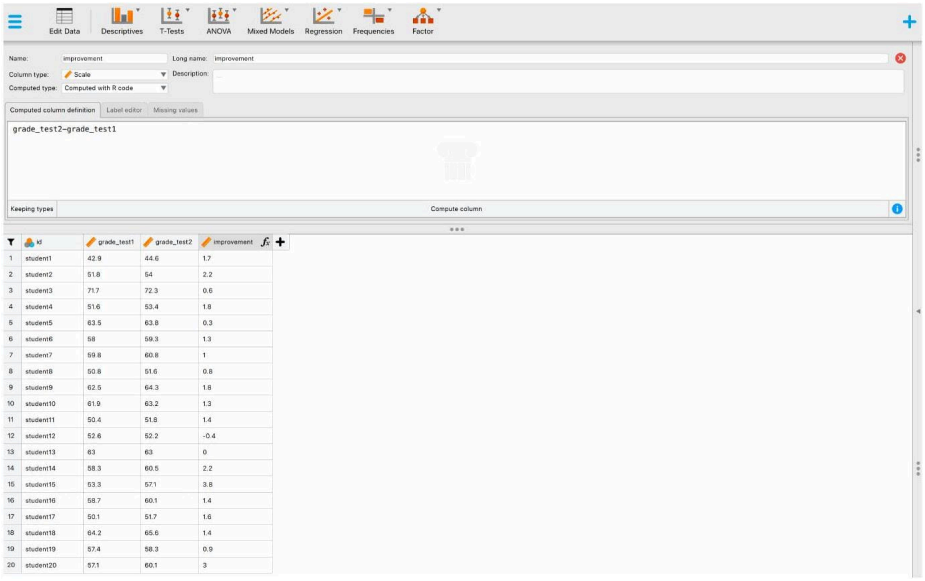


Figure 10.15: Using R code to compute an improvement score in JASP.

10.5.2 What is the paired samples t-test?

In light of the previous exploration, let's think about how to construct an appropriate $t$test. One possibility would be to try to run an independent samples t-test using grade_test1 and grade_test2 as the variables of interest. However, this is clearly the wrong thing to do as the independent samples t-test assumes that there is no particular relationship between the two samples. Yet clearly that's not true in this case because of the repeated measures structure in the data. To use the language that I introduced in the last section, if we were to try to do an independent samples t-test, we would be conflating the within subject differences (which is what we're interested in testing) with the between subject variability (which we are not).

The solution to the problem is obvious, I hope, since we already did all the hard work in the previous section. Instead of running an independent samples t-test on grade_test1 and grade_test2, we run a one-sample t-test on the within-subject difference variable, improvement. To formalise this slightly, if X_{i1} is the score that the i-th participant obtained on the first variable, and X_{i2} is the score that the same person obtained on the second one, then the difference score is:

$$D_i = X_{i1} - X_{i2}$$

Notice that the difference scores is variable 1 minus variable 2 and not the other way around, so if we want improvement to correspond to a positive valued difference, we actually want "test 2" to be our "variable 1". Equally, we would say that $\mu_D = \mu_1 - \mu_2$ is the population mean for this difference variable. So, to convert this to a hypothesis test, our null hypothesis is that this mean difference is zero and the alternative hypothesis is that it is not

$$
\begin{aligned}
H_0: &\quad \mu_D = 0 \\
H_1: &\quad \mu_D \neq 0
\end{aligned}
$$

This is assuming we're talking about a two-sided test here. This is more or less identical to the way we described the hypotheses for the one-sample t-test. The only difference is that the specific value that the null hypothesis predicts is 0. And so our t-statistic is defined in more or less the same way too. If we let $\bar{D}$ denote the mean of the difference scores, then

$$
t = \frac{\bar{D}}{\mathrm{se}(\bar{D})}
$$

which is

$$
t = \frac{\bar{D}}{\hat{\sigma}_D / \sqrt{N}}
$$

where $\hat{\sigma}_D$ is the standard deviation of the difference scores. Since this is just an ordinary, one-sample t-test, with nothing special about it, the degrees of freedom are still $N - 1$. And that's it. The paired samples t-test really isn't a new test at all. It's a one-sample t-test, but applied to the difference between two variables. It's actually very simple. The only reason it merits a discussion as long as the one we've just gone through is that you need to be able to recognise when a paired samples test is appropriate, and to understand why it's better than an independent samples $t$test.

10.5.3 Doing the test in JASP

How do you do a paired samples t-test in JASP? One possibility is to follow the process I outlined above. That is, create a "difference" variable and then run a one sample t-test on that. Since we've already created a variable called improvement, let's do that and see what we get, Figure 10.16.

The output shown in Figure 10.16 is (obviously) formatted exactly the same was as it was the last time we used the 'One Sample T-Test' analysis, and it confirms our intuition. There's an average improvement of 1.4 points from test 1 to test 2 , and this is significantly different from 0 ($t(19) = 6.48, p < .001$).

However, suppose you're lazy and you don't want to go to all the effort of creating a new variable. Or perhaps you just want to keep the difference between one-sample and paired-samples tests clear in your head. If so, you can use the JASP 'Paired Samples T-Test' analysis. As you will see, the numbers are identical to those that come from the one sample test, which of course they have to be given that the paired samples t-test is just a one sample test under the hood.

One Sample T-Test

One Sample T-Test

	t	df	p	Mean Difference	95% CI for Mean Difference		Cohen's d	SE Cohen's d
					Lower	Upper		
improvement	6.475	19	< . 001	1.405	0.951	1.859	1.448	0.320

Figure 10.16: Results showing a one sample t-test on paired difference scores

Note. For the Student t-test, the alternative hypothesis specifies that the mean is different from 0.
Note. Student's t-test.

10.6 One sided tests

When introducing the theory of null hypothesis tests, I mentioned that there are some situations when it's appropriate to specify a one-sided test. So far all of the t-tests have been two-sided tests. For instance, when we specified a one sample t-test for the grades in Dr Zeppo's class the null hypothesis was that the true mean was 67.5%. The alternative hypothesis was that the true mean was greater than or less than 67.5%. Suppose we were only interested in finding out if the true mean is greater than 67.5% , and have no interest whatsoever in testing to find out if the true mean is lower than 67.5%. If so, our null hypothesis would be that the true mean is 67.5% or less, and the alternative hypothesis would be that the true mean is greater than 67.5%. In JASP, for the 'One Sample T-Test' analysis, you can specify this by clicking on the '> Test Value' option, under 'Alt. Hypothesis'. When you have done this, you will get the results as shown in 10.17.

One Sample T-Test

One Sample T-Test

	t	df	p	Mean Difference	95% CI for Mean Difference		Cohen's d	SE Cohen's d
					Lower	Upper		
x	2.255	19	0.018	4.800	1.119	∞	0.504	0.237

Figure 10.17: JASP results showing a 'One Sample T-Test' where the actual hypothesis is one sided, i.e. that the true mean is greater than 67.5%

Note. For the Student t-test, the alternative hypothesis specifies that the mean is greater than 67.5.
Note. Student's t-test.

Notice that there are a few changes from the output that we saw last time. Most important is the fact that the actual hypothesis has changed, to reflect the different test. The second thing to note is that although the t-statistic and degrees of freedom have not changed, the p-value has. This is because the one-sided test has a different rejection region from the two-sided test. The third thing to note is that the confidence interval is different too: it now reports a "one-sided" confidence interval rather than a two-sided one. A two-sided 95% confidence interval provides both a lower and an upper bound. Across repeated studies, approximately 95% of intervals constructed this way would contain the true population mean. A one-sided 95% confidence interval provides one bound: across repeated studies, approximately 95% of lower bounds would fall below the true mean, or approximately 95% of upper bounds would fall above it.

So that's how to do a one-sided one sample t-test. However, all versions of the t-test can be one-sided. For an independent samples $t$test, you could have a one-sided test if you're only interested in testing to see if group A has higher scores than group B, but have no interest in finding out if group B has higher scores than group A. Let's suppose that, for Dr Harpo's class, you wanted to see if Anastasia's students had higher grades than Bernadette's. For this analysis, in the 'Alt. Hypothesis' options, specify that 'Group 1 > Group2'. You should get the results shown in Figure 10.18.

Independent Samples T-Test

						95% CI for Mean Difference		Cohen's d	SE Cohen's d
	t	df	p	Mean Difference	SE Difference	Lower	Upper		
grade	2.115	31	0.021	5.478	2.589	1.087	∞	0.740	0.375

Note. For all tests, the alternative hypothesis specifies that group Anastasia is greater than group Bernadette .

Note. Student's t-test.

Figure 10.18: JASP results showing an 'Independent Samples T-Test' where the actual hypothesis is one sided, i.e. that Anastasia's students had higher grades than Bernadette's

Again, the output changes in a predictable way. The definition of the alternative hypothesis has changed, the p-value has changed, and it now reports a one-sided confidence interval rather than a two-sided one.

What about the paired samples t-test? Suppose we wanted to test the hypothesis that grades go up from test 1 to test 2 in Dr Chico's class, and are not prepared to consider the idea that the grades go down. In JASP you would do this by specifying, under the 'Alt. Hypotheses' option, that grade_test2 ('Measure 1' in JASP, because we copied this first into the paired variables box) > grade_test1 ('Measure 2' in JASP). You should get the results shown in Figure 10.19.

Yet again, the output changes in a predictable way. The hypothesis has changed, the p-value has changed, and the confidence interval is now one-sided.

Paired Samples T-Test

Paired Samples T-Test

Measure 1		Measure 2	t	df	p	Mean Difference	SE Difference	95% CI for Mean Difference		Cohen's d	SE Cohen's d
grade_test2	-	grade_test1	6.475	19	< .001	1.405	0.217	1.030	∞	1.448	0.047

Note. For all tests, the alternative hypothesis specifies that grade_test2 is greater than grade_test1.
Note. Student's t-test.

Figure 10.19: JASP results showing a 'Paired Samples T-Test' where the actual hypothesis is one sided, i.e. that grade_test2 ('Measure 1') > grade_test1 ('Measure 2')

10.7 Effect size

The most commonly used measure of effect size for a t-test is Cohen's d(Cohen 1988). It's a very simple measure in principle, with quite a few wrinkles when you start digging into the details. Cohen himself defined it primarily in the context of an independent samples t-test, specifically the Student test. In that context, a natural way of defining the effect size is to divide the difference between the means by an estimate of the standard deviation. In other words, we're looking to calculate something along the lines of this:

$$d = \frac{(\text{ mean } 1) - (\text{ mean } 2)}{\text{std dev}}$$

and he suggested a rough guide for interpreting $d$in Table 10.1. You'd think that this would be pretty unambiguous, but it's not. This is largely because Cohen wasn't too specific on what he thought should be used as the measure of the standard deviation (in his defence he was trying to make a broader point in his book, not nitpick about tiny details). As discussed by McGrath and Meyer (2006), there are several different versions in common usage, and each author tends to adopt slightly different notation. For the sake of simplicity (as opposed to accuracy), I'll use $d$to refer to any statistic that you calculate from the sample, and use δ to refer to a theoretical population effect. Obviously, that does mean that there are several different things all called d.

My suspicion is that the only time that you would want Cohen's $d$is when you're running a t-test, and JASP has an option to calculate the effect size for all the different flavours of t-test it provides.

10.7.1 Cohen's d from one sample

The simplest situation to consider is the one corresponding to a one-sample t-test. In this case, there is the one sample mean $\bar{X}$ and one (hypothesised) population mean μ_o to compare it to. Not only that, there's really only one sensible way to estimate the population standard deviation. We just use our usual estimate $\hat{\sigma}$. Therefore, we end up with the following as the only way to calculate \(d$

$$d = \frac{\bar{X} - \mu_0}{\hat{\sigma}}$$

Table 10.1: A (very) rough guide to interpreting Cohen's d. My personal recommendation is to not use these blindly. The $d$statistic has a natural interpretation in and of itself. It re-describes the difference in means as the number of standard deviations that separates those means. So it's generally a good idea to think about what that means in practical terms. In some contexts a "small"

effect could be of big practical importance. In other situations a "large" effect may not be all that interesting.

d-value	rough interpretation
about 0.2	"small" effect
about 0.5	"moderate" effect
about 0.8	"large" effect

When we look back at the results in Figure 10.6, the effect size value is Cohen's $\mathrm{d} = 0.504$. Overall, then, the psychology students in Dr Zeppo's class are achieving grades $(\text{mean } = 72.3\%)$that are about 0.5 standard deviations higher than the level that you'd expect (67.5%) if they were performing at the same level as other students. Judged against Cohen's rough guide, this is a moderate effect size.

10.7.2 Cohen's $d$from a Student's $t$test

The majority of discussions of Cohen's $d$focus on a situation that is analogous to Student's independent samples $t$test, and it's in this context that the story becomes messier, since there are several different versions of $d$that you might want to use in this situation. To understand why there are multiple versions of d, it helps to take the time to write down a formula that corresponds to the true population effect size δ. It's pretty straightforward,

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

where, as usual, μ_1 and μ_2 are the population means corresponding to group 1 and group 2 respectively, and $\sigma$is the standard deviation (the same for both populations). The obvious way to estimate $\delta$is to do exactly the same thing that we did in the t-test itself, i.e., use the sample means as the top line and a pooled standard deviation estimate for the bottom line

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_p}$$

where $\hat{\sigma}_p$ is the exact same pooled standard deviation measure that appears in the t-test. This is the most commonly used version of Cohen's $d$when applied to the outcome of a Student t-test, and is the one provided in JASP. It is sometimes referred to as Hedges' $g$statistic (Hedges 1981).

However, there are other possibilities which I'll briefly describe. Firstly, you may have reason to want to use only one of the two groups as the basis for calculating the standard deviation. This approach (often called Glass' Δ, pronounced delta) only makes most sense when you have good reason to treat one of the two groups as a purer reflection of "natural variation" than the other. This can happen if, for instance, one of the two groups is a control group. Secondly, recall that in the usual calculation of the pooled standard deviation we divide by $N - 2$to correct for the bias in the sample variance. In one version of Cohen's $d$this correction is omitted, and instead we divide by N. This version makes sense primarily when you're trying to calculate the effect size in the sample rather than estimating an effect size in the population. Finally, there is a version based on Hedges and Olkin (1985), who point out there is a small bias in the usual (pooled) estimation for Cohen's d. Thus they introduce a small correction by multiplying the usual value of $d$by $(N - 3)/(N - 2.25)$.

In any case, ignoring all those variations that you could make use of if you wanted, let's have a look at the default version in JASP. In Figure 10.10 Cohen's $\mathrm{d} = 0.740$, indicating that the grade scores for students in Anastasia's class are, on average, 0.74 standard deviations higher than the grade scores for students in Bernadette's class. For the Welch test shown in Figure 10.12, JASP reports a slightly different effect size, d = 0.724.

10.7.3 Cohen's d from a paired-samples test

Finally, what should we do for a paired samples t-test? In this case, the answer depends on what it is you're trying to do. JASP assumes that you want to measure your effect sizes relative to the distribution of difference scores, and the measure of $d$that you calculate is:

$$d = \frac{\bar{D}}{\hat{\sigma}_D}$$

where $\hat{\sigma}_D$ is the estimate of the standard deviation of the differences. In Figure 10.16, Cohen's d is approximately 1.45: the mean improvement is 1.45 times the standard deviation of the difference scores.

This is the version of Cohen's $d$that gets reported by the JASP 'Paired Samples T-Test' analysis. The only wrinkle is figuring out whether this is the measure you want or not. To the extent that you care about the practical consequences of your research, you often want to measure the effect size relative to the original variables, not the difference scores (e.g., the 1 point improvement in Dr Chico's class over time is pretty small when measured against the amount of between-student variation in grades), in which case you use the same versions of Cohen's $d$that you would use for a Student or Welch test. It's not so straightforward to do this in JASP; essentially you have to change the structure of the data in the spreadsheet view so I won't go into that here.

10.8 Checking the normality of a sample

The normality assumption concerns the population scores for a one-sample test, the population scores within each group for an independent-samples test, and the population difference scores for a paired-samples test. This assumption is often quite reasonable, because the central limit theorem does tend to ensure that many real world quantities are normally distributed. Any time that you suspect that your variable is actually an average of lots of different things, there's a pretty good chance that it will be normally distributed, or at least close enough to normal that you can get away with using t-tests. However, life doesn't come with guarantees, and besides there are lots of ways in which you can end up with variables that are highly non-normal. For example, any time you think that your variable is actually the minimum of lots of different things, there's a very good chance it will end up quite skewed. In psychology, response time (RT) data is a good example of this. If you suppose that there are lots of things that could trigger a response from a human participant, then the actual response will occur the first time one of these trigger events occurs. [11] This means that RT data are systematically non-normal. Okay, so if normality is assumed by all the tests, and is mostly but not always satisfied (at least approximately) by real world data, how can we check the normality of a sample? In this section I discuss two methods: QQ plots and the Shapiro-Wilk test.

10.8.1 QQ plots

One way to check whether a sample violates the normality assumption is to draw a "QQ plot" (Quantile-Quantile plot). This allows you to visually check whether you're seeing any systematic violations. In a QQ plot, each observation is plotted as a single dot. The x co-ordinate is the theoretical quantile that the observation should fall in if the data were normally distributed (with mean and variance estimated from the sample), and on the y co-ordinate is the actual quantile of the data within the sample. If the data are normal, the dots should form a straight line. For instance, lets see what happens if we generate data by sampling from a normal distribution, and then drawing a QQ plot. The results are shown in Figure 10.20. As you can see, these data form a pretty straight line; which is no surprise given that we sampled them from a normal distribution! In contrast, have a look at the two data sets shown in Figure 10.21. The top panels show the histogram and a QQ plot for a data set that is highly skewed: the QQ plot curves upwards. The lower panels show the same plots for a heavy tailed (i.e., high kurtosis) data set: in this case the QQ plot flattens in the middle and curves sharply at either end.

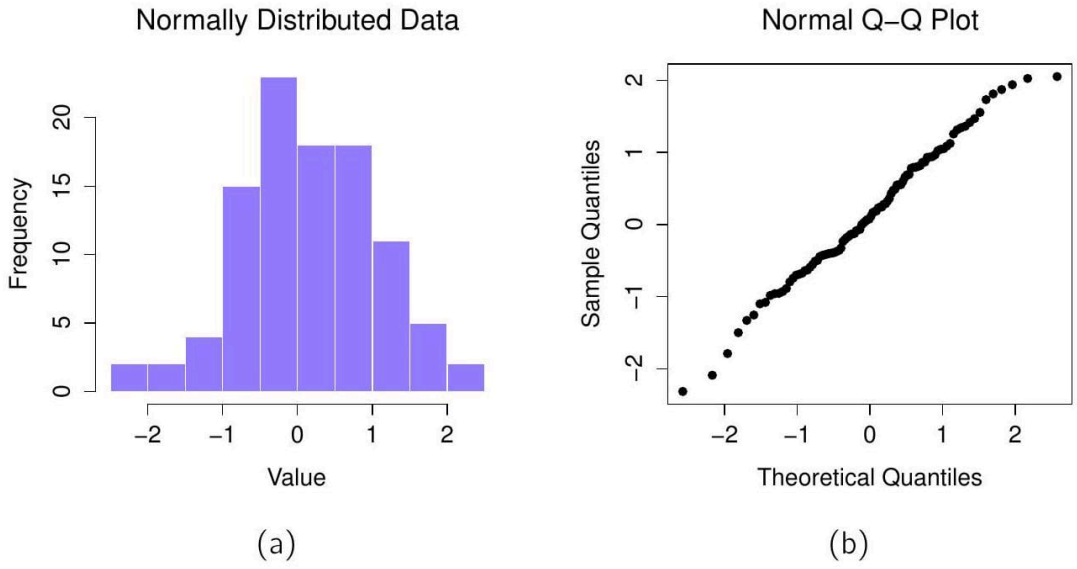


Figure 10.20: Histogram (panel a) and normal QQ plot (panel b) of normal.data, a normally distributed sample with 100 observations. The Shapiro-Wilk statistic associated with these data is $W = .99$, indicating that no significant departures from normality were detected $(p = .73)$.

10.8.2 Shapiro-Wilk tests

QQ plots provide a nice way to informally check the normality of your data, but sometimes you'll want to do something a bit more formal and the Shapiro-Wilk test (Shapiro and Wilk 1965) is probably what you're looking for. [12] As you'd expect, the null hypothesis being tested is that a set of Nobservations is normally distributed.

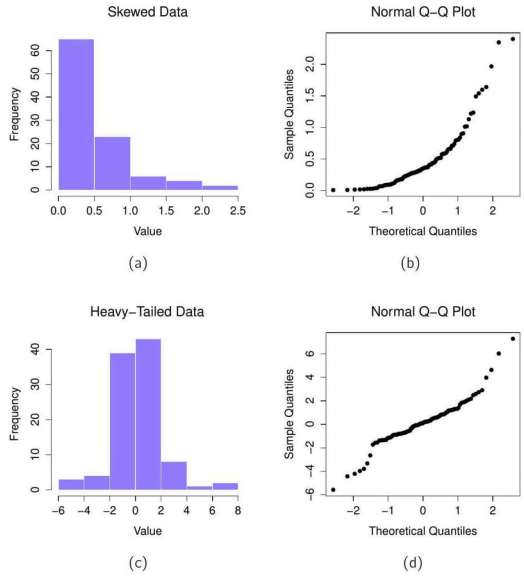


Figure 10.21: In the top row, a histogram (panel a) and normal QQ plot (panel b) of the 100 observations in a skewed.data set. The skewness of the data here is 1.94, and is reflected in a QQ plot that curves upwards. As a consequence, the Shapiro-Wilk statistic is $W = .80$, reflecting a significant departure from normality ($p < .001$). The bottom row shows the same plots for a heavy tailed data set, again consisting of 100 observations. In this case the heavy tails in the data produce a high kurtosis (2.80), and cause the QQ plot to flatten in the middle, and curve away sharply on either side. The resulting Shapiro-Wilk statistic is $W = .93$, again reflecting significant non-normality ($p < .001$).

The test statistic that it calculates is conventionally denoted as W, and it's calculated as follows. First, we sort the observations in order of increasing size, and let X_1 be the smallest value in the sample, X_2 be the second smallest and so on. Then the value of $W$is given by

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

where $\bar{X}$ is the mean of the observations, and the a_i values are ... mumble, mumble ... something complicated that is a bit beyond the scope of an introductory text.

Because it's a little hard to explain the maths behind the Wstatistic, a better idea is to give a broad brush description of how it behaves. Unlike most of the test statistics that we'll encounter in this book, it's actually small values of Wthat indicate departure from normality. The Wstatistic has a maximum value of 1 , which occurs when the data look "perfectly normal". The smaller the value of $W$the less normal the data are. However, the sampling distribution for W, which is not one of the standard ones and is in fact a complete pain in the arse to work with, does depend on the sample size N. To give you a feel for what these sampling distributions look like, I've plotted three of them in Figure 10.22. Notice that, as the sample size starts to get large, the sampling distribution becomes very tightly clumped up near $W = 1$, and as a consequence, for larger samples Wdoesn't have to be very much smaller than 1 in order for the test to be significant.

To get the Shapiro-Wilk statistic in JASP t-tests, check the option for 'Normality' listed under 'Assumption checks'. In the randomly

sampled data $(\mathrm{N} = 100)$we used for the QQ plot, the value for the Shapiro-Wilk normality test statistic was $\mathrm{W} = 0.99$with a p -value of 0.69 . So, not surprisingly, we have no evidence that these data depart from normality. When reporting the results for a Shapiro-Wilk test, you should (as usual) make sure to include the test statistic $W$and the $p$value, though given that the sampling distribution depends so heavily on Nit would probably be a politeness to include Nas well.

10.8.3 Example

In the meantime, it's probably worth showing you an example of what happens to the QQ plot and the Shapiro-Wilk test when the data turn out to be non-normal. For that, let's look at the distribution of our AFL winning margins data, which didn't look like they came from a normal distribution at all. Here's what happens to the QQ plot:

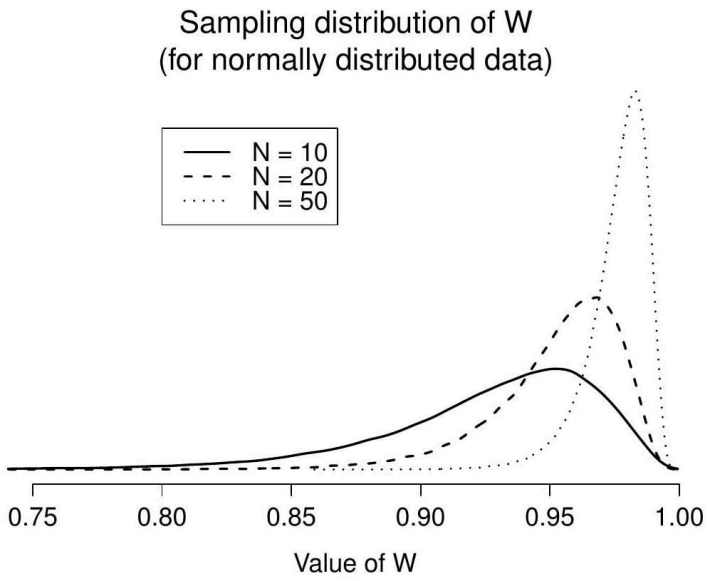
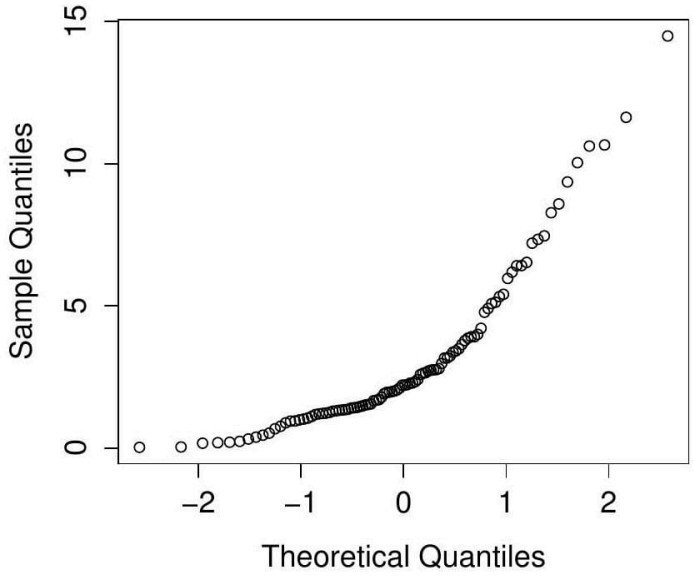


Figure 10.22: Sampling distribution of the Shapiro-Wilk Wstatistic, under the null hypothesis that the data are normally distributed, for samples of size 10, 20 and 50. Note that small values of Windicate departure from normality.

Normal Q-Q Plot



And when we run the Shapiro-Wilk test on the AFL margins data, we get a value for the Shapiro-Wilk normality test statistic of $\mathrm{W} = 0.94$, and p-value $= 9.481\mathrm{e}-07$. This is clearly a significant departure from normality!

10.9 Testing non-normal data with Wilcoxon tests

Okay, suppose your data turn out to be pretty substantially non-normal, but you still want to run something like a t-test? This situation occurs a lot in real life. For the AFL winning margins data, for instance, the Shapiro-Wilk test made it very clear that the normality assumption is violated. This is the situation where you want to use Wilcoxon tests.

Like the t-test, the Wilcoxon test comes in two forms, one-sample and two-sample, and they're used in more or less the exact same situations as the corresponding t-tests. Unlike the t-test, the Wilcoxon test doesn't assume normality, which is nice. These are nonparametric tests, but they still have assumptions. Observations must be independent across participants. For a location interpretation, the signed-rank test assumes a symmetric distribution of differences, while the Mann–Whitney test requires similarly shaped group distributions. Their power relative to a t-test depends on the population distributions.I won't discuss the Wilcoxon tests in as much detail as the t-tests, but I'll give you a brief overview.

10.9.1 Two sample Mann-Whitney U test

I'll start by describing the Mann-Whitney U test, since it's actually simpler than the one sample version. Suppose we're looking at the scores of 10 people on some test. Since my imagination has now failed me completely, let's pretend it's a "test of awesomeness" and there are two groups of people, "A" and "B". I'm curious to know which group is more awesome. The data are included in the file awesome.csv, and there are two variables apart from the usual ID variable: scores and group.

As long as there are no ties (i.e., people with the exact same awesomeness score) then the test that we want to do is surprisingly simple. All we have to do is construct a table that compares every observation in group Aagainst every observation in group B. Whenever the group Adatum is larger, we place a check mark in the table:

Group A score ↓ / Group B score →	14.5	10.4	12.4	11.7	13.0
6.4	—	—	—	—	—
10.7	—	✓	—	—	—
11.9	—	✓	—	✓	—
7.3	—	—	—	—	—
10.0	—	—	—	—	—

We then count up the number of checkmarks. This is our test statistic, W. [13] The actual sampling distribution for $W$is somewhat complicated, and I'll skip the details. For our purposes, it's sufficient to note that the interpretation of $W$is qualitatively the same as the interpretation of $t$or z. That is, if we want a two-sided test then we reject the null hypothesis when $W$is very large or very small, but if we have a directional (i.e., one-sided) hypothesis then we only use one or the other.

In JASP, if we run an 'Independent Samples T-Test' with scores as the dependent variable. and group as the grouping variable, and then under the options for 'tests' check the option for 'Mann-Whitney', we will get results showing that $\mathrm{U} = 3$(i.e., the same number of checkmarks as shown above), and a $\mathrm{p} - \mathrm{value} = 0.05556$.

10.9.2 One sample Wilcoxon test

What about the one sample Wilcoxon test (or equivalently, the paired samples Wilcoxon test)? Suppose I'm interested in finding out whether taking a statistics class has any effect on the happiness of students. My data is in the happiness.csv file. What I've measured here is the happiness of each student before taking the class and after taking the class, and the change score is the difference between the two. Just like we saw with the t-test, there's no fundamental difference between doing a paired-samples test using before and after, versus doing a one-sample test using the change scores. As before, the simplest way to think about the test is to construct a tabulation. The way to do it

this time is to take those change scores that are positive differences, and tabulate them against all the complete sample. What you end up with is a table that looks like this:

Positive difference ↓ / Sample difference →	−24	−14	−10	7	−6	−38	2	−35	−30	5
7	–	–	–	✓	✓	–	✓	–	–	✓
2	–	–	–	–	–	–	✓	–	–	–
5	–	–	–	–	–	–	✓	–	–	✓

Counting up the tick marks this time we get a test statistic of $W = 7$. As before, if our test is two sided, then we reject the null hypothesis when $W$is very large or very small. As far as running it in JASP goes, it's pretty much what you'd expect. For the one-sample version, you specify the 'Wilcoxon signed-rank' option under 'Tests' in the 'One Sample T-Test' analysis window. This gives you Wilcoxon $\mathrm{W} = 7, p$-value $= 0.037$. As this shows, we have a significant effect. Evidently, taking a statistics class does have an effect on your happiness. Switching to a paired samples version of the test won't give us a different answer, of course; see Figure 10.23.

10.10 Summary

- -A one sample t-test is used to compare a single sample mean against a hypothesised value for the population mean.
- -An independent samples t-test is used to compare the means of two groups, and tests the null hypothesis that they have the same mean. It comes in two forms: the Student test assumes that the groups have the same standard deviation, the Welch test does not.
- -A paired samples t-test is used when you have two scores from each person, and you want to test the null hypothesis that the two scores have the same mean. It is equivalent to taking the difference between the two scores for each person, and then running a one sample t-test on the difference scores.
- -One sided tests are perfectly legitimate as long as they are pre-planned.
- -Effect size calculations for the difference between means can be calculated via the Cohen's $d$statistic
- -If your data are non-normal, you can use Mann-Whitney or Wilcoxon tests instead of t-tests.

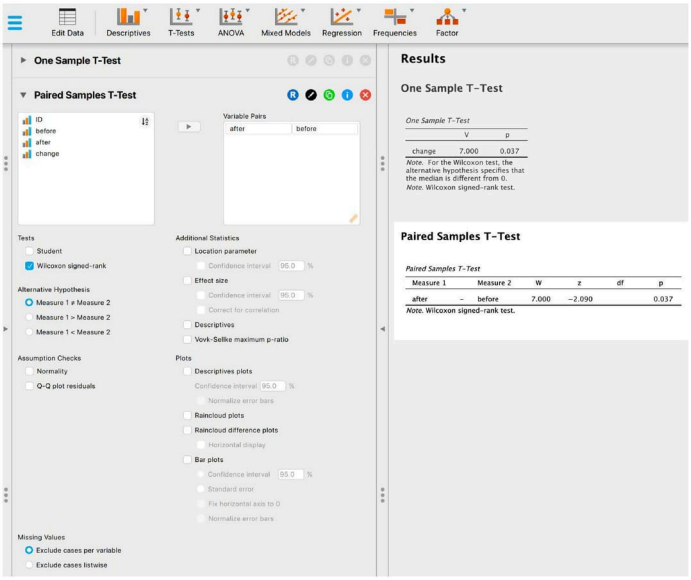


Figure 10.23: JASP screen showing results for one sample and paired sample Wilcoxon non-parametric tests

---

1. [1] Informal experimentation in my garden suggests that yes, it does. Australian natives are adapted to low phosphorus levels relative to everywhere else on Earth, so if you've bought a house with a bunch of exotics and you want to plant natives, keep them separate; nutrients to European plants are poison to Australian ones.
2. [2] The z-test was added as an option in the "One Sample T-Test" module with release 0.8.6.
3. [3] Actually this is too strong. Strictly speaking the $z$test only requires that the sampling distribution of the mean be normally distributed. If the population is normal then it necessarily follows that the sampling distribution of the mean is also normal. However, as we saw when talking about the central limit theorem, it's quite possible (even commonplace) for the sampling

distribution to be normal even if the population distribution itself is non-normal. However, in light of the sheer ridiculousness of the assumption that the true standard deviation is known, there really isn't much point in going into details on this front!

4. [4] Well, sort of. As I understand the history, Gosset only provided a partial solution; the general solution to the problem was provided by Sir Ronald Fisher.

5. [5] More seriously, I tend to think the reverse is true. I get very suspicious of technical reports that fill their results sections with nothing except the numbers. It might just be that I'm an arrogant jerk, but I often feel like an author that makes no attempt to explain and interpret their analysis to the reader either doesn't understand it themselves, or is being a bit lazy. Your readers are smart, but not infinitely patient. Don't annoy them if you can help it.
   [6] A technical comment. In the same way that we can weaken the assumptions of the z-test so that we're only talking about the sampling distribution, we can weaken the t-test assumptions so that we don't have to assume normality of the population. However, for the t-test it's trickier to do this. As before, we can replace the assumption of population normality with an assumption that the sampling distribution of $\bar{X}$ is normal. However, remember that we're also relying on a sample estimate of the standard deviation, and so we also require the sampling distribution of $\hat{\sigma}$ to be chi-square. That makes things nastier, and this version is rarely used in practice. Fortunately, if the population distribution is normal, then both of these two assumptions are met.
   [7] Although it is the simplest, which is why I started with it.

6. [8] A funny question almost always pops up at this point: what the heck is the population being referred to in this case? Is it the set of students actually taking Dr Harpo's class (all 33 of them)? The set of people who might take the class (an unknown number of them)? Or something else? Does it matter which of these we pick? It's traditional in an introductory behavioural stats class to mumble a lot at this point, but since I get asked this question every year by my students, I'll give a brief answer. Technically yes, it does matter. If you change your definition of what the "real world" population actually is, then the sampling distribution of your observed mean $\bar{X}$ changes too. The t-test relies on an assumption that the observations are sampled at random from an infinitely large population and, to the extent that real life isn't like that, then the t-test can be wrong. In practice, however, this isn't usually a big deal. Even though the assumption is almost always wrong, it doesn't lead to a lot of pathological behaviour from the test, so we tend to just ignore it.
7. [a] A more correct notation will be introduced in a later chapter.
8. [9] Well, I guess you can average apples and oranges, and what you end up with is a delicious fruit smoothie. But no one really thinks that a fruit smoothie is a very good way to describe the original fruits, do they?
9. [10] At this point we have Drs Harpo, Chico and Zeppo. No prizes for guessing who Dr Groucho is.
10. [11] This is a massive oversimplification.
11. [12] Either that, or the Kolmogorov-Smirnov test, which is probably more traditional than the Shapiro-Wilk. Although most things I've read seem to suggest Shapiro-Wilk is the better test of normality, the Kolomogorov- Smirnov is a general purpose test of distributional equivalence that can be adapted to handle other kinds of distribution tests. In JASP the Shapiro-Wilk test is preferred.
12. [13] Actually, there are two different versions of the test statistic that differ from each other by a constant value. The version that I've described is the one that JASP calculates.

This page titled 10.1: Comparing two means is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 10: Comparing two means by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

10.2: Independent t-test and Mann-Whitney U

Independent *t*-test

The independent *t*-test is used for analyses with a continuous dependent variable and a categorical independent variable but only when a **between-participants (independent measures) design** is used.

Note: The *t*-test can only be used to compare the differences in the means of **two groups**. For analyses with 3 groups or more, you can use another parametric test such as a one-way ANOVA.

How do I do that?

1. In the 'T-Test' menu select **Independent Samples T-Test** from the 'Classical' list.

image

2. A new window will then open, like so:

image

3. Move your **continuous dependent variable** (indicated by a ruler symbol) into the **Dependent Variables** box and your **categorical independent variable** (indicated by a Venn diagram symbol) into the **Grouping Variable** box.

image

4. There are many options in the expandable menus for this type of analysis which can initially be overwhelming. However, it is unlikely that you will need all of them. Here are some of the *t*-test options which you are likely to use:

Tests. The 'Student' test is the default, but you can also select the 'Welch' test which is reported if the assumption of equality of variance has been violated. If a rank-based comparison is appropriate, select Mann–Whitney (see next section).

image

Alternative Hypothesis. The default is the hypothesis that Group 1 ≠ Group 2. This is the non-directional hypothesis and for most cases, you will not change this. You can see that there are two options for the directional hypotheses.

Assumption Checks. These help you decide whether any of the assumptions for conducting an independent t-test have been violated; normality of distribution and equality of variances. For equality of variances select 'Levene's'.

The output for the **Assumption Checks** looks like this:

image

- The **Test of Normality** statistic is Shapiro Wilk (*W*). In this output, you can see that neither of the groups significantly differs from a 'normal' distribution.
- The **Equality of Variances** statistic is Levene's (*F*). In this output,t you can see that there is no significant difference in the group variances.

Additional Statistics. These provide you with additional information about your results. Unless instructed otherwise, it's usually a good idea to generate the following additional statistics:**image**

- Select **Location parameter** and **Confidence Interval.**
- Selecting **Effect Size** will select **Cohen's *d*** as the default. Glass' *delta* or Hedges' *g* are the other options that you can ignore for now.
- Selecting **Descriptives** and **Descriptive plots** will give you some basic descriptives and an error bar plot to help visualise the data.
- Selecting **Raincloud plots** gives you a rather fantastic set of plots that helps visualises the data in scatterplots, boxplots and overlayed histograms.

image

Mann-Whitney *U* Test

The **Mann-Whitney *U*** is a non-parametric alternative to the independent *t*-test for comparing two independent groups. It compares ranks rather than means. A significant Shapiro–Wilk result alone does not automatically require switching to this test; consider the sample sizes, outliers, and distribution shapes.

How do I do that?

1. Follow Steps 1-3 from the independent *t*-test section and then select the **Mann-Whitney U** option in the 'Tests' menu.

image

2. As with the independent t-test, the default for **Alternative Hypotheses** is Group 1 ≠ Group 2. This is the non-directional hypothesis and for most cases, you will not need to change this.

3. Under the **Analysis** options, you will only need to select **Effect Size** For the Mann–Whitney test, JASP reports the rank-biserial correlation rather than Cohen's d. Its sign indicates direction, and its magnitude indicates the strength of the rank-based difference between groups.

4. You may also want to select 'Raincloud plots' as this helps to visualise the data.

This page titled 10.2: Independent t-test and Mann-Whitney U is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss & Stanislava Naneva via source content that was edited to the style and standards of the LibreTexts platform.

10.3: Paired t-test and Wilcoxon Signed Ranks

Paired *t*-test

The paired *t*-test is used for analyses with a continuous dependent variable and a categorical independent variable but only when a **within-participants (repeated measures) design** is used.

Note: The paired *t*-test can only be used to compare the differences in the means of two variables for the same group of participants. More advanced tests are needed for designs with more than one group and/or more than two repeated measurements.

How do I do that?

1. In the 'T-Test' menu select **Paired Samples T-Test** from the 'Classical' list.

2. A new window will open, like so:

image

3. Move the two variables you wish to compare over to the **Variable Pairs** box.

image

4. There are many options in the expandable menus for this type of analysis which can initially be overwhelming. However, it is unlikely that you will need all of them. Here are some of the t-test options which you are likely to use:

Tests. The 'Student' test is the default but if you have non-parametric data, you need to select the 'Wilcoxon Signed Ranks' test.

Alternative Hypothesis. The default is the hypothesis that Group 1 ≠ Group 2. This is the non-directional hypothesis and for most cases, you will not change this. You can see that there are two options for the directional hypotheses.

Assumption Checks. These help you decide whether any of the assumptions for conducting a paired *t*-test have been violated. As this is a paired *t*-test which looks at the difference between the same group of people, the test for equality of variances is not needed so you only need to check 'Normality'. The Test of Normality statistic is **Shapiro Wilk (*W*)**. This check concerns the distribution of the difference scores. A significant result indicates a departure from normality, but does not automatically require a different test. Consider the sample size, outliers, and the shape of the difference-score distribution.

image **Additional Statistics**. These provide you with additional information about your results. Unless instructed otherwise, it is usually a good idea to generate the following additional statistics:

- Select **Location parameter** and **Confidence Interval.**
- Selecting **Effect Size** will select **Cohen's *d*** (although this is not stated).
- Selecting **Descriptives** and **Descriptive plots** will give you some basic descriptives and an error bar plot to help visualise the data.
- Selecting **Raincloud plots** will give you a rather fantastic set of plots that allows you to visualise the data in a scatterplots, boxplots, and overlayed histograms.

Wilcoxon Signed Ranks test

If you find that your data is not normally distributed (e.g., you have a significant Shapiro-Wilk test result), you need to conduct the

equivalent non-parametric paired t-test called **Wilcoxon Signed Ranks test**. The Wilcoxon signed-rank test is a nonparametric alternative for paired observations. For a location interpretation, it assumes that the distribution of difference scores is symmetric.

How do I do that?

1. Follow Steps 1-3 from the paired *t*-test and select the **Wilcoxon Signed Rank** option in the 'Tests' menu.

2. As with the paired *t*-test, the default for **Alternative Hypotheses** is Group 1 ≠ Group 2. This is the non-directional hypothesis and for most cases, you will not need to change this.

3. In the **Analysis** options, you will only need to select **Effect Size** and leave the default as Cohen's *d*. In the output table the effect size will be given by the **Rank-Biserial Correlation.** For the Wilcoxon signed-rank test, JASP reports the matched-pairs rank-biserial correlation rather than Cohen's d. Its sign indicates direction, and its magnitude indicates the strength of the rank-based change.

4. You may also want to select 'Raincloud plots' as this helps to visualise the data.

This page titled 10.3: Paired t-test and Wilcoxon Signed Ranks is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss & Stanislava Naneva via source content that was edited to the style and standards of the LibreTexts platform.

10.4: Worked Example- Calculating an Independent Samples t-Test

Example: Movies and Mood

We are interested in whether the type of movie someone sees at the theater affects their mood when they leave. We decide to ask people about their mood as they leave one of two movies: a comedy (Group 1, $n = 35$) or a horror film (Group 2, $n = 29$). Our data are coded so that higher scores indicate a more positive mood. We have good reason to believe that people leaving the comedy will be in a better mood, so we use a one-tailed test at $\alpha = .05$ to test our hypothesis.

Step 1: State the Hypotheses

As always, we start with hypotheses:

$$H_0 : \text{There is no difference in average mood between the two movie types}$$
$$H_0 : \mu_1 - \mu_2 = 0$$
$$\text{or}$$
$$H_0 : \mu_1 = \mu_2$$

$$H_A : \text{The comedy film will produce a better average mood than the horror film}$$
$$H_A : \mu_1 - \mu_2 > 0$$
$$\text{or}$$
$$H_A : \mu_1 > \mu_2$$

Notice that in the first formulation of the alternative hypothesis we say that the first mean minus the second mean will be greater than zero. This is based on how we code the data (higher is better), so we suspect that the mean of the first group will be higher. Thus, we will have a larger number minus a smaller number, which will be greater than zero. Be sure to pay attention to which group is which and how your data are coded (higher is almost always used as better outcomes) to make sure your hypothesis makes sense!

Step 2: Find the Critical Values

Just like before, we will need critical values, which come from our t table. In this example, we have a one-tailed test at $\alpha = .05$ and expect a positive answer (because we expect the difference between the means to be greater than zero). Our degrees of freedom for our independent samples t test is just the degrees of freedom from each group added together: 35 + 29 – 2 = 62. From our t table, we find that our critical value is $t^* = 1.671$. Note that because 62 does not appear on the table, we use the next lowest value, which in this case is 60.

Step 3: Compute the Test Statistic and Effect Size

The data from our two groups are presented in the tables below. Table 10.1 shows the values for the Comedy group, and Table 10.2 shows the values for the Horror group. Values for both have already been placed in the sum of squares tables since we will need to use them for our further calculations. As always, the column on the left is our raw data.

Table 10.1. Raw scores and sum of squares for Group 1 (comedy).

X	$X - M$	$(X - M)^2$
39.10	15.10	228.01

X	$X - M$	$(X - M)^2$
38.00	14.00	196.00
14.90	−9.10	82.81
20.70	−3.30	10.89
19.50	−4.50	20.25
32.20	8.20	67.24
11.00	−13.00	169.00
20.70	−3.30	10.89
26.40	2.40	5.76
35.70	11.70	136.89
26.40	2.40	5.76
28.80	4.80	23.04
33.40	9.40	88.36
13.70	−10.30	106.09
46.10	22.10	488.41
13.70	−10.30	106.09
23.00	−1.00	1.00

X	$X - M$	$(X - M)^2$
20.70	−3.30	10.89
19.50	−4.50	20.25
11.40	−12.60	158.76
24.10	0.10	0.01
17.20	−6.80	46.24
38.00	14.00	196.00
10.30	−13.70	187.69
35.70	11.70	136.89
41.50	17.50	306.25
18.40	−5.60	31.36
36.80	12.80	163.84
54.10	30.10	906.01
11.40	−12.60	158.76
8.70	−15.30	234.09
23.00	−1.00	1.00
14.30	−9.70	94.09

X	$X - M$	$(X - M)^2$
5.30	–18.70	349.69
6.30	–17.70	313.29
$\Sigma = 840$	$\Sigma = 0$	$\Sigma = 5061.60$

Table 10.2. Raw scores and sum of squares for Group 2 (horror).

X	$X - M$	$(X - M)^2$
24.00	7.50	56.25
17.00	0.50	0.25
35.80	19.30	372.49
18.00	1.50	2.25
–1.70	–18.20	331.24
11.10	–5.40	29.16
10.10	–6.40	40.96
16.10	–0.40	0.16
–0.70	–17.20	295.84
14.10	–2.40	5.76
25.90	9.40	88.36

X	$X - M$	$(X - M)^2$
23.00	6.50	42.25
20.00	3.50	12.25
14.10	−2.40	5.76
−1.70	−18.20	331.24
19.00	2.50	6.25
20.00	3.50	12.25
30.90	14.40	207.36
30.90	14.40	207.36
22.00	5.50	30.25
6.20	−10.30	106.09
27.90	11.40	129.96
14.10	−2.40	5.76
33.80	17.30	299.29
26.90	10.40	108.16
5.20	−11.30	127.69
13.10	−3.40	11.56

X	$X - M$	$(X - M)^2$
19.00	2.50	6.25
−15.50	−32.00	1024.00
$\Sigma = 478.6$	$\Sigma = 0.10$	$\Sigma = 3896.45$

Using the sum of the first column for each table, we can calculate the mean for each group:

$$M_1 = \frac{840}{35} = 24.00$$

and

$$M_2 = \frac{478.60}{29} = 16.50$$

These values were used to calculate the middle rows of each table, which sum to zero as they should (the middle column for Group 2 sums to a very small value instead of zero due to rounding error—the exact mean is 16.50344827586207, but that's far more than we need for our purposes). Squaring each of the deviation scores in the middle columns gives us the values in the third columns, which sum to our next important value: the sum of squares for each group: $SS_1 = 5061.60$ and $SS_2 = 3896.45$. Before we move on to the pooled variance that will allow us to calculate standard error, let's compute our standard deviation for each group; even though we will not use them in our calculation of the test statistic, they are still important descriptors of our data:

$$s_1 = \sqrt{\frac{5061.60}{34}} = 12.20$$

and

$$s_2 = \sqrt{\frac{3896.45}{28}} = 11.80$$

Now we can move on to our new calculation, the pooled variance, which is just the sums of squares that we calculated from our table and the degrees of freedom, which is just $n - 1$ for each group:

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{5061.60 + 3896.45}{34 + 28} = \frac{8958.05}{62} = 144.48$$

As you can see, if you follow the regular process of calculating standard deviation using the sum of squares table, finding the pooled variance is very easy. Now we can use that value to calculate our standard error, the last step before we can find our test statistic:

$$s_{M_1 - M_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{144.48}{35} + \frac{144.48}{29}} = \sqrt{4.13 + 4.98} = \sqrt{9.11} = 3.02$$

Finally, we can use our standard error and the means we calculated earlier to compute our test statistic. Because the null hypothesis value of $\mu_1 - \mu_2$ is 0.00, we will leave that portion out of the equation for simplicity:

$$t = \frac{(M_1 - M_2)}{s_{M_1 - M_2}} = \frac{24.00 - 16.50}{3.02} = 2.48$$

The process of calculating our obtained test statistic $t = 2.48$ followed the same sequence of steps as before: use raw data to compute the mean and sum of squares (this time for two groups instead of one), use the sum of squares and degrees of freedom to calculate standard error (this time using pooled variance instead of standard deviation), and use that standard error and the observed means to get t.

Effect Sizes and Confidence Intervals

Even a statistically significant effect needs to be interpreted along with an effect size to see if it is practically meaningful. We have also seen that our sample means, as a point estimate, are not perfect and would be better represented by a range of values that we call a confidence interval. As with all other topics, this is also true of our independent samples t tests.

Our effect size for the independent samples t test is still Cohen's d, and it is still just our observed effect divided by the standard deviation. Remember that standard deviation is just the square root of the variance, and because we work with pooled variance in our test statistic, we will use the square root of the pooled variance as our denominator in the formula for Cohen's d. This gives us:

$$d = \frac{M_1 - M_2}{\sqrt{s_p^2}}$$

For our example above, we can calculate the effect size to be:

$$d = \frac{24.00 - 16.50}{\sqrt{144.48}} = \frac{7.50}{12.02} = 0.62$$

We interpret this using the same guidelines as before, so we would consider this a moderate or moderately large effect.

Our confidence intervals also take on the same form and interpretation as they have in the past. The value we are interested in is the difference between the two means, so our point estimate is the value of one mean minus the other, or $M_1 - M_2$. Just like before, this is our observed effect and is the same value as the one we place in the numerator of our test statistic. We calculate this value then place the margin of error—still our critical value multiplied by our standard error—above and below it. That is:

$$\text{Confidence interval } = (M_1 - M_2) \pm t^* \left(s_{M_1 - M_2}\right)$$

A one-tailed hypothesis test does not prevent us from calculating a confidence interval. We can report a one-sided interval corresponding to the test, or a two-sided interval if we clearly identify it. For a separate example of a two-sided interval, let's use summary statistics on the average life satisfaction of people from two different towns.

Our sample data are $M_1 = 28.65, s_1 = 12.40, n_1 = 40$ and $M_2 = 25.40, s_2 = 15.68, n_2 = 42$. At face value, it looks like the people from the first town have higher life satisfaction (28.65 vs. 25.40), but it will take a confidence interval (or complete hypothesis testing process) to see if that is true or just due to random chance. First, we want to calculate the difference between our sample means, which is 28.65 – 25.40 = 3.25. Next, we need a critical value from our t table. If we want to test at the normal 95% level of confidence, then our sample sizes will yield degrees of freedom equal to 40 + 42 – 2 = 80. From our table, that gives us a critical value of $t^* = 1.990$. Finally, we need our standard error. Recall that our standard error for an independent samples t test uses pooled variance, which requires the sum of squares and degrees of freedom. Up to this point, we have calculated the sum of squares using raw data, but in this situation, we do not have access to it. So, what are we to do?

If we have summary data like standard deviation and sample size, it is very easy to calculate the pooled variance, and the key lies in rearranging the formulas to work backward through them. We need the sum of squares and degrees of freedom to calculate our pooled variance. Degrees of freedom is very simple: we just take the sample size minus 1.00 for each group. Getting the sum of squares is also easy: remember that variance is standard deviation squared and is the sum of squares divided by the degrees of freedom. That is:

$$s^2 = (s)^2 = \frac{SS}{df}$$

To get the sum of squares, we just multiply both sides of the above equation to get:

$$(s)^2(df) = SS$$

which is the squared standard deviation multiplied by the degrees of freedom $(n-1)$ equals the sum of squares.

Using our example data:

$$(s_1)^2\,(df_1) = SS_1$$
$$(12.40)^2(40-1) = 5996.64$$

$$(s_2)^2\,(df_2) = SS_2$$
$$(15.68)^2(42-1) = 10080.36$$

And thus our pooled variance equals:

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{5996.64 + 10080.36}{39 + 41} = \frac{16077}{80} = 200.96$$

And our standard error equals:

$$s_{M_1-M_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{200.96248}{40} + \frac{200.96248}{42}} = \sqrt{9.808883} \approx 3.13 \tag{10.4.1}$$

All of these steps are just slightly different ways of using the same formulas, numbers, and ideas we have worked with up to this point. Once we get out standard error, it's time to build our confidence interval.

$$95\%\ \text{CI} = 3.25 \pm 1.990(3.1319) \tag{10.4.2}$$
$$\text{Upper bound} = 3.25 + 6.23 = 9.48 \tag{10.4.3}$$
$$\text{Lower bound} = 3.25 - 6.23 = -2.98 \tag{10.4.4}$$
$$95\%\ \text{CI} = (-2.98,\ 9.48) \tag{10.4.5}$$

Our confidence interval represents a range of plausible values for the population mean difference based on our observed data. In this instance, our interval (–2.98, 9.48) contains zero. These data therefore do not provide sufficient evidence of a difference in mean life satisfaction between the two towns.

Now we can move on to the final step of the hypothesis-testing procedure.

Step 4: Make the Decision

Our test statistic has a value of $t = 2.48$, and in Step 2 we found that the critical value is $t^* = 1.671$. Because 2.48 > 1.671, we reject the null hypothesis:

Reject H_0. Based on our sample data from people who watched different kinds of movies, we can say that the average mood after a comedy movie ($M_1 = 24.00, SD_1 = 12.20$) is better than the average mood after a horror movie ($M_2 = 16.50, SD_2 = 11.80$), $t(62) = 2.48, p < .05, d = 0.62$.

The results calculated from the Movies and Mood raw scores are shown below. The one-sided independent-samples t-test indicates that people leaving the comedy reported a more positive average mood (M = 24.00, SD = 12.20) than people leaving the horror film (M = 16.50, SD = 11.80), t(62) = 2.48, p = .008, d = .62.

Group	N	Mean	SD
Comedy	35	24.000	12.201
Horror	29	16.503	11.797

Figure 10.1. Output from JASP for the independent-samples t test described in the Movies and Mood example. The output provides the t value (2.478), degrees of freedom (62), the exact p value (.008, which is less than .05), and Cohen's d for effect size (0.622). Note that the means and standard deviations for both samples are also provided. Based on our sample

data from people who watched different kinds of movies, we can say that the average mood after a comedy movie ($M = 24.0, SD = 12.2$) is better than the average mood after a horror movie ($M = 16.5; SD = 11.8$), $t(62) = 2.48, p = .008, d = 0.62$. ("JASP independent-samples t test" by Rupa G. Gordon/Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

This page titled 10.4: Worked Example- Calculating an Independent Samples t-Test is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida via source content that was edited to the style and standards of the LibreTexts platform.

- 12.3: Isoquants by Staff is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-economic-analysis/.

11: CORRELATION AND LINEAR REGRESSION

11.1: Correlation and linear regression
11.2: Pearson and Spearman Correlations
11.3: Calculating and Reporting Pearson's r

This page titled 11: Correlation and Linear Regression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

11.1: Correlation and linear regression

The goal in this chapter is to introduce correlation and linear regression. These are the standard tools that statisticians rely on when analysing the relationship between continuous predictors and continuous outcomes.

11.1 Correlations

In this section we'll talk about how to describe the relationships between variables in the data. To do that, we want to talk mostly about the correlation between variables. But first, we need some data.
11.1.1 The data

Table 11.1: Descriptive statistics for the parenthood data.

variable	min	max	mean	median	std. dev	IQR
Dan's grumpiness	41	91	63.71	62	10.05	14
Dan's hours slept	4.84	9.00	6.97	7.03	1.02	1.45
Dan's son's hours slept	3.25	12.07	8.05	7.95	2.07	3.21

Let's turn to a topic close to every parent's heart: sleep. The data set we'll use is fictitious, but based on real events. Suppose I'm curious to find out how much my infant son's sleeping habits affect my mood. Let's say that I can rate my grumpiness very precisely, on a scale from 0 (not at all grumpy) to 100 (grumpy as a very, very grumpy old man or woman). And lets also assume that I've been measuring my grumpiness, my sleeping patterns and my son's sleeping patterns for
quite some time now. Let's say, for 100 days. And, being a nerd, I've saved the data as a file called parenthood.csv. If we load the data into JASP we can see that the file contains four variables dan.sleep, baby.sleep, dan.grump and day. Note that when you first load this data set JASP may not have guessed the data type for each variable correctly, in which case you should fix it: dan.sleep, baby.sleep, dan.grump and day can be specified as continuous variables, and ID is a nominal(integer) variable.

Next, I'll take a look at some basic descriptive statistics and, to give a graphical depiction of what each of the three interesting variables looks like, Figure 11.1 plots histograms. One thing to note: just because JASP can calculate dozens of different statistics doesn't mean you should report all of them. If I were writing this up for a report, I'd probably pick out those statistics that are of most interest to me (and to my readership), and then put them into a nice, simple table like the one in Table 11.1. [1] Notice that when I put it into a table, I gave everything "human readable" names. This is always good practice. Notice also that I'm not getting enough sleep. This isn't good practice, but other parents tell me that it's pretty standard.

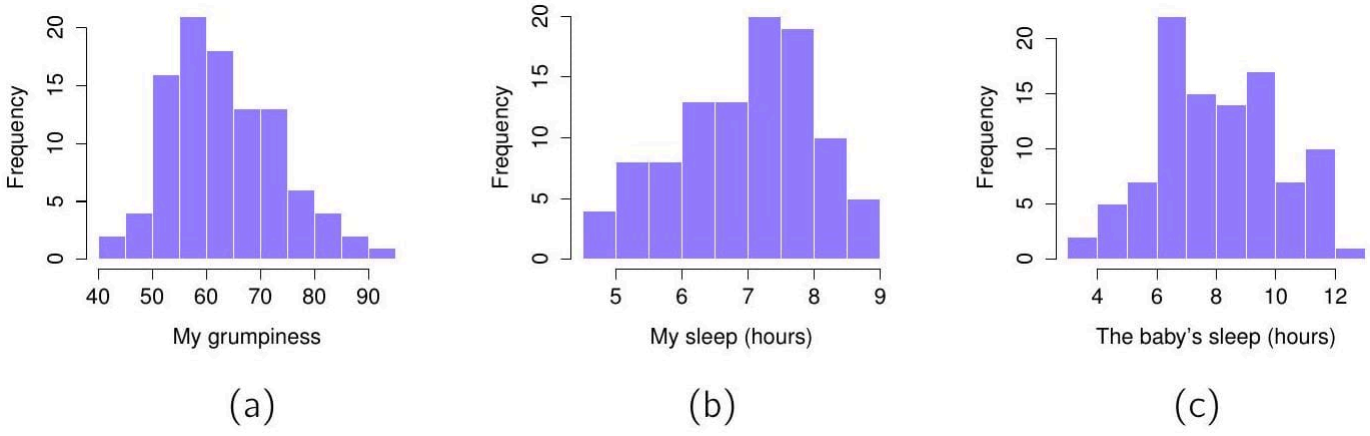


Figure 11.1: Histograms for the three interesting variables in the parenthood data set.

11.1.2 The strength and direction of a relationship

We can draw scatterplots to give us a general sense of how closely related two variables are. Ideally though, we might want to say a bit more about it than that. For instance, let's compare the relationship between dan.sleep and dan.grump (Figure 11.2, left) with that between baby.sleep and dan.grump (Figure 11.2, right). When looking at these two plots side by side, it's clear that the relationship is qualitatively the same in both cases: more sleep equals less grump! However, it's also pretty obvious that the relationship between dan.sleep and dan.grump is stronger than the relationship between baby.sleep and dan.grump. The plot on the left is "neater" than the one on the right. What it feels like is that if you want to predict what my mood is, it'd help you a little bit to know how many hours my son slept, but it'd be more helpful to know how many hours I slept.

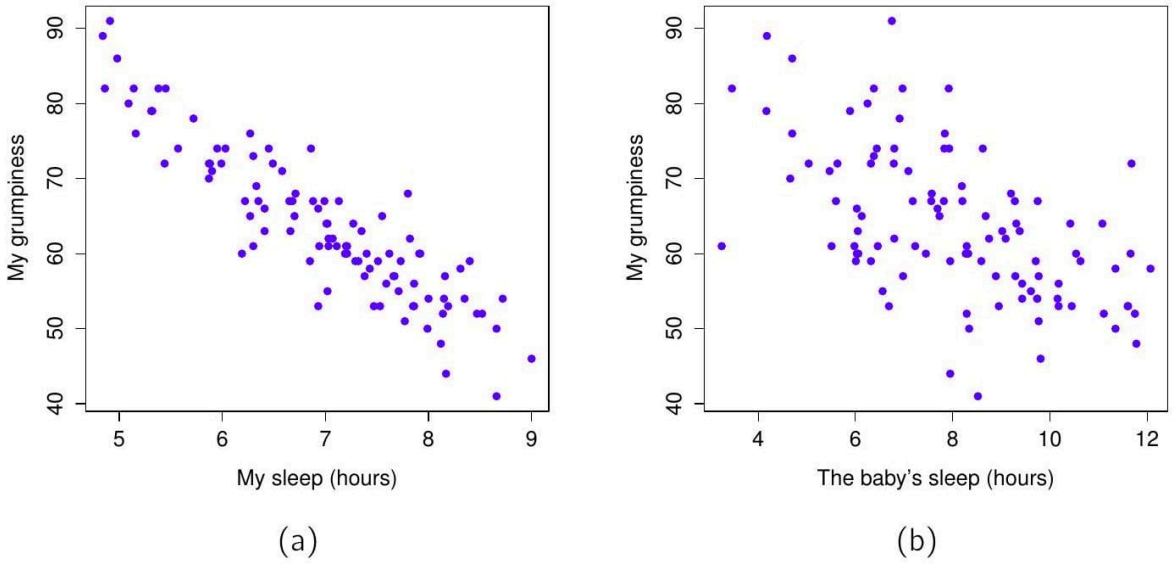


Figure 11.2: Scatterplots showing the relationship between dan.sleep and dan.grump (left) and the relationship between baby.sleep and dan.grump (right).

In contrast, let's consider the two scatterplots shown in Figure 11.3. If we compare the scatterplot of "baby.sleep v dan.grump" (left) to the scatterplot of "'baby.sleep v dan.sleep" (right), the overall strength of the relationship is the same, but the direction is different. That is, if my son sleeps more, I get more sleep (positive relationship, right hand side), but if he sleeps more then I get less grumpy (negative relationship, left hand side).

11.1.3 The correlation coefficient

We can make these ideas a bit more explicit by introducing the idea of a correlation coefficient (or, more specifically, Pearson's correlation coefficient), which is traditionally denoted as r. The correlation coefficient between two variables X and Y (sometimes denoted r_{XY}), which we'll define more precisely in the next section, is a measure that varies from -1 to 1. When $r = -1$ it means that we have a perfect negative relationship, and when $r = 1$ it means we have a perfect positive relationship. When $r = 0$, there is no linear relationship. A nonlinear relationship may still be present. If you look at Figure 11.4, you can see several plots showing what different correlations look like.

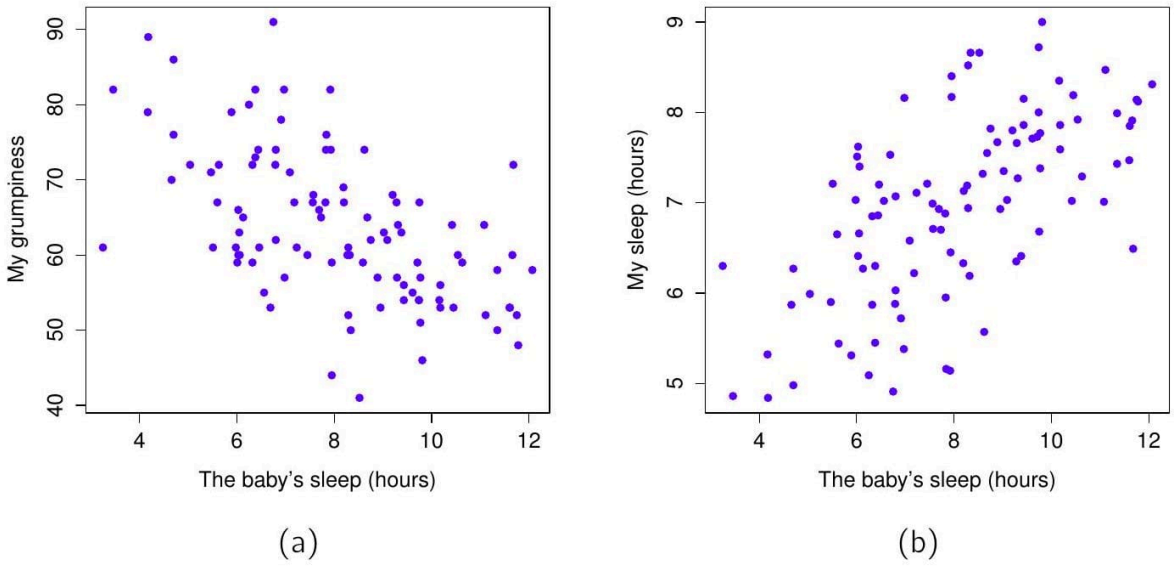
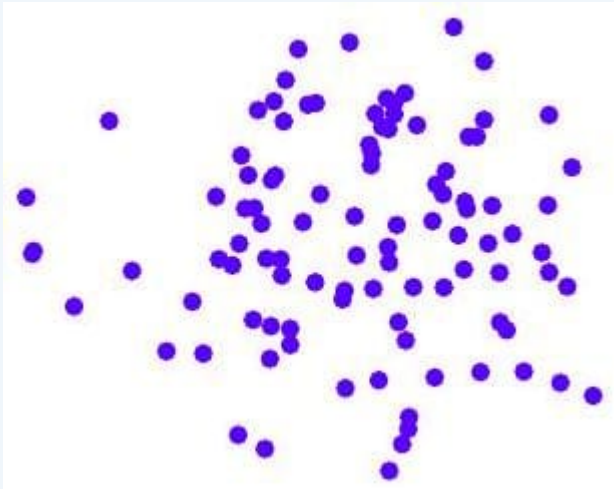


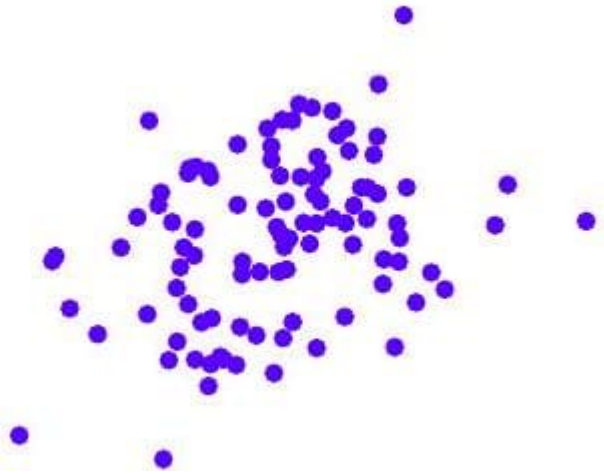
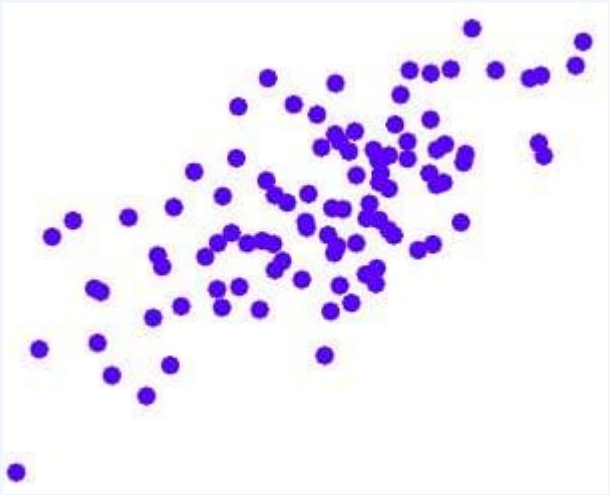
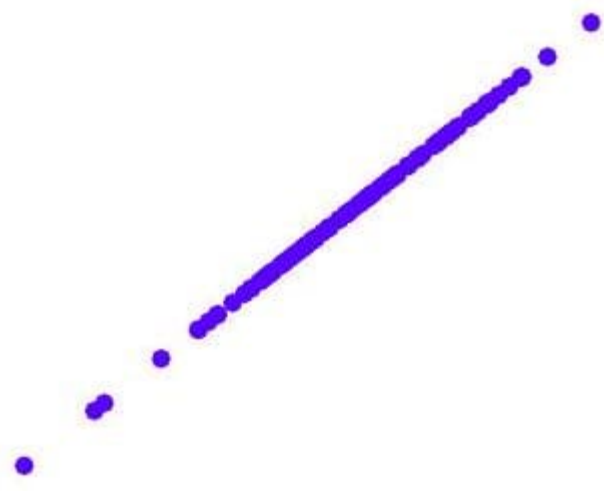
Figure 11.3: Scatterplots showing the relationship between baby.sleep and dan.grump (left), as compared to the relationship between baby.sleep and dan.sleep (right).

The formula for the Pearson's correlation coefficient can be written in several different ways. I think the simplest way to write down the formula is to break it into two steps. Firstly, let's introduce the idea of a covariance. The covariance between two variables X and Y is a generalisation of the notion of the variance amd is a mathematically simple way of describing the relationship between two variables that isn't terribly informative to humans

$$\mathrm{Cov}(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

Because we're multiplying a quantity that depends on X by a quantity that depends on Y and then averaging, you can think of the formula for the covariance as an "average cross product" between X and Y.

positive correlations		negative correlations
0.0		0.0

0.33		-0.33
0.66		-0.66
1.0		-1.0

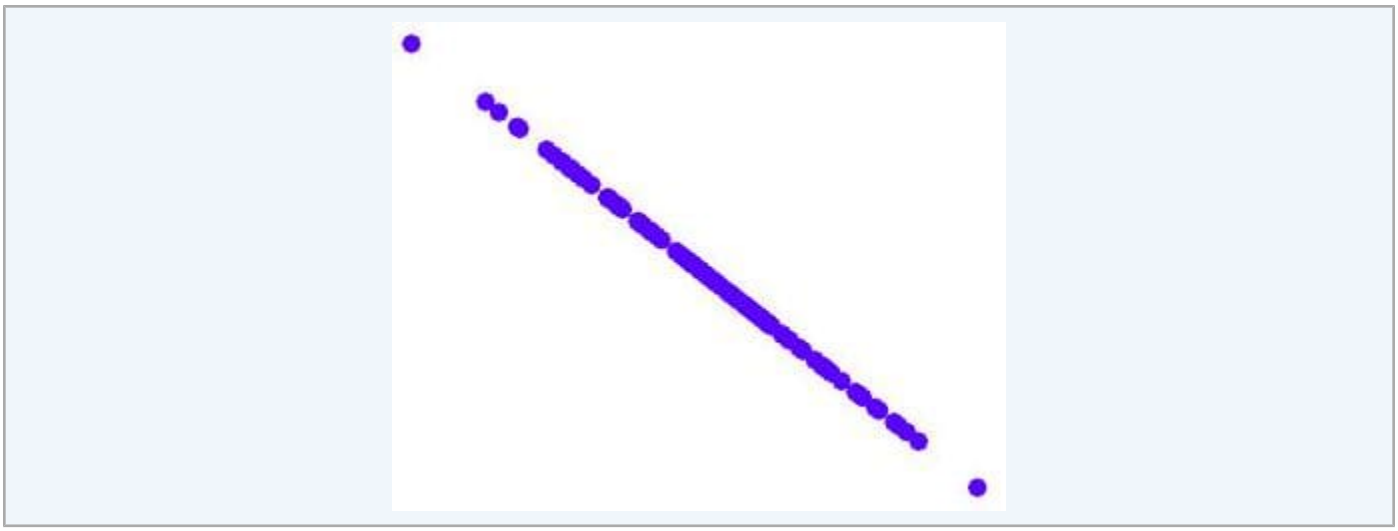


Figure 11.4: Illustration of the effect of varying the strength and direction of a correlation. In the left hand column, the correlations are 0, .33, . 66 and 1. In the right hand column, the correlations are 0, -.33, -. 66 and -1.

The covariance has the nice property that, if X and Y are entirely unrelated, then the covariance is exactly zero. If the relationship between them is positive (in the sense shown in Figure 11.4) then the covariance is also positive, and if the relationship is negative then the covariance is also negative. In other words, the covariance captures the basic qualitative idea of correlation. Unfortunately, the raw magnitude of the covariance isn't easy to interpret as it depends on the units in which X and Y are expressed and, worse yet, the actual units that the covariance itself is expressed in are really weird. For instance, if X refers to the dan.sleep variable (units: hours) and Y refers to the dan.grump variable (units: grumps), then the units for their covariance are "hours × grumps". And I have no freaking idea what that would even mean.

The Pearson correlation coefficient r fixes this interpretation problem by standardising the covariance, in pretty much the exact same way that the z-score standardises a raw score, by dividing by the standard deviation. However, because we have two variables that contribute to the covariance, the standardisation only works if we divide by both standard deviations. [b] In other words, the correlation between X and Y can be written as follows:

$$r_{XY} = \frac{\text{Cov}(X, Y)}{\hat{\sigma}_X \hat{\sigma}_Y}$$

[a] Just like we saw with the variance and the standard deviation, in practice we divide by $N - 1$ rather than N.
[b] This is an oversimplification, but it'll do for our purposes.
By standardising the covariance, not only do we keep all of the nice properties of the covariance discussed earlier, but the actual values of r are on a meaningful scale: $r = 1$ implies a perfect positive relationship and $r = -1$ implies a perfect negative relationship. I'll expand a little more on this point later, in Section 11.1.5. But before I do, let's look at how to calculate correlations in JASP.

11.1.4 Calculating correlations in JASP

Calculating correlations in JASP can be done by clicking on the 'Regression → Classical → Correlation. Transfer all four continuous variables across into the box on the right to get the output in Figure 11.5. Notice that each correlation (denoted 'Pearson's r ') is paired with a p-value. Clearly, something is being tested here, but ignore it for now. We'll talk more about that soon!

11.1.5 Interpreting a correlation

Naturally, in real life you don't see many correlations of 1. So how should you interpret a correlation of, say, $r = .4$? The honest answer is that it really depends on what you want to use the data for, and on how strong the correlations in your field tend to be. A friend of mine in engineering once argued that any correlation less than . 95 is completely useless (I think he was exaggerating, even for engineering).

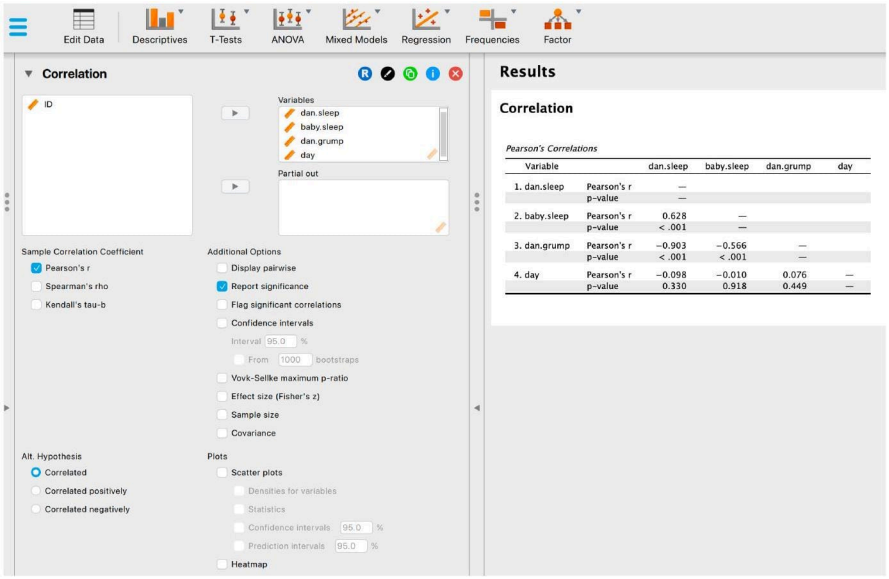


Figure 11.5: A JASP screenshot showing correlations between variables in the parenthood.csv file

On the other hand, there are real cases, even in psychology, where you should really expect correlations that strong. For instance, one of the benchmark data sets used to test theories of how people judge similarities is so clean that any theory that can't achieve a correlation of at least .9 really isn't deemed to be successful. However, when looking for (say) elementary correlates of intelligence (e.g., inspection time, response time), if you get a correlation above . 3 you're doing very very well. In short, the interpretation of a correlation depends a lot on the context. That said, the rough guide in Table 11.2 is pretty typical.

However, something that can never be stressed enough is that you should always look at the scatterplot before attaching any interpretation to the data. A correlation might not mean what you think it means. The classic illustration of this is "Anscombe's Quartet" (Anscombe 1973), a collection of four data sets. Each data set has two variables, an X and a Y. For all four data sets the mean value for X is 9 and the mean for Y is 7.5. The standard deviations for all X variables are almost identical, as are those for the Y variables. And in each case the correlation between X and Y is $r = 0.816$. You can verify this yourself, since I happen to have saved it in a file called anscombe.csv.

You'd think that these four data sets would look pretty similar to one another. They do not. If we draw scatterplots of X against Y for all four variables, as shown in Figure 11.6, we see that all four of these are spectacularly different to each other. The lesson here, which so very many people seem to forget in real life, is "always graph your raw data."

Table 11.2: A rough guide to interpreting correlations. Note that I say a rough guide. There aren't hard and fast rules for what counts as strong or weak relationships. It depends on the context.

Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive

11.1.6 Spearman's rank correlations

The Pearson correlation coefficient is useful for a lot of things, but it does have shortcomings. One issue in particular stands out: what it actually measures is the strength of the linear relationship between two variables. In other words, what it gives you is a measure of the extent to which the data all tend to fall on a single, perfectly straight line. Often, this is a pretty good approximation to what we mean when we say "relationship", and so the Pearson correlation is a good thing to calculate. Sometimes though, it isn't.

One very common situation where the Pearson correlation isn't quite the right thing to use arises when an increase in one variable X really is reflected in an increase in another variable Y, but the nature of the relationship isn't necessarily linear. An example of this might be the relationship between effort and reward when studying for an exam. If you put zero effort (X) into learning a subject then you should expect a grade of 0% (Y). However, a little bit of effort will cause a massive improvement. Just turning up to lectures means that you learn a fair bit, and if you just turn up to classes and scribble a few things down your grade might rise to 35%, all without a lot of effort. However, you just don't get the same effect at the other end of the scale. As everyone knows, it takes a lot more effort to get a grade of 90% than it takes to get a grade of 55%. What this means is that, if I've got data looking at study effort and grades, there's a pretty good chance that Pearson correlations will be misleading.

To illustrate, consider the data plotted in Figure 11.7, showing the relationship between hours worked and grade received for 10 students taking some class. The curious thing about this (highly fictitious) data set is that increasing your effort always increases your grade. It might be by a lot or it might be by a little, but increasing effort will never decrease your grade.

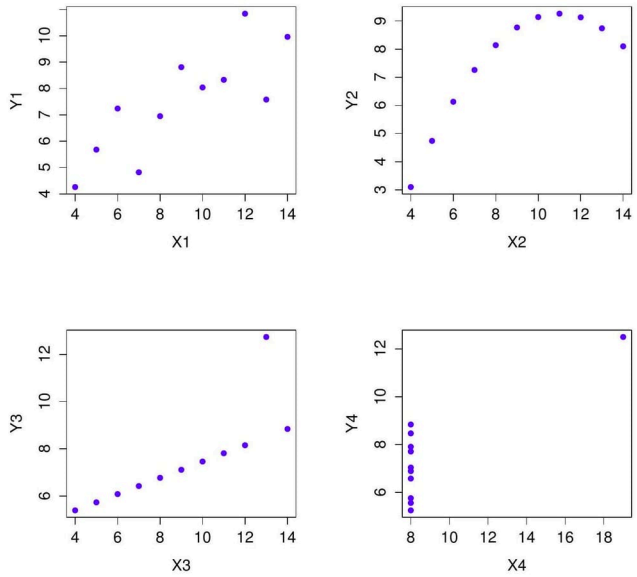


Figure 11.6: Anscombe's quartet. All four of these data sets have a Pearson correlation of $r = .816$, but they are qualitatively different from one another.

If we run a standard Pearson correlation, it shows a strong relationship between hours worked and grade received, with a correlation coefficient of 0.91. However, this doesn't actually capture the observation that increasing hours worked always increases the grade. There's a sense here in which we want to be able to say that the correlation is perfect but for a somewhat different notion of what a "relationship" is. What we're looking for is something that captures the fact that there is a perfect ordinal relationship here.

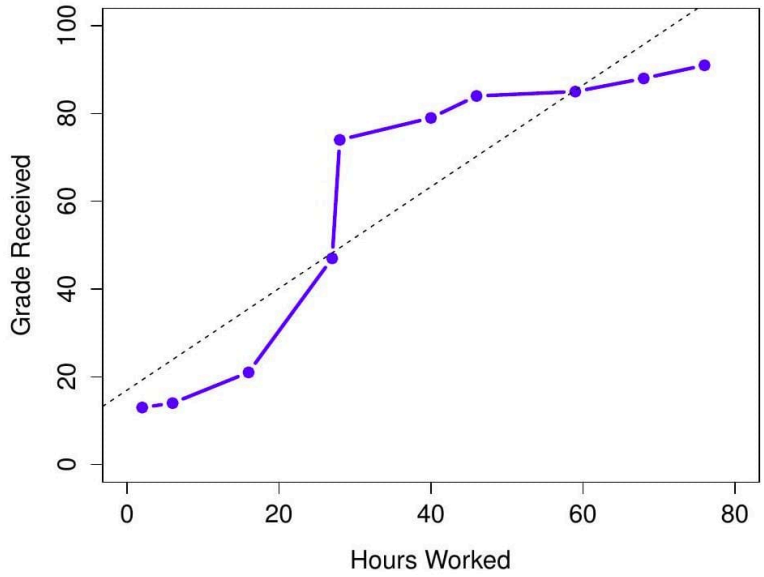


Figure 11.7: The relationship between hours worked and grade received for a toy data set consisting of only 10 students (each circle corresponds to one student). The dashed line through the middle shows the linear relationship between the two variables. This produces a strong Pearson correlation of $r = .91$. However, the interesting thing to note here is that there's actually a perfect monotonic relationship between the two variables. In this toy example, increasing the hours worked always increases the grade received, as illustrated by the solid line. This is reflected in a Spearman correlation of $\rho = 1$. With such a small data set, however, it's an open question as to which version better describes the actual relationship involved.

That is, if student 1 works more hours than student 2 , then we can guarantee that student 1 will get the better grade. That's not what a correlation of $r = .91$ says at all.

How should we address this? Actually, it's really easy. If we're looking for ordinal relationships all we have to do is treat the data as if it were ordinal scale! So, instead of measuring effort in terms of "hours worked", lets rank all 10 of our students in order of hours worked. That is, student 1 did the least work out of anyone (2 hours) so they get the lowest rank (rank $= 1$). Student 4 was the next laziest, putting in only 6 hours of work over the whole semester, so they get the next lowest rank (rank $= 2$). Notice that I'm using "rank $= 1$ " to mean "low rank". Sometimes in everyday language we talk about "rank = 1" to mean "top rank" rather than "bottom rank". So be careful, you can rank "from smallest value to largest value" (i.e., small equals rank 1) or you can rank "from largest value to smallest value" (i.e., large equals rank 1). In this case, I'm ranking from smallest to largest, but as it's really easy to forget which way you set things up you have to put a bit of effort into remembering!

Okay, so let's have a look at our students when we rank them from worst to best in terms of effort and reward:

	rank (hours worked)	rank (grade received)
student 1	1	1
student 2	10	10
student 3	6	6
student 4	2	2
student 5	3	3
student 6	5	5
student 7	4	4
student 8	8	8
student 9	7	7
student 10	9	9

Hmm. These are identical. The student who put in the most effort got the best grade, the student with the least effort got the worst grade, etc. As the table above shows, these two rankings are identical, so if we now correlate them we get a perfect relationship, with a correlation of 1.0.

What we've just re-invented is Spearman's rank order correlation, usually denoted ρ to distinguish it from the Pearson correlation r. We can calculate Spearman's ρ using JASP simply by clicking the 'Spearman' check box in the 'Correlation Matrix' screen.

11.2 Scatterplots

Scatterplots are a simple but effective tool for visualising the relationship between two variables, like we saw with the figures in the section on correlation (Section 11.1). It's this latter application that we usually have in mind when we use the term "scatterplot". In this kind of plot each observation corresponds to one dot. The horizontal location of the dot plots the value of the observation on one variable, and the vertical location displays its value on the other variable. In many situations you don't really have a clear opinions about what the causal relationship is (e.g., does A cause B, or does B cause A, or does some other variable C control both A and B). If that's the case, it doesn't really matter which variable you plot on the x-axis and which one you plot on the y-axis. However, in many situations you do have a pretty strong idea which variable you think is most likely to be causal, or at least you have some suspicions in that direction. If so, then it's conventional to plot the cause variable on the x-axis, and the effect variable on the y-axis. With that in mind, let's look at how to draw scatterplots in JASP, using the same parenthood data set (i.e. parenthood.csv) that I used when introducing correlations.

Suppose my goal is to draw a scatterplot displaying the relationship between the amount of sleep that I get (dan.sleep) and how grumpy I am the next day (dan.grump). The way in which we can use JASP to get this plot is to use the 'Plots' option under the 'Regression' - 'Correlation Matrix' button, giving us the output shown in Figure 11.8.

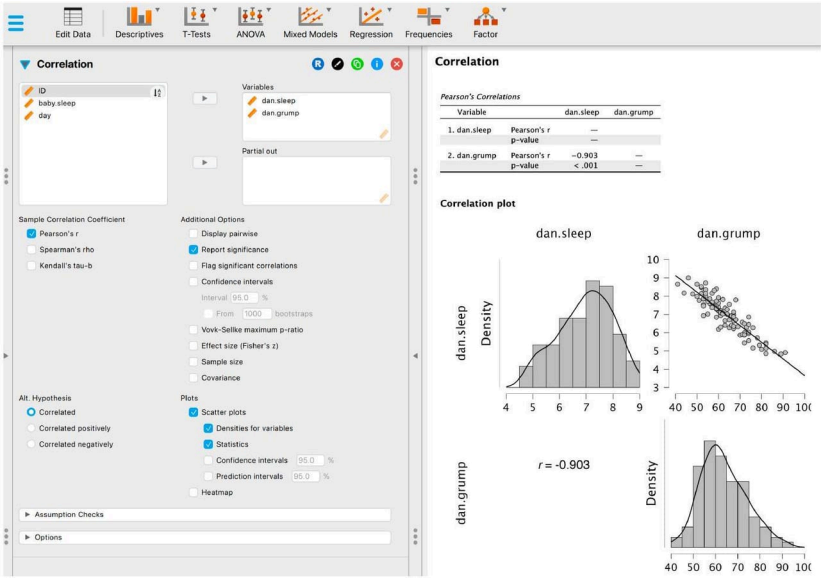


Figure 11.8: Scatterplot via the 'Correlation Matrix' method in JASP

Note that JASP draws a line through the points, we'll come onto this a bit later in Section (11.3). Plotting a scatterplot in this way also allow you to specify 'Densities for variables', which adds a histogram and density curve showing how the data in each variable is distributed. You can also specify the 'Statistics' option, which provides an estimate of the correlation along with a 95% confidence interval.

11.3 What is a linear regression model?

Stripped to its bare essentials, linear regression models are basically a slightly fancier version of the Pearson correlation (Section 11.1), though as we'll see regression models are much more powerful tools.

Since the basic ideas in regression are closely tied to correlation, we'll return to the parenthood.csv file that we were using

to illustrate how correlations work. Recall that, in this data set we were trying to find out why Dan is so very grumpy all the time and our working hypothesis was that I'm not getting enough sleep. We drew some scatterplots to help us examine the relationship between the amount of sleep I get and my grumpiness the following day, as in Figure 11.8, and as we saw previously this corresponds to a correlation of $r = -.90$, but what we find ourselves secretly imagining is something that looks closer to Figure 11.9a. That is, we mentally draw a straight line through the middle of the data. In statistics, this line that we're drawing is called a regression line. Notice that, since we're not idiots, the regression line goes through the middle of the data. We don't find ourselves imagining anything like the rather silly plot shown in Figure 11.9b.

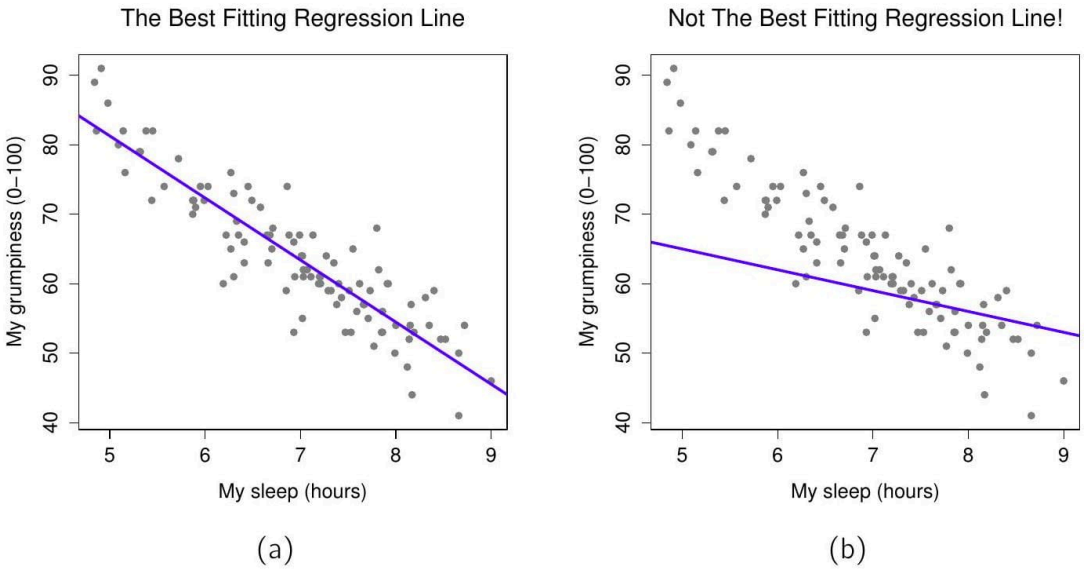


Figure 11.9: Panel a shows the sleep-grumpiness scatterplot from Figure 11.8 with the best fitting regression line drawn over the top. Not surprisingly, the line goes through the middle of the data. In contrast, panel b shows the same data, but with a very poor choice of regression line drawn over the top.

This is not highly surprising. The line that I've drawn in Figure 11.9b doesn't "fit" the data very well, so it doesn't make a lot of sense to propose it as a way of summarising the data, right? This is a very simple observation to make, but it turns out to be very powerful when we start trying to wrap just a little bit of maths around it. To do so, let's start with a refresher of some high school maths. The formula for a straight line is usually written like this

$$y = a + bx$$

Or, at least, that's what it was when I went to high school all those years ago. The two variables are x and y, and we have two coefficients, a and b.[2] The coefficient a represents the y-intercept

of the line, and coefficient b represents the slope of the line. Digging further back into our decaying memories of high school (sorry, for some of us high school was a long time ago), we remember that the intercept is interpreted as "the value of y that you get when $x = 0$ ". Similarly, a slope of b means that if you increase the x-value by 1 unit, then the y-value goes up by b units, and a negative slope means that the y-value would go down rather than up. Ah yes, it's all coming back to me now. Now that we've remembered that it should come as no surprise to discover that we use the exact same formula for a regression line. If Y is the outcome variable (the DV) and X is the predictor variable (the IV), then the formula that describes our regression is written like this

$$\hat{Y}_i = b_0 + b_1 X_i$$

Hmm. Looks like the same formula, but there's some extra frilly bits in this version. Let's make sure we understand them. Firstly, notice that I've written X_i and Y_i rather than just plain old X and Y. This is because we want to remember that we're dealing with actual data. In this equation, X_i is the value of predictor variable for the i th observation (i.e., the number of hours of sleep that I got on day i of my little study), and Y_i is the corresponding value of the outcome variable (i.e., my grumpiness on that day). And although I haven't said so explicitly in the equation, what we're assuming is that this formula works for all observations in the data set (i.e., for all i). Secondly, notice that I wrote $\hat{Y}_i$ and not Y_i. This is because we want to make the distinction between the actual data Y_i, and the estimate $\hat{Y}_i$ (i.e., the prediction that our regression line is making). Thirdly, I changed the letters used to describe

the coefficients from a and b to b_0 and b_1. That's just the way that statisticians like to refer to the coefficients in a regression model. I've no idea why they chose b, but that's what they did. In any case b_0 always refers to the intercept term, and b_1 refers to the slope.

Excellent, excellent. Next, I can't help but notice that, regardless of whether we're talking about the good regression line or the bad one, the data don't fall perfectly on the line. Or, to say it another way, the data Y_i are not identical to the predictions of the regression model $\hat{Y}_i$. Since statisticians love to attach letters, names and numbers to everything, let's refer to the difference between the model prediction and that actual data point as a residual, and we'll refer to it as ε_i.[3] Written using mathematics, the residuals are defined as

$$\varepsilon_i = Y_i - \hat{Y}_i$$

which in turn means that we can write down the complete linear regression model as

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

11.4 Estimating a linear regression model

Okay, now let's redraw our pictures but this time I'll add some lines to show the size of the residual for all observations. When the regression line is good, our residuals (the lengths of the solid black lines) all look pretty small, as shown in Figure 11.10a, but when the regression line is a bad one the residuals are a lot larger, as you can see from looking at Figure 11.10b.

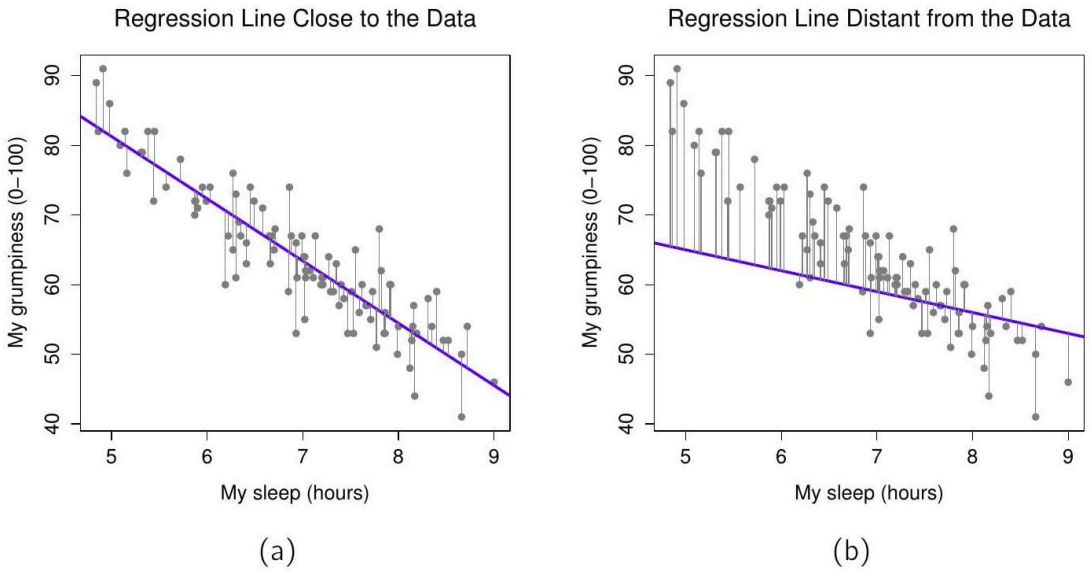


Figure 11.10: A depiction of the residuals associated with the best fitting regression line (panel a), and the residuals associated with a poor regression line (panel b). The residuals are much smaller for the good regression line. Again, this is no surprise given that the good line is the one that goes right through the middle of the data.

Hmm. Maybe what we "want" in a regression model is small residuals. Yes, that does seem to make sense. In fact, I think I'll go so far as to say that the "best fitting" regression line is the one that has the smallest residuals. Or, better yet, since statisticians seem to like to take squares of everything why not say that:

> The estimated regression coefficients, $\hat{b}_0$ and $\hat{b}_1$, are those that minimise the sum of the squared residuals, which we could either write as $\sum_i \left(Y_i - \hat{Y}_i\right)^2$ or as $\sum_i \varepsilon_i^2$.

Yes, yes that sounds even better. And since I've indented it like that, it probably means that this is the right answer. And since this is the right answer, it's probably worth making a note of the fact that our regression coefficients are estimates (we're trying to guess the parameters that describe a population!), which is why I've added the little hats, so that we get $\hat{b}_0$ and $\hat{b}_1$ rather than b_0 and b_1. Finally, I should also note that, since there's actually more than one way to estimate a regression model, the more technical name for this estimation process is ordinary least squares (OLS) regression.

At this point, we now have a concrete definition for what counts as our "best" choice of regression coefficients, $\hat{b}_0$ and $\hat{b}_1$. The natural question to ask next is, if our optimal regression coefficients are those that minimise the sum squared residuals, how do we find these wonderful numbers? The actual
answer to this question is complicated and doesn't help you understand the logic of regression. [4] This time I'm going to let you off the hook. Instead of showing you the long and tedious way first and then "revealing" the wonderful shortcut that JASP provides, let's cut straight to the chase and just use JASP to do all the heavy lifting.

11.4.1 Linear regression in JASP

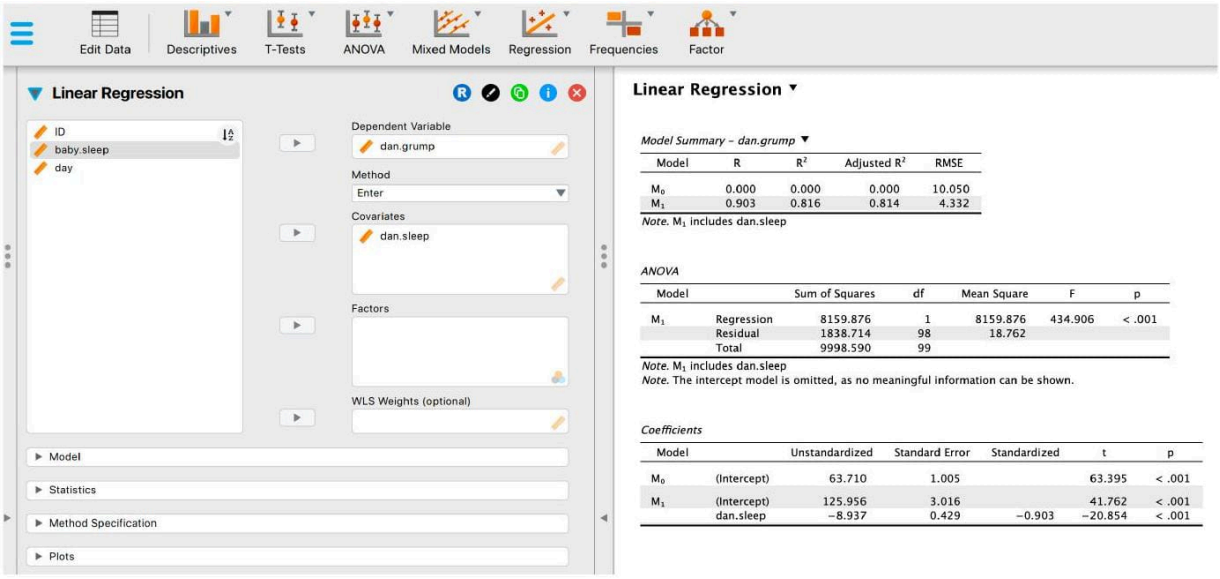


Figure 11.11: A JASP screenshot showing a simple linear regression analysis.

To run my linear regression, open up the 'Regression' - 'Linear Regression' analysis in JASP, using the parenthood.csv data file. Then specify dan.grump as the 'Dependent Variable' and dan.sleep as the variable entered in the 'Covariates' box. This gives the results shown in Figure 11.11, showing an intercept $\hat{b}_0 = 125.956$ and the slope $\hat{b}_1 = -8.937$. In other words, the best-fitting regression line that I plotted in Figure 11.9 has this formula:

$$\hat{Y}_i = 125.956 + (-8.937 X_i)$$

11.4.2 Interpreting the estimated model

The most important thing to be able to understand is how to interpret these coefficients. Let's start with $\hat{b}_1$, the slope. If we remember the definition of the slope, a regression coefficient of $\hat{b}_1 = -8.94$ means that if I increase X_i by 1, then I'm decreasing Y_i by 8.94. That is, each additional hour of sleep is associated with a predicted decrease of 8.94 grumpiness points. What about the intercept? Well, since $\hat{b}_0$ corresponds to "the expected value of Y_i when X_i equals 0", it's pretty straightforward. At zero hours of sleep, the model predicts a grumpiness score of 125.96. However, zero hours is outside the observed range of sleep values, so this prediction is an extrapolation and should not be treated as a reliable estimate.

11.5 Multiple linear regression

The simple linear regression model that we've discussed up to this point assumes that there's a single predictor variable that you're interested in, in this case dan.sleep. In fact, up to this point every statistical tool that we've talked about has assumed that your analysis uses one predictor variable and one outcome variable. However, in many (perhaps most) research projects you actually have multiple predictors that you want to examine. If so, it would be nice to be able to extend the linear regression framework to be able to include multiple predictors. Perhaps some kind of multiple regression model would be in order?

Multiple regression is conceptually very simple. All we do is add more terms to our regression equation. Let's suppose that we've

got two variables that we're interested in; perhaps we want to use both dan.sleep and baby.sleep to predict the dan.grump variable. As before, we let Y_i refer to my grumpiness on the i-th day. But now we have two X variables: the first corresponding to the amount of sleep I got and the second corresponding to the amount of sleep my son got. So we'll let X_{i1} refer to the hours I slept on the i-th day and X_{i2} refers to the hours that the baby slept on that day. If so, then we can write our regression model like this:

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \varepsilon_i$$

As before, ε_i is the residual associated with the i-th observation, $\varepsilon_i = Y_i - \hat{Y}_i$. In this model, we now have three coefficients that need to be estimated: b_0 is the intercept, b_1 is the coefficient associated with my sleep, and b_2 is the coefficient associated with my son's sleep. However, although the number of coefficients that need to be estimated has changed, the basic idea of how the estimation works is unchanged: our estimated coefficients $\hat{b}_0, \hat{b}_1$ and $\hat{b}_2$ are those that minimise the sum squared residuals.

11.5.1 Doing it in JASP

Multiple regression in JASP is no different to simple regression. All we have to do is add additional variables to the 'Covariates' box in JASP. For example, if we want to use both dan.sleep and baby.sleep as predictors in our attempt to explain why I'm so grumpy, then move baby.sleep across into the 'Covariates' box alongside dan.sleep.

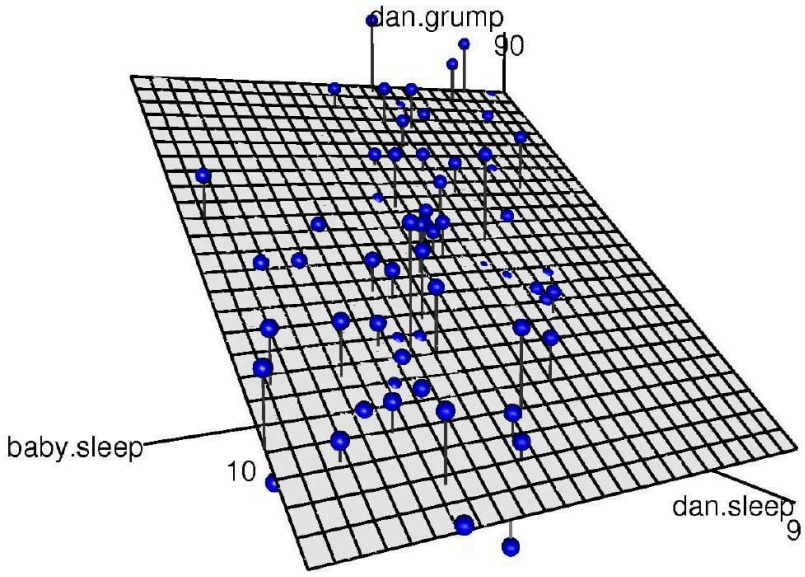


Figure 11.12: A 3D visualisation of a multiple regression model. There are two predictors in the model, dan.sleep and baby.sleep and the outcome variable is dan.grump. Together, these three variables form a 3D space. Each observation (dot) is a point in this space. In much the same way that a simple linear regression model forms a line in 2D space, this multiple regression model forms a plane in 3D space. When we estimate the regression coefficients what we're trying to do is find a plane that is as close to all the blue dots as possible.

By default, JASP assumes that the model should include an intercept. The coefficients we get this time are:

(Intercept)	dan.sleep	baby. sleep
125.966	-8.950	0.011

The coefficient associated with dan.sleep is quite large, suggesting that every hour of sleep I lose makes me a lot grumpier. However, the coefficient for baby.sleep is very small, suggesting that it doesn't really matter how much sleep my son gets. What matters as far as my grumpiness goes is how much sleep / get. To get a sense of what this multiple regression model looks like, Figure 11.12 shows a 3D plot that plots all three variables, along with the regression model itself.

11.5.2 Formula for the general case

The equation that I gave above shows you what a multiple regression model looks like when you include two predictors. Not

surprisingly, then, if you want more than two predictors all you have to do is add more X terms and more b coefficients. In other words, if you have K predictor variables in the model then the regression equation looks like this

$$Y_i = b_0 + \left(\sum_{k=1}^{K} b_k X_{ik} \right) + \varepsilon_i$$

11.6 Quantifying the fit of the regression model

So we now know how to estimate the coefficients of a linear regression model. The problem is, we don't yet know if this regression model is any good. For example, the regression model that we constructed in section 11.4.1 claims that every hour of sleep will improve my mood by quite a lot, but it might just be rubbish. Remember, the regression model only produces a prediction $\hat{Y}_i$ about what my mood is like, but my actual mood is Y_i. If these two are very close, then the regression model has done a good job. If they are very different, then it has done a bad job.

11.6.1 The R^2 value

Once again, let's wrap a little bit of mathematics around this. Firstly, we've got the sum of the squared residuals

$$\text{SS}_{\text{res}} = \sum_i \left(Y_i - \hat{Y}_i \right)^2$$

which we would hope to be pretty small. Specifically, what we'd like is for it to be very small in comparison to the total variability in the outcome variable

$$\text{SS}_{\text{tot}} = \sum_i \left(Y_i - \bar{Y} \right)^2$$

While we're here, let's calculate these values ourselves, not by hand though. I have constructed a JASP file called parenthood_rsquared.jasp, which you can open from the book's data folder. You'll notice that this data file has 5 variables; two of them are the original dan.sleep and dan.grump variables that we've already been using. The other three are calculated variables:

1. Y.pred is the predicted value of grumpiness using the regression equation. It is calculated using the formula ' 125.97 + (-8.94 * dan.sleep)'.
2. resid is a measure of the residual error $\varepsilon_i = Y_i - \hat{Y}_i$, which represents the difference between our predicted value of grumpiness and our actual value of grumpiness. It is calculated using the formula 'dan.grump - Y.pred'.
3. sq.resid is the square of the residual and is calculated using the formula `resid^2`.

Since SS_{res} is the sum of these squared residuals, we can use JASP to find the sum of the sq.resid column. Simply click 'Descriptives' - 'Descriptive Statistics' and move sq.resid to the 'Variables' box. You'll then need to select 'Sum' from the 'Statistics' options below. This should give you a value of '1838.714'.

Wonderful. A big number that doesn't mean very much. Still, let's forge boldly onwards anyway and calculate the total sum of squares as well. That's also pretty simple. Let's calculate SS_{tot} similarly. This time, you'll need to create a new computed column. Click the "+" symbol to start. For 'Name', let's type 'sq.resid2' (you'll see why in a minute). Be sure to select the "R" button, then click 'Create Column'. For your R code, type the following (see Figure 11.13):

```
(dan.grump - mean(dan.grump))^2
```

Then click 'Compute column'. This will produce a column of values that are themselves residuals, but they are residuals (errors) against a really bad predictive model; that is, the model that just predicts grumpiness using the mean of all grumpiness values. To find SS_{tot} , we need to compute the sum of sq.resid2, just like we did above.

This should give you a value of '9998.590'. Hmm. Well, it's a much bigger number than the last one, so this does suggest that our regression model was making good predictions (that is, it has greatly reduced the residual error compared to the model that uses the mean as a single predictor). But it's not very interpretable.

Perhaps we can fix this. What we'd like to do is to convert these two fairly meaningless numbers into one number. A nice,

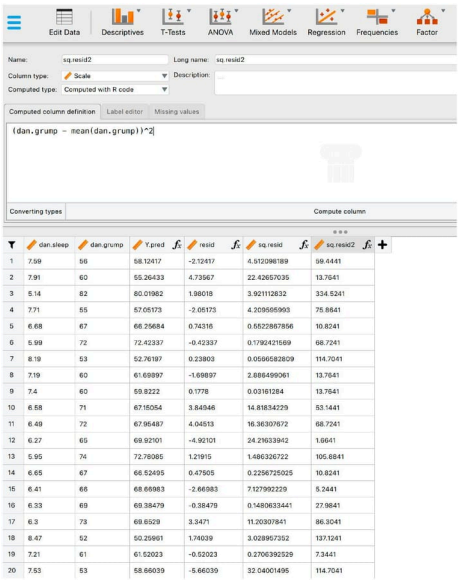
interpretable number, which for no particular reason we'll call R^2. What we would like is for the value of R^2 to be equal to 1 if the regression model makes no errors in predicting the data. In other words, if it turns out that the residual errors are zero. That is, if $\mathrm{SS}_{\mathrm{res}} = 0$ then we expect $R^2 = 1$. Similarly, if the model is completely useless, we would like R^2 to be equal to 0 . What do I mean by "useless"? Tempting as it is to demand that the regression model move out of the house, cut its hair and get a real job, I'm probably going to have to pick a more practical definition. In this case, all I mean is that the residual sum of squares is no smaller than the total sum of squares, $\mathrm{SS}_{\mathrm{res}} = \mathrm{SS}_{\mathrm{tot}}$. Wait, why don't we do exactly that? The formula that provides us with our R^2 value is pretty simple to write down,

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$$

and equally simple to calculate by hand:

$$\begin{aligned} R^2 &= 1 - \frac{SS_{\mathrm{res}}}{SS_{\mathrm{tot}}} \\ &= 1 - \frac{1838.714}{9998.590} \\ &= 1 - 0.1839 \\ &= 0.8161. \end{aligned}$$

The R^2 value, sometimes called the coefficient of determination [5] has a simple interpretation: it is the proportion of the variance in the outcome variable that can be accounted for by the predictor. So, in this case the fact that we have obtained $R^2 = .8161$ means that the predictor (dan.sleep) explains 81.61% of the variance in the outcome (dan.grump).



Y	dan.sleep	dan.grump	Y-Pred $f_(x)$	resid $f_(x)$	sq- resid $f_(x)$	sq-resid2 $f_(x)$
1	7.59	56	58.12417	-2.12417	4.512098	59.4441
2	7.91	60	55.26433	4.73567	22.42657	13.7641
3	5.14	82	80.01982	1.98018	3.921113	334.5241
4	7.71	55	57.05173	-2.05173	4.209596	75.8641
5	6.68	67	66.25684	0.74316	0.552287	10.8241
6	5.99	72	72.42337	-0.42337	0.179242	68.7241
7	8.19	53	52.76197	0.23803	0.056658	114.7041
8	7.19	60	61.69897	-1.69897	2.886499	13.7641
9	7.4	60	59.8222	0.1778	0.031613	13.7641

Y	dan.sleep	dan.grump	Y-Pred $f(x)$	resid $f(x)$	sq- resid $f(x)$	sq-resid2 $f(x)$
10	6.58	71	67.15054	3.84946	14.81834	53.1441
11	6.49	72	67.95487	4.04513	16.36308	68.7241
12	6.27	65	69.92101	-4.92101	24.21634	1.6641
13	5.95	74	72.78085	1.21915	1.486327	105.8841
14	6.65	67	66.52495	0.47505	0.225673	10.8241
15	6.41	66	68.66983	-2.66983	7.127992	5.2441
16	6.33	69	69.38479	-0.38479	0.148063	27.9841
17	6.3	73	69.6529	3.3471	11.20308	86.3041
18	8.47	52	50.25961	1.74039	3.028957	137.1241
19	7.21	61	61.52023	-0.52023	0.270639	7.3441
20	7.53	53	58.66039	-5.66039	32.04001	114.7041

Figure 11.13: A JASP screenshot showing a residual computation in R.

Naturally, you don't actually need to do all these computations by hand if you want to obtain the R^2 value for your regression model. It turns out that JASP gives you this by default! Take a look at Figure 11.11 again; notice that in the top table labeled 'Model Summary', the value of R^2 is already there!

11.6.2 The relationship between regression and correlation

At this point we can revisit my earlier claim that regression, in this very simple form that I've discussed so far, is basically the same thing as a correlation. Previously, we used the symbol r to denote a Pearson correlation. Might there be some relationship between the value of the correlation coefficient r and the R^2 value from linear regression? Of course there is: the squared correlation r^2 is identical to the R^2 value for a linear regression with only a single predictor. In other words, running a Pearson correlation is more or less equivalent to running a linear regression model that uses only one predictor variable.

11.6.3 The adjusted R^2 value

One final thing to point out before moving on. It's quite common for people to report a slightly different measure of model performance, known as "adjusted R^2". The motivation behind calculating the adjusted R^2 value is the observation that adding more predictors into the model will always cause the R^2 value to increase (or at least not decrease).

The adjusted R^2 value introduces a slight change to the calculation, as follows. For a regression model with K predictors, fit to a data set containing N observations, the adjusted R^2 is:

$$\text{adj. } R^2 = 1 - \left(\frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}} \times \frac{N-1}{N-K-1} \right)$$

This adjustment is an attempt to take the degrees of freedom into account. The big advantage of the adjusted R^2 value is that when you add more predictors to the model, the adjusted R^2 value will only increase if the new variables improve the model performance more than you'd expect by chance. The big disadvantage is that the adjusted R^2 value can't be interpreted in the elegant way that R^2 can. R^2 has a simple interpretation as the proportion of variance in the outcome variable that is explained by the regression model. To my knowledge, no equivalent interpretation exists for adjusted R^2.

An obvious question then is whether you should report R^2 or adjusted R^2. This is probably a matter of personal preference. If you care more about interpretability, then R^2 is better. If you care more about correcting for bias, then adjusted R^2 is probably better. Speaking just for myself, I prefer R^2. My feeling is that it's more important to be able to interpret your measure of model performance. Besides, as we'll see in Section 11.7, if you're worried that the improvement in R^2 that you get by adding a predictor is just due to chance and not because it's a better model, well we've got hypothesis tests for that.

11.7 Hypothesis tests for regression models

So far we've talked about what a regression model is, how the coefficients of a regression model are estimated, and how we quantify the performance of the model (the last of these, incidentally, is basically our measure of effect size). The next thing we need to talk about is hypothesis tests. There are two different (but related) kinds of hypothesis tests that we need to talk about: those in which we test whether the regression model as a whole is performing significantly better than a null model, and those in which we test whether a particular regression coefficient is significantly different from zero.

11.7.1 Testing the model as a whole

Okay, suppose you've estimated your regression model. The overall F-test evaluates the null hypothesis that all population slope coefficients are zero. The alternative hypothesis is that at least one population slope coefficient is not zero.

Formally, our "null model" corresponds to the fairly trivial "regression" model in which we include 0 predictors and only include the intercept term b_0 :

$$H_0 : Y_i = b_0 + \varepsilon_i$$

If our regression model has K predictors, the "alternative model" is described using the usual formula for a multiple regression model:

$$H_1 : Y_i = b_0 + \left(\sum_{k=1}^{K} b_k X_{ik} \right) + \varepsilon_i$$

How can we test these two hypotheses against each other? The trick is to understand that it's possible to divide up the total variance $\mathrm{SS}_{\mathrm{tot}}$ into the sum of the residual variance $\mathrm{SS}_{\mathrm{res}}$ and the regression model variance $\mathrm{SS}_{\mathrm{mod}}$. I'll skip over the technicalities.

$$\mathrm{SS}_{\mathrm{mod}} = \mathrm{SS}_{\mathrm{tot}} - \mathrm{SS}_{\mathrm{res}}$$

And we can convert the sums of squares into mean squares by dividing by the degrees of freedom.

$$\mathrm{MS}_{\mathrm{mod}} = \frac{\mathrm{SS}_{\mathrm{mod}}}{df_{\mathrm{mod}}}$$
$$\mathrm{MS}_{\mathrm{res}} = \frac{\mathrm{SS}_{\mathrm{res}}}{df_{\mathrm{res}}}$$

So, how many degrees of freedom do we have? As you might expect the df associated with the model is closely tied to the number of predictors that we've included. In fact, it turns out that $df_{\mathrm{mod}} = K$. For the residuals the total degrees of freedom is $df_{\mathrm{res}} = N - K - 1$.

Now that we've got our mean square values we can calculate an F-statistic like this

$$F = \frac{\mathrm{MS}_{\mathrm{mod}}}{\mathrm{MS}_{\mathrm{res}}}$$

and the degrees of freedom associated with this are K and $N - K - 1$.
Large F values indicate that the null model performs poorly compared with the model containing predictors, but for now just know that we can interpret large F values as indicating that the null hypothesis is performing poorly in comparison to the alternative hypothesis. In a moment I'll show you how to do the test in JASP the easy way, but first let's have a look at the tests for the individual regression coefficients.

11.7.2 Tests for individual coefficients

The F-test that we've just introduced is useful for checking that the model as a whole is performing better than chance. If your regression model doesn't produce a significant result for the F-test then you probably don't have a very good regression model (or, quite possibly, you don't have very good data). However, while failing this test is a pretty strong indicator that the model has problems, passing the test (i.e., rejecting the null) doesn't imply that the model is good! Why is that, you might be wondering? The answer to that can be found by looking at the coefficients for the multiple regression model we have already looked at in section

11.5 above, where the coefficients we got were:

(Intercept)	dan.sleep	baby.sleep
125.966	-8.950	0.011

I can't help but notice that the estimated regression coefficient for the baby.sleep variable is tiny (0.011), relative to the value that we get for dan.sleep (-8.950). Given that these two variables are absolutely on the same scale (they're both measured in "hours slept"), I find this illuminating. In fact, I'm beginning to suspect that it's really only the amount of sleep that I get that matters in order to predict my grumpiness.

We can re-use a hypothesis test that we discussed earlier, the t-test. The test that we're interested in has a null hypothesis that the true regression coefficient is zero $(b = 0)$, which is to be tested against the alternative hypothesis that it isn't $(b \neq 0)$. That is:

$$
\begin{aligned}
H_0 &: \quad b = 0 \\
H_1 &: \quad b \neq 0
\end{aligned}
$$

How can we test this? Well, if the central limit theorem is kind to us we might be able to guess that the sampling distribution of $\hat{b}$, the estimated regression coefficient, is a normal distribution with mean centred on b. What that would mean is that if the null hypothesis were true, then the sampling distribution of $\hat{b}$ has mean zero and unknown standard deviation. Assuming that we can come up with a good estimate for the standard error of the regression coefficient, $\mathrm{se}(\hat{b})$, then we're in luck. This gives us the familiar structure of a t statistic: an estimate divided by its standard error. So let's define a t-statistic like this

$$
t = \frac{\hat{b}}{\mathrm{se}(\hat{b})}
$$

I'll skip over the reasons why, but our degrees of freedom in this case are $df = N - K - 1$. Irritatingly, the estimate of the standard error of the regression coefficient, $\mathrm{se}(\hat{b})$, is not as easy to calculate as the standard error of the mean that we used for the simpler t-tests. In fact, the formula is somewhat ugly, and not terribly helpful to look at. [6] For our purposes it's sufficient to point out that the standard error of the estimated regression coefficient depends on both the predictor

and outcome variables, and it is somewhat sensitive to violations of the homogeneity of variance assumption (discussed shortly).

In any case, this t-statistic can be interpreted in the same way as the t-statistics. Assuming that you have a two-sided alternative (i.e., you don't really care if $b > 0$ or $b < 0$), then it's the extreme values of t (i.e., a lot less than zero or a lot greater than zero) that suggest that you should reject the null hypothesis.

11.7.3 Running the hypothesis tests in JASP

To compute all of the statistics that we have talked about so far, all you need to do is make sure the relevant options are checked in JASP and then run the regression. Fortunately, these options are usually selected by default. As you can see in Figure 11.14, we get a whole bunch of useful output.

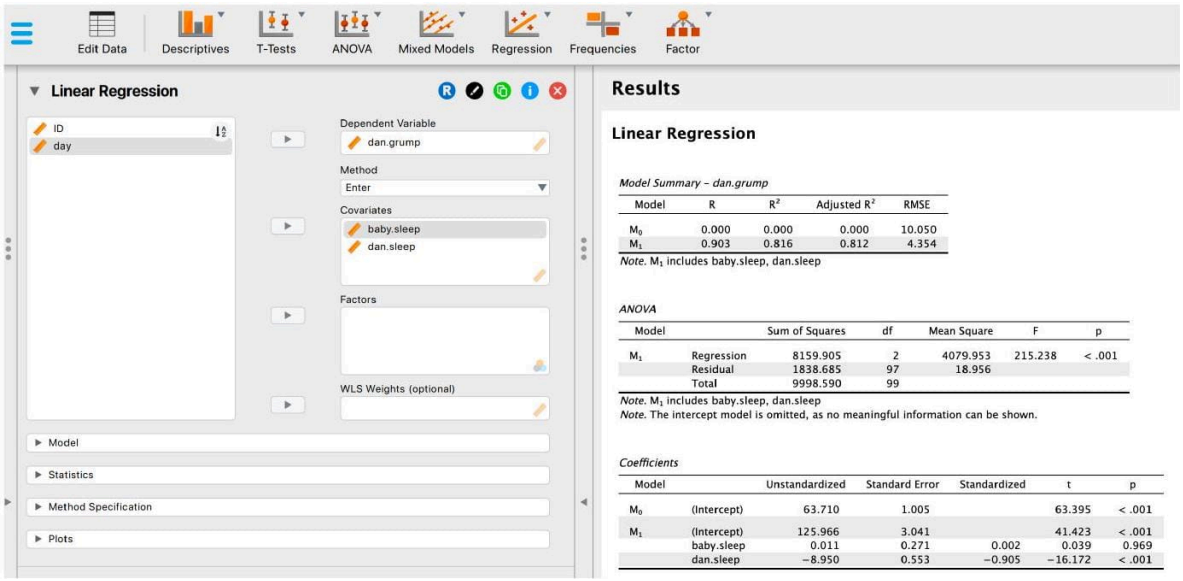


Figure 11.14: A JASP screenshot showing a multiple linear regression analysis, including relevant hypothesis tests.

The 'Coefficients' at the bottom of the JASP analysis results shown in 11.14 provides the coefficients of the regression model. Each row in this table refers to one of the coefficients in the regression model. The first row is the intercept term, and the later ones look at each of the predictors. The columns give you all of the relevant information. The first column (labeled 'Unstandardized') is the actual estimate of b (e.g., 125.966 for the intercept, and -8.950 for the dan.sleep predictor). The second column is the standard error estimate $\hat{\sigma}_b$. The third column provides a 'Standardized' regression coefficient; more about this in Section 11.8. The fourth column gives you the t-statistic, and it's worth noticing that in this table $t = \hat{b}/\text{se}(\hat{b})$ every time. Finally, the last column gives you the actual
p-value for each of these tests. [7]

The only thing that the coefficients table itself doesn't list is the degrees of freedom used in the t test, which is always $N - K - 1$ and is listed in the table in the middle of the output, labelled 'ANOVA'. We can see from this table that the model performs significantly better than you'd expect by chance ($F(2, 97) = 215.238, p < .001$), which isn't all that surprising: the $R^2 = .816$ value indicate that the regression model accounts for 81.6% of the variability in the outcome measure. However, when we look back up at the t-tests for each of the individual coefficients, we can see that the baby.sleep variable seems to have no significant effect. All the work in this model is being done by the dan.sleep variable. Taken together, these results suggest that this regression model is actually the wrong model for the data. You'd probably be better off dropping the baby.sleep predictor entirely. In other words, the simple regression model that we started with is likely the better model.

11.8 Regarding regression coefficients

Before moving on to discuss the assumptions underlying linear regression and what you can do to check if they're being met, there's two more topics I want to briefly discuss, both of which relate to the regression coefficients. The first thing to talk about is calculating confidence intervals for the coefficients. After that, I'll discuss the somewhat murky question of how to determine which predictor is most important.

11.8.1 Confidence intervals for the coefficients

Like any population parameter, the regression coefficients b cannot be estimated with complete precision from a sample of data; that's part of why we need hypothesis tests. Given this, it's quite useful to be able to report confidence intervals that capture our uncertainty about the true value of b. This is especially useful when the research question focuses heavily on an attempt to find out how strongly variable X is related to variable Y, since in those situations the interest is primarily in the regression weight b.

Fortunately, confidence intervals for the regression weights can be constructed in the usual fashion

$$\text{CI}(b) = \hat{b} \pm \Big(t_{\text{crit}} \times \text{se}(\hat{b}) \Big)$$

where $\mathrm{se}(\hat{b})$ is the standard error of the regression coefficient, and t_{crit} is the relevant critical value of the appropriate t distribution. For instance, if it's a 95% confidence interval that we want, then the critical value is the 97.5th quantile of a t distribution with $N - K - 1$ degrees of freedom. In other words, this is basically the same approach to calculating confidence intervals that we've used throughout.

In JASP we can display confidence intervals by selecting 'Confidence intervals' from the 'Statistics' menu in our regression model dialog. The default is 95% CI, but we could easily choose something different, say 99%, if that is what we decided on.

11.8.2 Calculating standardised regression coefficients

One more thing that you might want to do is to calculate "standardised" regression coefficients, often denoted β. The rationale behind standardised coefficients goes like this. In a lot of situations, your variables are on fundamentally different scales. Suppose, for example, my regression model aims to predict people's IQ scores using their educational attainment (number of years of education) and their income as predictors. Obviously, educational attainment and income are not on the same scales. The number of years of schooling might only vary by 10s of years, whereas income can vary by 10,000s of dollars (or more). The units of measurement have a big influence on the regression coefficients. The b coefficients only make sense when interpreted in light of the units, both of the predictor variables and the outcome variable. This makes it very difficult to compare the coefficients of different predictors. Yet there are situations where you really do want to make comparisons between different coefficients. Specifically, you might want some kind of standard measure of which predictors have the strongest relationship to the outcome. This is what standardised coefficients aim to do.

The basic idea is quite simple; the standardised coefficients are the coefficients that you would have obtained if you'd converted all the variables to z-scores before running the regression. [8] The idea here is that, by converting all the predictors to z-scores, they all go into the regression on the same scale, thereby removing the problem of having variables on different scales. Regardless of what the original variables were, a β value of 1 means that an increase in the predictor of 1 standard deviation will produce a corresponding 1 standard deviation increase in the outcome variable. Therefore, if variable A has a larger absolute value of β than variable B, it is deemed to have a stronger relationship with the outcome. Or at least that's the idea. It's worth being a little cautious here, since this does rely very heavily on the assumption that "a 1 standard deviation change" is fundamentally the same kind of thing for all variables. It's not always obvious that this is true.

Leaving aside the interpretation issues, let's look at how it's calculated. What you could do is standardise all the variables yourself and then run a regression, but there's a much simpler way to do it. As it turns out, the β coefficient for a predictor X and outcome Y has a very simple formula, namely

$$\beta_X = b_X \times \frac{\sigma_X}{\sigma_Y}$$

where σ_X is the standard deviation of the predictor, and σ_Y is the standard deviation of the outcome variable Y. This makes matters a lot simpler.

To make things even simpler, JASP computes the β coefficients by default, as you can see the third column of the 'Coefficients' table in Figure 11.14. This clearly shows that the dan.sleep variable has a much stronger effect than the baby.sleep variable. However, this is a perfect example of a situation where it would probably make sense to use the original coefficients b rather than the standardised coefficients β. After all, my sleep and the baby's sleep are already on the same scale: number of hours slept. Why complicate matters by converting these to z-scores?

11.9 Assumptions of regression

The linear regression model that I've been discussing relies on several assumptions. In Section 11.10 we'll talk a lot more about how to check that these assumptions are being met, but first let's have a look at each of them.

- Normality. Like many of the models in statistics, basic simple or multiple linear regression relies on an assumption of normality. Specifically, it assumes that the residuals are normally distributed. It's actually okay if the predictors X and the outcome Y are non-normal, so long as the residuals ε are normal. See Section 11.10.3.
- Linearity. A pretty fundamental assumption of the linear regression model is that the relationship between X and Y actually is linear! Regardless of whether it's a simple regression or a multiple regression, we assume that the relationships involved are linear.
- Homogeneity of variance. Strictly speaking, the regression model assumes that each residual ε_i is generated from a normal

distribution with mean 0, and (more importantly for the current purposes) with a standard deviation σ that is the same for every single residual. In practice, it's impossible to test the assumption that every residual is identically distributed. Instead, what we care about is that the standard deviation of the residual is the same for all values of $\hat{Y}$, and (if we're being especially paranoid) all values of every predictor X in the model.
- No perfect multicollinearity. In multiple regression, no predictor can be an exact linear combination of the others. Predictors can be correlated, but very strong correlations can make individual coefficient estimates unstable and difficult to interpret.
- Residuals are independent of each other. This is really just a "catch all" assumption, to the effect that "there's nothing else funny going on in the residuals". If there is something weird (e.g., the residuals all depend heavily on some other unmeasured variable) going on, it might screw things up.
- No "bad" outliers. Again, not actually a technical assumption of the model (or rather, it's sort of implied by all the others), but there is an implicit assumption that your regression model isn't being too strongly influenced by one or two anomalous data points because this raises questions about the adequacy of the model and the trustworthiness of the data in some cases. See Section 11.10.2.

11.10 Model checking

The main focus of this section is regression diagnostics, a term that refers to the art of checking that the assumptions of your regression model have been met, figuring out how to fix the model if the assumptions are violated, and generally to check that nothing "funny" is going on. I refer to this as the "art" of model checking with good reason. It's not easy, and while there are a lot of fairly standardised tools that you can use to diagnose and maybe even cure the problems that ail your model (if there are any, that is!), you really do need to exercise a certain amount of judgement when doing this. It's easy to get lost in all the details of checking this thing or that thing, and it's quite exhausting to try to remember what all the different things are. This has the very nasty side effect that a lot of people get frustrated when trying to learn all the tools, so instead they decide not to do any model checking. This is a bit of a worry!

In this section I describe several different things you can do to check that your regression model is doing what it's supposed to. It doesn't cover the full space of things you could do, but it's still much more detailed than what I see a lot of people doing in practice, and even I don't usually cover all of this in my intro stats class either. However, I do think it's important that you get a sense of what tools are at your disposal, so I'll try to introduce a bunch of them here. Finally, I should note that this section draws quite heavily from the Fox and Weisberg (2011) text, the book associated with the car package that is used to conduct regression analysis in R. The car package is notable for providing some excellent tools for regression diagnostics, and the book itself talks about them in an admirably clear fashion. I don't want to sound too gushy about it, but I do think that Fox et al. (2011) is well worth reading, even if some of the advanced diagnostic techniques are only available in R and not JASP.

11.10.1 Three kinds of residuals

The majority of regression diagnostics revolve around looking at the residuals, and by now you've probably formed a sufficiently pessimistic theory of statistics to be able to guess that, precisely because of the fact that we care a lot about the residuals, there are several different kinds of residual that we might consider. In particular, the following three kinds of residuals are referred to in this section: "ordinary residuals", "standardised residuals", and "Studentised residuals". There is a fourth kind that you'll see referred to in some of the Figures, and that's the "Pearson residual". However, for the models that we're talking about in this chapter the Pearson residual is identical to the ordinary residual.

The first and simplest kind of residuals that we care about are ordinary residuals. These are the actual raw residuals that I've been talking about throughout this chapter so far. The ordinary residual is just the difference between the fitted value $\hat{Y}_i$ and the observed value Y_i. I've been using the notation ε_i to refer to the i-th ordinary residual, and darn it, I'm going to stick to it. With this in mind, we have the very simple equation

$$\varepsilon_i = Y_i - \hat{Y}_i$$

This is of course what we saw earlier, and unless I specifically refer to some other kind of residual, this is the one I'm talking about. So there's nothing new here. I just wanted to repeat myself. One drawback to using ordinary residuals is that they're always on a different scale, depending on what the outcome variable is and how good the regression model is. That is, unless you've decided to run a regression model without an intercept term, the ordinary residuals will have mean 0 but the variance is different for every regression. In a lot of contexts, especially where you're only interested in the pattern of the residuals and not their actual values, it's convenient to estimate the standardised residuals, which are normalised in such a way as to have standard deviation 1.

The way we calculate these is to divide the ordinary residual by an estimate of the (population) standard deviation of these residuals. For technical reasons, mumble mumble, the formula for this is

$$\varepsilon_i' = \frac{\varepsilon_i}{\hat{\sigma}\sqrt{1-h_i}}$$

where $\hat{\sigma}$ in this context is the estimated population standard deviation of the ordinary residuals, and h_i is the "hat value" of the ith observation. I haven't explained hat values to you yet (but have no fear, [a] it's coming shortly), so this won't make a lot of sense. For now, it's enough to interpret the standardised residuals as if we'd converted the ordinary residuals to z-scores. In fact, that is more or less the truth, it's just that we're being a bit fancier.

The third kind of residuals are Studentised residuals (also called "jackknifed residuals") and they're even fancier than standardised residuals. Again, the idea is to take the ordinary residual and divide it by some quantity in order to estimate some standardised notion of the residual.

The formula for doing the calculations this time is subtly different

$$\varepsilon_i^* = \frac{\varepsilon_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}}$$

Notice that our estimate of the standard deviation here is written $\hat{\sigma}_{(-i)}$. What this corresponds to is the estimate of the residual standard deviation that you would have obtained if you just deleted the ith observation from the data set. This sounds like the sort of thing that would be a nightmare to calculate, since it seems to be saying that you have to run N new regression models (even a modern computer might grumble a bit at that, especially if you've got a large data set). Fortunately, some terribly clever person has shown that this standard deviation estimate is actually given by the following equation:

$$\hat{\sigma}_{(-i)} = \hat{\sigma}\sqrt{\frac{N-K-1-\varepsilon_i'^2}{N-K-2}}$$

Isn't that a pip?
Before moving on, I should point out that you don't often need to obtain these residuals yourself, even though they are at the heart of almost all regression diagnostics. Most of the time the various options that provide the diagnostics, or assumption checks, will take care of these calculations for you. Even so, it's always nice to know how to actually get hold of these things yourself in case you ever need to do something non-standard.

11.10.2 Three kinds of anomalous data

One danger that you can run into with linear regression models is that your analysis might be disproportionately sensitive to a smallish number of "unusual" or "anomalous" observations. We have already encountered outliers in the context of boxplots, but regression requires a more precise distinction between different kinds of unusual observations. In the context of linear regression, there are three conceptually distinct ways in which an observation might be called "anomalous". All three are interesting, but they have rather different implications for your analysis.

The first kind of unusual observation is an outlier. The definition of an outlier (in this context) is an observation that is very different from what the regression model predicts. An example is shown in Figure 11.15. In practice, we operationalise this concept by saying that an outlier is an observation that has a very large Studentised residual, ε_i^*. Outliers are interesting: a big outlier might correspond to junk data, e.g., the variables might have been recorded incorrectly in the data set, or some other defect may be detectable. Note that you shouldn't throw an observation away just because it's an outlier. But the fact that it's an outlier is often a cue to look more closely at that case and try to find out why it's so different.

The second way in which an observation can be unusual is if it has high leverage, which happens when the observation is very different from all the other observations. This doesn't necessarily have to correspond to a large residual. If the observation happens to be unusual on all variables in precisely the same way, it can actually lie very close to the regression line. An example of this is shown in Figure 11.16. The leverage of an observation is operationalised in terms of its hat value, usually written h_i. The formula for the hat value is rather complicated [9] but its interpretation is not: h_i is a measure of the extent to which the i-th observation is "in control" of where the regression line ends up going.

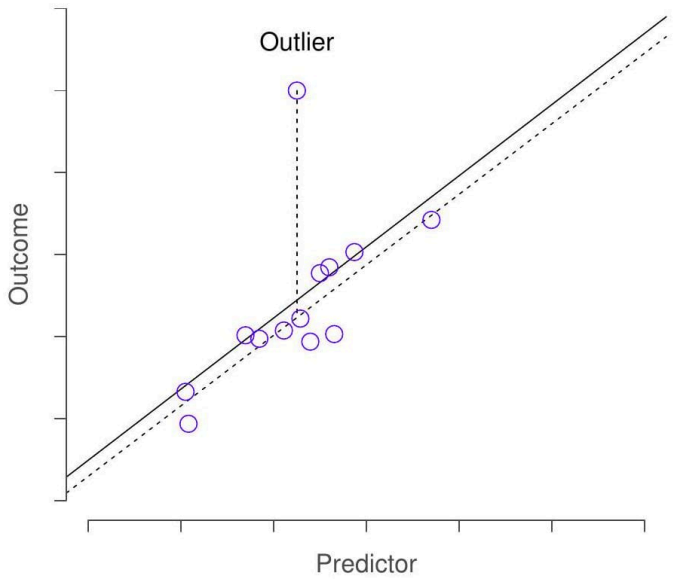


Figure 11.15: An illustration of outliers. The dotted lines plot the regression line that would have been estimated without the anomalous observation included, and the corresponding residual (i.e., the Studentised residual). The solid line shows the regression line with the anomalous observation included. The outlier has an unusual value on the outcome (y axis location) but not the predictor (x axis location), and lies a long way from the regression line.

In general, if an observation lies far away from the other ones in terms of the predictor variables, it will have a large hat value (as a rough guide, high leverage is when the hat value is more than 2-3 times the average; and note that the sum of the hat values is constrained to be equal to $K + 1$). High leverage points are also worth looking at in more detail, but they're much less likely to be a cause for concern unless they are also outliers.

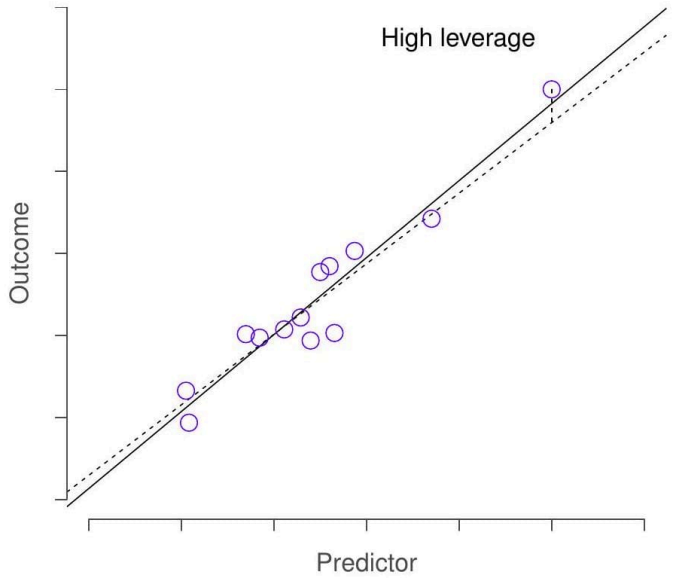


Figure 11.16: An illustration of high leverage points. The anomalous observation in this case is unusual both in terms of the predictor (x axis) and the outcome (y axis), but this unusualness is highly consistent with the pattern of correlations that exists among the other observations. The observation falls very close to the regression line and does not distort it.

This brings us to our third measure of unusualness, the influence of an observation. A high influence observation is an outlier that has high leverage. That is, it is an observation that is very different to all the other ones in some respect, and also lies a long way from the regression line. This is illustrated in Figure 11.17. Notice the contrast to the previous two figures. An observation's influence depends on the combination of its residual size and leverage. That's why we call these points high influence, and it's why they're the biggest worry. We operationalise influence in terms of a measure known as Cook's distance.

$$D_i = \frac{\varepsilon_i^{*2}}{K+1} \times \frac{h_i}{1-h_i}$$

Notice that this is a multiplication of something that measures the outlier-ness of the observation (the bit on the left), and something that measures the leverage of the observation (the bit on the right).

Cook's distance combines residual size and leverage to measure how strongly an observation affects the fitted model. As a rough guide, Cook's distance greater than 1 is often considered large (that's what I typically use as a quick and dirty rule).

In JASP, information about Cook's distance can be calculated by clicking on 'Casewise diagnostics' under the 'Statistics' menu. There are two ways to visualize these data. First, you can select 'All' to see Cook's distance for each case (i.e., row of data); see Figure 11.18. Alternatively, you can opt to display only those cases for which Cook's distance is greater than some threshold; the default in JASP is 1. In either case, you can see that we have no data cases that are beyond this threshold.

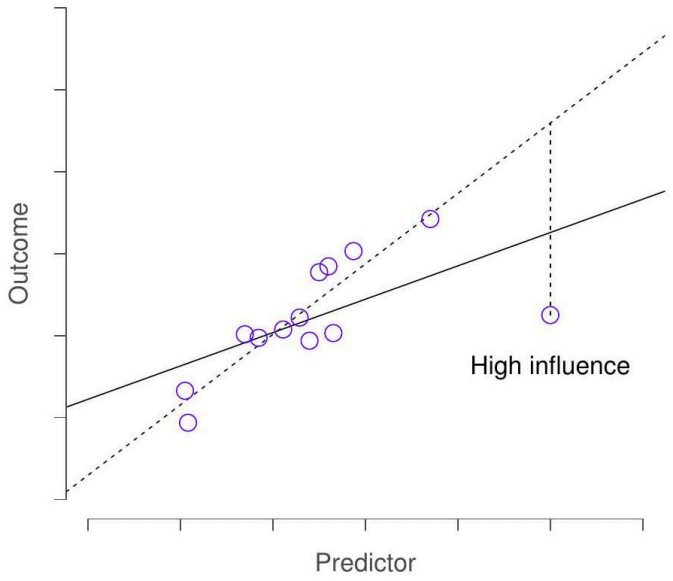


Figure 11.17: An illustration of high influence points. In this case, the anomalous observation is highly unusual on the predictor variable (x axis), and falls a long way from the regression line. As a consequence, the regression line is highly distorted, even though (in this case) the anomalous observation is entirely typical in terms of the outcome variable (y axis).

Influential Cases

Case Number	Std. Residual	dan.grump	Predicted Value	Residual	Cook's Distance
1	-0.497	56.000	58.140	-2.140	0.002
2	1.104	60.000	55.292	4.708	0.017
3	0.464	82.000	80.045	1.955	0.005
4	-0.477	55.000	57.060	-2.060	0.001
5	0.168	67.000	66.281	0.719	2.632×10^{-4}
6	-0.095	72.000	72.407	-0.407	9.698×10^{-5}
7	0.053	53.000	52.773	0.227	2.620×10^{-5}
8	-0.393	60.000	61.700	-1.700	5.458×10^{-4}
9	0.047	60.000	59.797	0.203	2.876×10^{-5}
10	0.890	71.000	67.148	3.852	0.003
11	0.959	72.000	68.001	3.999	0.027
12	-1.139	65.000	69.912	-4.912	0.008

Figure 11.18: JASP output showing Cook's distance for each case/row of data

An obvious question to ask next is, if you do have large values of Cook's distance what should you do? As always, there's no hard and fast rule. Probably the first thing to do is to try running the regression with the outlier with the greatest Cook's distance [10] excluded and see what happens to the model performance and to the regression coefficients. If they really are substantially different, it's time to start digging into your data set and your notes that you no doubt were scribbling as your ran your study. Try to figure out why the point is so different. If you start to become convinced that this one data point is badly distorting your results then you might consider excluding it, but that's less than ideal unless you have a solid explanation for why this particular case is qualitatively different from the others and therefore deserves to be handled separately.

11.10.3 Checking the normality of the residuals

Like many of the statistical tools we've discussed in this book, regression models rely on a normality assumption. In this case, we assume that the residuals are normally distributed. The first thing we can do is draw a QQ-plot in JASP via the 'Plots' - 'Q-Q plot standardized residuals' option. The output is shown in Figure 11.19, showing the standardised residuals plotted as a function of their theoretical quantiles according to the regression model.

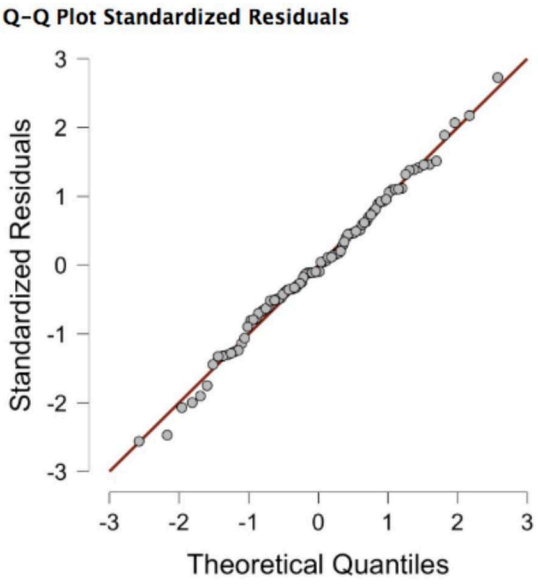


Figure 11.19: Plot of the theoretical quantiles according to the model, against the quantiles of the standardised residuals, produced in JASP.

Another thing we should check is the relationship between the residuals and the fitted values or predictors. JASP provides these plots through its Residuals Plots options; see Figure 11.20. We look for residuals scattered around zero, without a curved pattern or a clear change in their vertical spread. These plots help assess linearity and homogeneity of variance.

The residual-versus-outcome plot should not be judged by the same rule: a relationship between residuals and the observed outcome is expected and does not by itself indicate a problem.

If we were worried, then in a lot of cases the solution to this problem (and many others) is to transform one or more of the variables. Transformations are beyond the scope of this text.

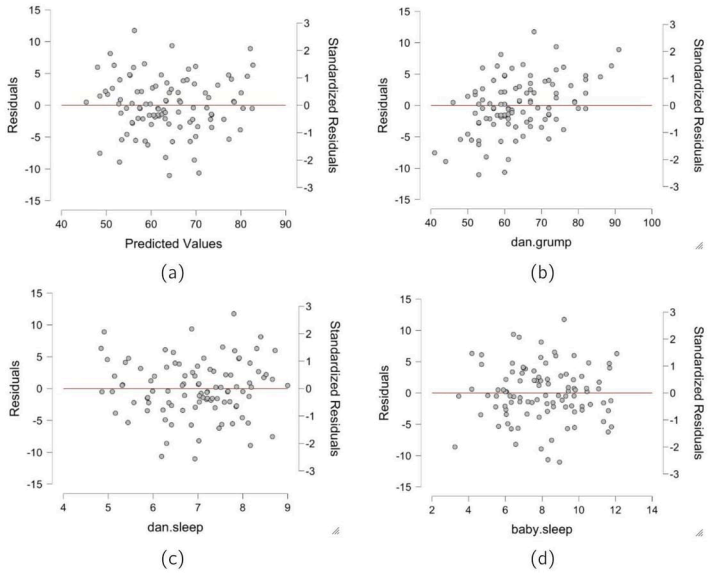


Figure 11.20: Residuals plots produced in JASP

11.11 Model selection

One fairly major problem that remains is the problem of "model selection". That is, if we have a data set that contains several variables, which ones should we include as predictors, and which ones should we not include? In other words, we have a problem of variable selection. In general, model selection is a complex business but it's made somewhat simpler if we restrict ourselves to the problem of choosing a subset of the variables that ought to be included in the model.

Nevertheless, I'm not going to try covering even this reduced topic in a lot of detail. Instead, I'll talk about two broad principles that you need to think about, and then discuss one concrete tool to help you select a subset of variables to include in your model. First, the two principles:

- It's nice to have an actual substantive basis for your choices. That is, in a lot of situations you the researcher have good reasons to pick out a smallish number of possible regression models that are of theoretical interest. These models will have a sensible interpretation in the context of your field. Never discount the importance of this. Statistics serves the scientific process, not the other way around.
- To the extent that your choices rely on statistical inference, there is a trade off between simplicity and goodness of fit. As you add more predictors to the model you make it more complex. Each predictor adds a new free parameter (i.e., a new regression coefficient), and each new parameter increases the model's capacity to "absorb" random variations. So the goodness of fit (e.g., R^2)

continues to rise, sometimes trivially or by chance, as you add more predictors no matter what. If you want your model to be able to generalise well to new observations you need to avoid throwing in too many variables.

This latter principle is often referred to as Ockham's razor and is often summarised in terms of the following pithy saying: do not multiply entities beyond necessity. In this context, it means don't chuck in a bunch of largely irrelevant predictors just to boost your R^2. Hmm. Yeah, the original was better.

In any case, what we need is an actual mathematical criterion that will implement the qualitative principle behind Ockham's razor in the context of selecting a regression model. As it turns out there are several possibilities. The one that I'll talk about is the Akaike information criterion (AIC; Akaike 1974); currently, it is not part of JASP's standard output, but it is quite easy to compute with the model fit data that JASP does produce.

In the context of a linear regression model, the AIC for a model that has n observations and K predictor variables (not including the intercept) can be computed [a] as

$$AIC = n \ln \left(SS_{\text{res}} \right) + 2K$$

[a] Strictly speaking, this formula is not completely correct. Akaike's original definition was in terms of something called a maximum likelihood estimate for the model, and as such, there are some other terms that appear in the computation. However, many of them don't depend on the model, and given that the purpose of the AIC is to compare models, these terms will be present in all models and will mathematically "wash out". Thus, I am presenting a 'bare bones' version of the formula that is sufficient for our purposes.

Here's the basic principle behind using AIC for model comparison: the smaller the AIC value, the better the model performance. If we ignore the low level details it's fairly obvious what the AIC does. On the left we have a term that decreases as the model predictions get better; on the right we have a term that increases as the model complexity increases. The best model is the one that both fits the data well (small SS_{res} , left hand side) and uses as few predictors as possible (small K, right hand side). In short, this is a simple mathematical implementation of Ockham's razor.

Let's demonstrate how AIC can be used to compare the two regression models we have computed in this chapter. Consider first the regression model with only one predictor: dan.sleep (see Figure 11.11). In this model, we have $n = 100$ observations, $K = 1$ predictor, and $SS_{\text{res}} = 1838.714$. Thus,

$$\begin{aligned} AIC_1 &= n \ln (SS_{\text{res}}) + 2K \\ &= 100 \ln(1838.714) + 2(1) \\ &= 753.68 \end{aligned}$$

Now consider the second model using two predictors: dan.sleep and baby.sleep (see Figure 11.14). In this model, we have $n = 100$ observations, $K = 2$ predictors, and $SS_{\text{res}} = 1838.685$. This gives us

$$\begin{aligned} AIC_2 &= n \ln (SS_{\text{res}}) + 2K \\ &= 100 \ln(1838.685) + 2(2) \\ &= 755.68. \end{aligned}$$

Since $AIC_1 < AIC_2$, this tells us that Model 1 is the better fit, which confirms our intuitions. Adding baby.sleep doesn't add much to the model fit, but it increases model complexity. AIC balances these two requirements; the penalty for adding an additional parameter is not outweighed by the meager improvement in model fit.

11.12 Summary

- -Want to know how strong the relationship is between two variables? Calculate a correlation (Section 11.1).
- -Drawing scatterplots (Section 11.2).
- -Basic ideas in linear regression and how regression models are estimated (Sections 11.3 and 11.4).
- -Multiple linear regression (Section 11.5).
- -Measuring the overall performance of a regression model using R^2 (Section 11.6).
- -Hypothesis tests for regression models (Section 11.7)
- -Calculating confidence intervals for regression coefficients and standardised coefficients (Section 11.8).
- -The assumptions of regression (Section 11.9) and how to check them (Section 11.10).
- -Selecting a regression model (Section 11.11).

---

1. [1] Actually, even that table is more than I'd bother with. In practice most people pick one measure of central tendency, and one measure of variability only.
2. [2] Also sometimes written as $y = mx + b$ where m is the slope coefficient and b is the intercept (constant) coefficient.
3. [3] The ε symbol is the Greek letter epsilon. It's traditional to use ε_i or e_i to denote a residual.
4. [4] Or at least, I'm assuming that it doesn't help most people. But on the off chance that someone reading this is a proper kung fu master of linear algebra (and to be fair, I always have a few of these people in my intro stats class), it will help you to know that the solution to the estimation problem turns out to be $\hat{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$, where $\hat{b}$ is a vector containing the estimated regression coefficients, X is the "design matrix" that contains the predictor variables (plus an additional column containing all ones; strictly $\mathbf{X}$ is a matrix of the regressors, but I haven't discussed the distinction yet), and y is a vector containing the outcome variable. For everyone else, this isn't exactly helpful and can be downright scary. However, since quite a few things in linear regression can be written in linear algebra terms, you'll see a bunch of footnotes like this one in this chapter. If you can follow the maths in them, great. If not, ignore it.
5. [5] And by "sometimes" I mean "almost never". In practice everyone just calls it " R-squared".
6. [6] For advanced readers only. The vector of residuals is $\varepsilon = y - \mathbf{X}\hat{b}$. For K predictors plus the intercept, the estimated residual

variance is $\hat{\sigma}^2 = \epsilon'\epsilon/(N-K-1)$. The estimated covariance matrix of the coefficients is $\hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}$, the main diagonal of which is $\text{se}(\hat{b})$, our estimated standard errors.

7. [7] Note that, although JASP has done multiple tests here, it hasn't done any sort of correction for multiple comparisons. These are standard one-sample t-tests with a two-sided alternative. If you want to make corrections for multiple tests, you need to do that yourself.
8. [8] Strictly, you standardise all the regressors. That is, every "thing" that has a regression coefficient associated with it in the model. For the regression models that I've talked about so far, each predictor variable maps onto exactly one regressor, and vice versa. However, that's not actually true in general. But, for now we don't need to care too much about this distinction.
9. [a] Or have no hope, as the case may be.
10. [9] Again, for the linear algebra fanatics: the "hat matrix" is defined to be that matrix $\mathbf{H}$ that converts the vector of observed values y into a vector of fitted values $\hat{y}$, such that $\hat{y} = \mathbf{H}y$. The name comes from the fact that this is the matrix that "puts a hat on y ". The hat value of the i-th observation is the i-th diagonal element of this matrix (so technically I should be writing it as h_{ii} rather than h_i). Oh, and in case you care, here's how it's calculated: $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Pretty, isn't it?
11. [10] although currently there isn't a very easy way to do this in JASP, so a more powerful regression program such as the car package in R would be better for this more advanced analysis

This page titled 11.1: Correlation and linear regression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 11: Correlation and linear regression by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

11.2: Pearson and Spearman Correlations

Pearson's *r* Correlation

How do I do that?

1. In the 'Regression' menu select **Correlation** from the 'Classical' list.

image

2. A new window will open, like so:

image

3. Move the variables for your analysis into the **Variables** box. You can move more than two variables into this box to obtain a matrix of pairwise correlations. Do not worry about the 'Partial out' box at this stage.

image

4. There are many options in the expandable menus for this type of analysis which can initially be overwhelming. However, it is unlikely that you will need all of them. Here are some of the options which you are likely to use:

Sample Correlation Coefficient. Pearson's *r* iis selected by default and measures linear association. Spearman's rho measures monotonic association using ranks and can be appropriate for ordinal data. A departure from normality alone does not automatically require switching to Spearman's *rho*.

Alt. Hypothesis. The default is 'Correlated' specifies a two-sided alternative hypothesis: the population correlation is not zero. The null hypothesis is that the population correlation is zero. The other two options specify directional alternatives.

Additional Options. These provide you with additional information about your results. Unless instructed otherwise, it's usually a good idea to generate the following additional statistics:

- **Display pairwise** – this is useful for 'simple' (i.e., with only two variable) correlations as the output is a single line, rather than a correlation matrix.
- **Report significance** – because you want the *p* value!
- **Flag significant correlations** – this is useful, particularly when you have a number of correlations in a matrix.

Plots. Select 'Scatter plots' to visualise the data.

Spearman's *rho*

How do I do that?

1. The only change you make from the previous instruction is to select **Spearman's *rho***.

image image

2. All other menu selections are the same.

This page titled 11.2: Pearson and Spearman Correlations is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rebecca Denniss & Stanislava Naneva via source content that was edited to the style and standards of the LibreTexts platform.

11.3: Calculating and Reporting Pearson's r

Example of How to Test a Hypothesis Using Correlation

Let's test the hypothesis and Data Set 12.1 below. Suppose a researcher collected data from 10 college students to test the hypothesis that sleep would positively relate to quiz scores. Assume that Data Set 12.1 includes data from the aforementioned sample. Let's follow the steps in hypothesis testing using these data.

Data Set 12.1. Hours of Sleep and Quiz Scores (n = 10)

Participant Number	Sleep Hours	Quiz Score
1	7	92
2	8	88
3	9	96
4	6	70
5	6	79
6	4	64
7	5	75
8	10	98
9	3	53
10	7	85

Steps in Hypothesis Testing

In order to test a hypothesis, we must follow these steps:

1. **State the hypothesis.**

A summary of the research hypothesis and corresponding null hypothesis in sentence and symbol format are shown below. However, researchers often only state the research hypothesis using a format like this: *It is hypothesized that hours of sleep will positively relate to quiz scores.* The format shown in the table below could also be used. Because the hypothesis is directional a one-tailed is needed.

Directional Hypothesis for a Bivariate Correlation

Research hypothesis	Hours of sleep will be positively related to quiz scores.	H_a: $\rho > 0$
Null hypothesis	Hours of sleep will not be positively related to quiz scores.	H_0: $\rho \leq 0$

2. **Choose the inferential test (formula) that best fits the hypothesis.**

The relationship between two quantitative variables is being tested so the appropriate test is a bivariate correlation.

3. **Determine the critical value.**

In order to determine the critical value for a bivariate correlation, three things must be identified:

1. the alpha level,
2. the degrees of freedom (df), and
3. whether the hypothesis is directional (requiring a one-tailed test) or non-directional (requiring a two tailed test).

The alpha level is often set at .05 unless there is reason to adjust it such as when multiple hypotheses are being tested in one study or when a Type I Error could be particularly problematic. The default alpha level can be used for this example because only one hypothesis is being tested and there is no clear indication that a Type I Error would be especially problematic. Thus, alpha can be set to 5%, which can be summarized as α = .05.

The df must also be calculated. The df is calculated as the sample size minus the number of variables being tested. In bivariate correlation, there are always 2 variables being tested so the formula is $df = n–2$. The sample size in Data Set 12.1 is 10. Thus, the df for Data Set 12.1 is as follows:

$$df = n - 2$$
$$df = 10 - 2$$
$$df = 8$$

The hypothesis is directional because it specified that the expected correlation would be positive. Thus, this hypothesis requires a one-tailed test of significance.

The alpha level, df, and determination of whether the hypothesis requires a one-tailed or two-tailed test of significance are used to locate the critical value from the test. Below is an excerpt of the section of the *r*-tables that fits the current hypothesis and data. Under the conditions of an alpha level of .05, df = 8, and using a one-tailed test, the critical value is.549.

Critical Values Table

Degrees of Freedom	**one-tailed test**		
	alpha level:	α **= 0.05**	α **= 0.01**
	8	.549	.716

The critical value represents the value which must be exceeded in order to declare a result significant. The obtained value (which is called r in correlation) is the magnitude of evidence present. Because the correlation is directional and states the correlation will be positive, the obtained value must both be in the hypothesized direction (i.e. indicate a positive relationship) and exceed the critical value in magnitude. Thus, in order for the result to significantly support the hypothesis is needs to be positive and exceed the critical value of .549.

4. **Calculate the test statistic.**

A test statistic can also be referred to as an obtained value. The formula needed to find the test statistic r for this scenario is as follows:

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2}\sqrt{\Sigma(Y - \bar{Y})^2}}$$

Section A: Preparation

Start each inferential formula by identifying and solving for the pieces that must go into the formula. For bivariate correlation, this preparatory work is as follows:

1. Find $\bar{X}$ (the mean for the X-variable scores)

This value is found using Data Set 12.1 and is summarized as $\bar{X}$ = 6.50

2. Find $\bar{Y}$ (the mean for the Y-variable scores)

This value is found using Data Set 12.1 and is summarized as $\bar{Y}$ = 80.00

Now that the pieces needed for the formula have been found, we can move to Section B.

Section B: Solving

The values from the preparatory steps can now be plugged into the correlation formula and used to find the r-value. Much of the work involves finding deviations from the mean. Therefore, a summary table will be used to show deviations for the two variables in this section.

1. Find $\sqrt{\Sigma(X-\bar{X})^2}$
   a. Find $((X-\bar{X}))$ by subtracting the mean of X from each x-value
   b. Find $\Sigma(X-\bar{X})^2$ by squaring each deviation and then summing those values, yielding SS_X
   c. Square root SS_X.
2. Find $\sqrt{\Sigma(Y-\bar{Y})^2}$
   a. Find $(Y-\bar{Y})$ by subtracting the mean of Y from each y-value
   b. Find $\Sigma(Y-\bar{Y})^2$ by squaring each deviation and then summing those values, yielding SS_Y
   c. Square root SS_Y.
3. Find $\Sigma(X-\bar{X})(Y-\bar{Y})$
   a. Find $(X-\bar{X})(Y-\bar{Y})$ by multiplying each deviation from X by its deviation from Y to get the product of the deviations for each case
   b. Find $\Sigma(X-\bar{X})(Y-\bar{Y})$ by summing the product of deviations for each case (i.e. sum the results from step 3a). This is the numerator for the formula.
4. Find the denominator by multiplying the square root of SS_X (which is the result of step 1c) by the square root of SS_Y (which is the result of step 2c)
5. Divide the numerator (which is the result of step 3b) by the denominator (which is the result of step 4).

Correlation Computations for Data Set 12.1.

Steps 1 Through 3							
Deviation Steps		**1a**	**1b**		**2a**	**2b**	**3a**
	Sleep Hours (X)	Deviation $(X-\bar{X})$	Dev. Squared $(X-\bar{X})^2$	Quiz Scores (Y)	Deviation $(Y-\bar{Y})$	Dev. Squared $(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
	7	0.50	0.25	92	12	144	6.00
	8	1.50	2.25	88	8	64	12.00
	9	2.50	6.25	96	16	256	40.00
	6	-0.50	0.25	70	-10	100	5.00
	6	-0.50	0.25	79	-1	1	0.50

	4	-2.50	6.25	64	-16	256	40.00
	5	-1.50	2.25	75	-5	25	7.50
	10	3.50	12.25	98	18	324	63.00
	3	-3.50	12.25	53	-27	729	94.50
	7	0.50	0.25	85	5	25	2.50
Summation Steps	$\bar{X}$ = 6.50		**Step 1c** $\Sigma(X-\bar{X})^2$ = 42.50	$\bar{Y}$ = 80.00		**Step 2c** $\Sigma(Y-\bar{Y})^2$ = 1,924.00	**Step 3b** $\Sigma(X-\bar{X})(Y-\bar{Y})$ = 271.00
Step 4							
Find the denominator	$\sqrt{\Sigma(X-\bar{X})^2}\sqrt{\Sigma(Y-\bar{Y})^2} = \sqrt{42.50}\sqrt{1924} = 285.9545\ldots$						
Step 5							
Put the pieces together to find r and then round to the hundredths place	$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{\sum(Y-\overline{Y})^2}}$ $r = \frac{271.00}{285.9545\ldots}$ $r = .9477\ldots$ $r \approx .95$						

The obtained value for this test is .95 when rounded to the hundredths place. This is a very strong, positive correlation.

5. **Apply a decision rule and determine whether the result is significant.**

Assess whether the obtained value for r exceeds the critical value in the proper direction as follows:

a. Check the direction. The hypothesis stated the relationship would be positive. The result was positive. Thus, the direction hypothesized is supported.
b. Check the magnitude. The critical value is .549. The obtained r-value is .95. The obtained r-value exceeds (i.e. is greater in magnitude than) the critical value. Thus, the magnitude is sufficient to support the hypothesis.
c. Decide whether the evidence is sufficient to support the hypothesis

The criteria has been met for both direction and magnitude. Thus, the result significantly supports the hypothesis.

Note

If the hypothesis had not been directional, the only comparison needed before concluding that the hypothesis was supported would be the magnitude of the obtained r-value compared to the critical value.

6. **Calculate the effect size and/or other relevant secondary analyses.**

Pearson's r is itself an effect-size measure. We can also calculate the coefficient of determination, r^2, to describe the proportion of variance accounted for by the corresponding simple linear regression model. Neither measure requires a statistically significant result.

The **coefficient of determination** is the percent of variation in the Y-variable that is accounted for by variance in the X-variable. It can also be described as a calculation of how well a model using one variable (X) can be used to estimate the other (Y). These refer to two ways of interpreting and describing the same thing with the former more applicable to correlation and the latter more applicable to regression. As the focus of this chapter is correlation, we will use the interpretation language that is most applicable to correlation.

The symbol and the formula for the coefficient of determination are the same and are written as follows:

$$r^2$$

To calculate this, the obtained r-value is squared and often reported as a percent. The greatest r^2 can be is 1.00 (or 100.00% when presented as a percent). This would occur if there was a perfect correlation and could be interpreted and reported as follows:

Approximately 100.00% of the variance in Y is accounted for by variance in X.

The lowest r^2 can be is 0.00 (or 0.00% when presented as a percent). This would occur if there was no correlation (r = 0.00) and could be interpreted and reported as follows:

Approximately 0.00% of the variance in Y is accounted for by variance in X.

Our result with Data Set 12.1 was r = .9477.... Squaring this value gives the coefficient of determination:

$$r^2 = (0.9477\ldots)^2$$
$$r^2 = 0.8981\ldots$$

This result is quite large and can be reported as a percent and interpreted as follows for Data Set 12.1:

Approximately 89.81% of the variance in quiz scores was accounted for by variance in hours of sleep.

7. **Report the results in American Psychological Association (APA) format.**

Results for inferential tests are often best summarized using a paragraph that states the following:

a. the hypothesis and specific inferential test used,
b. the main results of the test and whether they were significant,
c. any additional results that clarify or add details about the results,
d. whether the results support or refute the hypothesis.

Following this, the results for our hypothesis with Data Set 12.1 can be written as shown in the summary example below.

APA Formatted Summary Example

A Pearson correlation was used to test the hypothesis that hours of sleep would be positively related to quiz scores. Consistent with the hypothesis, students who reported more sleep tended to have higher quiz scores, r(8) = .95, p < .001, one-tailed. Approximately 89.81% of the variance in quiz scores was accounted for by the linear relationship with hours of sleep.

As always, the APA-formatted summary provides a lot of detail in a particular order. For a brief review of the structure for the APA-formatted summary of the test results, see the summary below.

Anatomy of the Evidence String

The following breaks down what each part represents in the evidence string for the correlation results in the APA-formatted paragraph above:

Symbol for the test	Degrees of Freedom	Obtained Value	p-Value
r	(8)	= .95,	p < .001

This page titled 11.3: Calculating and Reporting Pearson's r is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 10.4: Japan and Korea (North and South) by Staff is licensed CC BY-NC-SA 3.0. Original source: https://open.lib.umn.edu/worldgeography.

12: ONE-WAY ANOVA

12.1: Comparing several means (one-way ANOVA)

This page titled 12: One-Way ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

12.1: Comparing several means (one-way ANOVA)

This chapter introduces one of the most widely used tools in psychological statistics, known as "the analysis of variance", but usually referred to as ANOVA. The basic technique was developed by Sir Ronald Fisher in the early 20th century and it is to him that we owe the rather unfortunate terminology. The term ANOVA is a little misleading, in two respects. Firstly, although the name of the technique refers to variances, ANOVA is concerned with investigating differences in means. Secondly, there are several different things out there that are all referred to as ANOVAs, some of which have only a very tenuous connection to one another. Later on in the book we'll encounter a range of different ANOVA methods that apply in quite different situations, but for the purposes of this chapter we'll only consider the simplest form of ANOVA, in which we have several different groups of observations, and we're interested in finding out whether those groups differ in terms of some outcome variable of interest. This is the question that is addressed by a one-way ANOVA.

The structure of this chapter is as follows: in Section 12.1 I'll introduce a fictitious data set that we'll use as a running example throughout the chapter. After introducing the data, I'll describe the mechanics of how a one-way ANOVA actually works (Section 12.2) and then focus on how you can run one in JASP (Section 12.3). These two sections are the core of the chapter. The remainder of the chapter discusses a range of important topics that inevitably arise when running an ANOVA, namely how to calculate effect sizes (Section 12.4), post hoc tests and corrections for multiple comparisons (Section 12.5) and the assumptions that ANOVA relies upon (Section 12.6). We'll also talk about how to check those assumptions and some of the things you can do if the assumptions are violated (Sections 12.6.1 to 12.7). Then we'll cover repeated measures ANOVA in Sections 12.8 and 12.9. At the end of the chapter we'll talk a little about the relationship between ANOVA and other statistical tools (Section 12.10).

12.1 An illustrative data set

Suppose you've become involved in a clinical trial in which you are testing a new antidepressant drug called Joyzepam. In order to construct a fair test of the drug's effectiveness, the study involves three separate drugs to be administered. One is a placebo, and the other is an existing antidepressant / anti-anxiety drug called Anxifree. A collection of 18 participants with moderate to severe depression are recruited for your initial testing. Because the drugs are sometimes administered in conjunction with psychological therapy, your study includes 9 people undergoing cognitive behavioural therapy (CBT) and 9 who are not. Participants are randomly assigned (doubly blinded, of course) a treatment, such that there are 3 CBT people and 3 no-therapy people assigned to each of the 3 drugs. A psychologist assesses the mood of each person after a 3 month run with their assigned drug, and the overall improvement in each person's mood is assessed on a scale ranging from -5 to +5. With that as the study design, let's now load up the data file in clinicaltrial.csv. We can see that this data set contains the three variables drug, therapy and mood.gain.

For the purposes of this chapter, what we're really interested in is the effect of drug on mood.gain. The first thing to do is calculate some descriptive statistics and draw some graphs. You can calculate these descriptive statistics in JASP using the Split box in Descriptives → Descriptive Statistics. The results are shown in Figure 12.1.

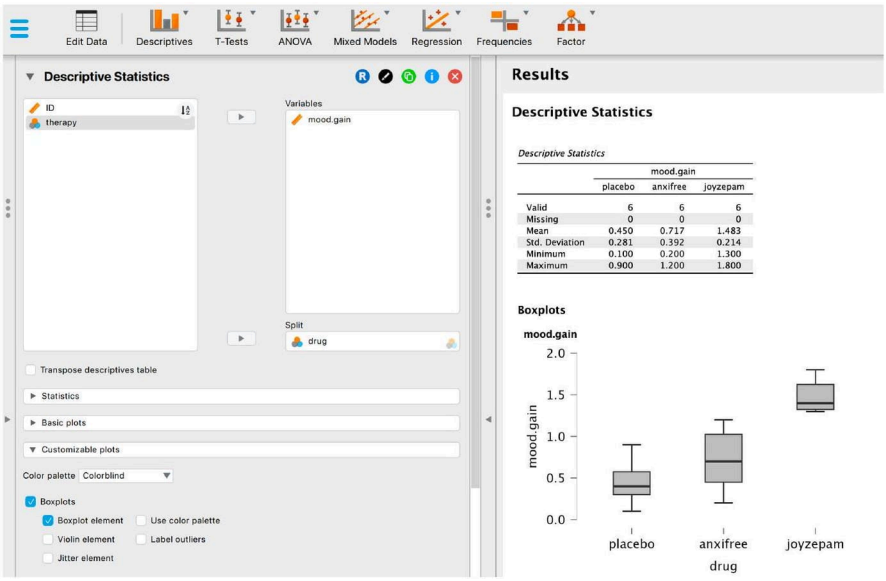


Figure 12.1: JASP screenshot showing descriptives and box plots for mood gain, split by drug administered

As the plot makes clear, there is a larger improvement in mood for participants in the Joyzepam group than for either the Anxifree group or the placebo group. The Anxifree group shows a larger mood gain than the control group, but the difference isn't as large. The question that we want to answer is are these difference "real", or are they just due to chance?

12.2 How ANOVA works

In order to answer the question posed by our clinical trial data we're going to run a one-way ANOVA. I'm going to start by showing you how to do it the hard way, building the statistical tool from the ground up and showing you how you could do it if you didn't have access to any of the cool built-in ANOVA functions in JASP. And I hope you'll read it carefully, try to do it the long way once or twice to make sure you really understand how ANOVA works, and then once you've grasped the concept never ever do it this way again.

The experimental design that I described in the previous section strongly suggests that we're interested in comparing the average mood change for the three different drugs. In that sense, we're talking about an analysis similar to the t- test (Chapter 10) but involving more than two groups. If we let μ_P denote the population mean for the mood change induced by the placebo, and let μ_A and μ_J denote the corresponding means for our two drugs, Anxifree and Joyzepam, then the (somewhat pessimistic) null hypothesis that we want to test is that all three population means are identical. That is, neither of the two drugs is any more effective than a placebo. We can write out this null hypothesis as:

$$H_0: \text{ it is true that } \mu_P = \mu_A = \mu_J$$

As a consequence, our alternative hypothesis is that at least one of the three different treatments is different from the others. It's a bit tricky to write this mathematically, because (as we'll discuss) there are quite a few different ways in which the null hypothesis can be false. So for now we'll just write the alternative hypothesis like this:

$$H_1: \text{ it is not true that } \mu_P = \mu_A = \mu_J$$

This null hypothesis is a lot trickier to test than any of the ones we've seen previously. How shall we do it? A sensible guess would be to "do an ANOVA", since that's the title of the chapter, but it's not particularly clear why an "analysis of variances" will help us learn anything useful about the means. In fact, this is one of the biggest conceptual difficulties that people have when first encountering ANOVA. To see how this works, I find it most helpful to start by talking about variances. In fact, what I'm going to do is start by playing some mathematical games with the formula that describes the variance. That is, we'll start out by playing around with variances and it will turn out that this gives us a useful tool for investigating means.

12.2.1 Two formulas for the variance of Y

First, let's start by introducing some notation. We'll use G to refer to the total number of groups. For our data set there are three drugs, so there are $G = 3$ groups. Next, we'll use N to refer to the total sample size; there are a total of $N = 18$ people in our data set. Similarly, let's use N_k to denote the number of people in the k- th group. In our fake clinical trial, the sample size is $N_k = 6$ for all three groups. [1] Finally, we'll use Y to denote the outcome variable. In our case, Y refers to mood change. Specifically, we'll use Y_{ik} to refer to the mood change experienced by the i- th member of the k- th group. Similarly, we'll use $\bar{Y}$ to be the average mood change, taken across all 18 people in the experiment, and $\bar{Y}_k$ to refer to the average mood change experienced by the 6 people in group k .

Now that we've got our notation sorted out we can start writing down formulas. To start with, let's recall the formula for the variance, way back in those kinder days when we were just doing descriptive statistics. The sample variance of Y is defined as follows

$$\mathrm{Var}(Y) = \frac{1}{N} \sum_{k=1}^{G} \sum_{i=1}^{N_k} \left(Y_{ik} - \bar{Y} \right)^2$$

This is the same variance calculation, written using group-specific notation. The only difference is that this time around I've got two summations here: I'm summing over groups (i.e., values for k) and over the people within the groups (i.e., values for i). This is purely a cosmetic detail. If I'd instead used the notation Y_p to refer to the value of the outcome variable for person p in the sample, then I'd only have a single summation. The only reason that we have a double summation here is that I've classified people into groups, and then assigned numbers to people within groups.

A concrete example might be useful here. Let's consider this table, in which we have a total of $N = 5$ people sorted into $G = 2$ groups. Arbitrarily, let's say that the "cool" people are group 1 and the "uncool" people are group 2. It turns out that we have three cool people $(N_1 = 3)$ and two uncool people $(N_2 = 2)$

name	person p	group	group num. k	index in group i	grumpiness Y_{ik} or Y_p
Ann	1	cool	1	1	20
Ben	2	cool	1	2	55
Cat	3	cool	1	3	21
Dan	4	uncool	2	1	91
Egg	5	uncool	2	2	22

Notice that I've constructed two different labelling schemes here. We have a "person" variable p so it would be perfectly sensible to refer to Y_p as the grumpiness of the p- th person in the sample. For instance, the table shows that Dan is the fourth so we'd say $p = 4$. So, when talking about the grumpiness Y of this "Dan" person, whoever he might be, we could refer to his grumpiness by saying that $Y_p = 91$, for person $p = 4$ that is. However, that's not the only way we could refer to Dan. As an alternative we could note that Dan belongs to the "uncool" group $(k = 2)$, and is in fact the first person listed in the uncool group $(i = 1)$. So it's equally valid to refer to Dan's grumpiness by saying that $Y_{ik} = 91$, where $k = 2$ and $i = 1$.

In other words, each person p corresponds to a unique ik combination, and so the formula that I gave above is actually identical to our original formula for the variance, which would be

$$\mathrm{Var}(Y) = \frac{1}{N} \sum_{p=1}^{N} \left(Y_p - \bar{Y} \right)^2$$

In both formulas, all we're doing is summing over all of the observations in the sample. Most of the time we would just use the simpler Y_p notation; the equation using Y_p is clearly the simpler of the two. However, when doing an ANOVA it's important to keep track of which participants belong in which groups, and we need to use the Y_{ik} notation to do this.

12.2.2 From variances to sums of squares

Okay, now that we've got a good grasp on how the variance is calculated, let's define something called the total sum of squares, which is denoted $\mathrm{SS}_{\mathrm{tot}}$. This is very simple. Instead of averaging the squared deviations, which is what we do when calculating

the variance, we just add them up.

So the formula for the total sum of squares is almost identical to the formula for the variance

$$\mathrm{SS}_{\mathrm{tot}} = \sum_{k=1}^{G} \sum_{i=1}^{N_k} \left(Y_{ik} - \bar{Y}\right)^2$$

When we talk about analysing variances in the context of ANOVA, what we're really doing is working with the total sums of squares rather than the actual variance. One very nice thing about the total sum of squares is that we can break it up into two different kinds of variation.

First, we can talk about the within-group sum of squares, in which we look to see how different each individual person is from their own group mean

$$\mathrm{SS}_w = \sum_{k=1}^{G} \sum_{i=1}^{N_k} \left(Y_{ik} - \bar{Y}_k\right)^2$$

where $\bar{Y}_k$ is a group mean. In our example, $\bar{Y}_k$ would be the average mood change experienced by those people given the k- th drug. So, instead of comparing individuals to the average of all people in the experiment, we're only comparing them to those people in the the same group. As a consequence, you'd expect the value of SS_w to be smaller than the total sum of squares, because it's completely ignoring any group differences, i.e., whether the drugs will have different effects on people's moods.

Next, we can define a third notion of variation which captures only the differences between groups. We do this by looking at the differences between the group means $\bar{Y}_k$ and grand mean $\bar{Y}$.

In order to quantify the extent of this variation, what we do is calculate the between-group sum of squares

$$\begin{aligned} \mathrm{SS}_b &= \sum_{k=1}^{G} \sum_{i=1}^{N_k} \left(\bar{Y}_k - \bar{Y}\right)^2 \\ &= \sum_{k=1}^{G} N_k \left(\bar{Y}_k - \bar{Y}\right)^2 \end{aligned}$$

It's not too difficult to show that the total variation among people in the experiment $\mathrm{SS}_{\mathrm{tot}}$ is actually the sum of the differences between the groups SS_b and the variation inside the groups SS_w. That is,

$$\mathrm{SS}_w + \mathrm{SS}_b = \mathrm{SS}_{\mathrm{tot}}$$

Yay.

Okay, so what have we found out? We've discovered that the total variability associated with the outcome variable $(\mathrm{SS}_{\mathrm{tot}})$ can be mathematically carved up into the sum of "the variation due to the differences in the sample means for the different groups" (SS_b) plus "all the rest of the variation" $(\mathrm{SS}_w)^2$. How does that help me find out whether the groups have different population means? Um. Wait. Hold on a second. Now that I think about it, this is exactly what we were looking for. If the null hypothesis is true then you'd expect all the sample means to be pretty similar to each other, right? And that would imply that you'd expect SS_b to be really small, or at least you'd expect it to be a lot smaller than "the variation associated with everything else", SS_w. Hmm. I detect a hypothesis test coming on.

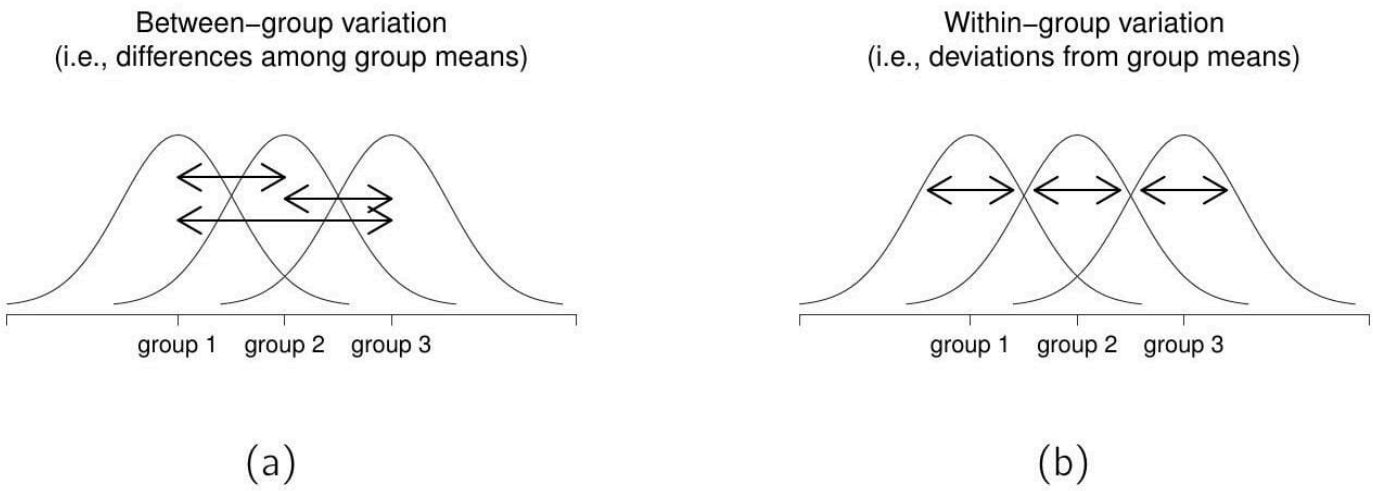


Figure 12.2: Graphical illustration of "between groups" variation (panel a) and "within groups" variation (panel b). On the left the arrows show the differences in the group means. On the right the arrows highlight the variability within each group.

12.2.3 From sums of squares to the F- test

As we saw in the last section, the qualitative idea behind ANOVA is to compare the two sums of squares values SS_b and SS_w to each other. If the between-group variation SS_b is large relative to the within-group variation SS_w then we have reason to suspect that the population means for the different groups aren't identical to each other. In order to convert this into a workable hypothesis test, there's a little bit of "fiddling around" needed. What I'll do is first show you what we do to calculate our test statistic, the F ratio, and then try to give you a feel for why we do it this way.

In order to convert our SS values into an F- ratio the first thing we need to calculate is the degrees of freedom associated with the SS_b and SS_w values. As usual, the degrees of freedom corresponds to the number of unique "data points" that contribute to a particular calculation, minus the number of "constraints" that they need to satisfy. For the within-groups variability what we're calculating is the variation of the individual observations (N data points) around the group means (G constraints). In contrast, for the between groups variability we're interested in the variation of the group means (G data points) around the grand mean (1 constraint). Therefore, the degrees of freedom here are:

$$\mathrm{df}_b = G - 1$$
$$\mathrm{df}_w = N - G$$

Okay, that seems simple enough. What we do next is convert our summed squares value into a "mean squares" value, which we do by dividing by the degrees of freedom:

$$\mathrm{MS}_b = \frac{\mathrm{SS}_b}{\mathrm{df}_b}$$
$$\mathrm{MS}_w = \frac{\mathrm{SS}_w}{\mathrm{df}_w}$$

Table 12.1: All of the key quantities involved in an ANOVA organised into a "standard" ANOVA table. The formulas for all quantities (except the p- value which has a very ugly formula and would be nightmarishly hard to calculate without a computer) are shown.

Source	df	Sum of squares	Mean square	F statistic
between groups	$\mathrm{df}_b = G - 1$	$\mathrm{SS}_b = \sum_{k=1}^{G} N_k \left(\bar{Y}_k - \bar{Y}\right)^2$	$\mathrm{MS}_b = \frac{\mathrm{SS}_b}{\mathrm{df}_b}$	$F = \frac{\mathrm{MS}_b}{\mathrm{MS}_w}$
within groups	$\mathrm{df}_w = N - G$	$\mathrm{SS}_w = \sum_{k=1}^{G} \sum_{i=1}^{N_k} \left(Y_{ik} - \bar{Y}_k\right)^2$	$\mathrm{MS}_w = \frac{\mathrm{SS}_w}{\mathrm{df}_w}$	-

Finally, we calculate the F- ratio by dividing the between-groups MS by the within-groups MS:

$$F = \frac{\mathrm{MS}_b}{\mathrm{MS}_w}$$

At a very general level, the intuition behind the F statistic is straightforward. Bigger values of F means that the between-groups variation is large relative to the within-groups variation. As a consequence, the larger the value of F the more evidence we have against the null hypothesis. But how large does F have to be in order to actually reject H_0 ? In order to understand this, you need a slightly deeper understanding of what ANOVA is and what the mean squares values actually are.

The next section discusses that in a bit of detail, but for readers that aren't interested in the details of what the test is actually measuring I'll cut to the chase. In order to complete our hypothesis test we need to know the sampling distribution for F if the null hypothesis is true. Not surprisingly, the sampling distribution for the F statistic under the null hypothesis is an F distribution. The F distribution has two parameters, corresponding to the two degrees of freedom involved. The first one df_1 is the between groups degrees of freedom df_b, and the second one df_2 is the within groups degrees of freedom df_w .

A summary of all the key quantities involved in a one-way ANOVA, including the formulas showing how they are calculated, is shown in Table 12.1.

12.2.4 The model for the data and the meaning of F

At a fundamental level ANOVA is a competition between two different statistical models, H_0 and H_1. When I described the null and alternative hypotheses at the start of the section, I was a little imprecise about what these models actually are. I'll remedy that now, though you probably won't like me for doing so. If you recall, our null hypothesis was that all of the group means are identical to one another. If so, then a natural way to think about the outcome variable Y_{ik} is to describe individual scores in terms of a single population mean μ, plus the deviation from that population mean. This deviation is usually denoted ε_{ik} and is traditionally called the error or residual associated with that observation. Be careful though. Just like we saw with the word "significant", the word "error" has a technical meaning in statistics that isn't quite the same as its everyday English definition. In everyday language, "error" implies a mistake of some kind, but in statistics it doesn't (or at least, not necessarily). With that in mind, the word "residual" is a better term than the word "error". In statistics both words mean "leftover variability", that is "stuff" that the model can't explain.

In any case, here's what the null hypothesis looks like when we write it as a statistical model

$$Y_{ik} = \mu + \varepsilon_{ik}$$

where we make the assumption (discussed later) that the residual values ε_{ik} are normally distributed, with mean 0 and a standard deviation σ that is the same for all groups. We would write this assumption like this

$$\varepsilon_{ik} \sim \mathrm{Normal}\left(0, \sigma^2\right)$$

What about the alternative hypothesis, H_1 ? The only difference between the null hypothesis and the alternative hypothesis is that we allow each group to have a different population mean. So, if we let μ_k denote the population mean for the k- th group in our experiment, then the statistical model corresponding to H_1 is

$$Y_{ik} = \mu_k + \varepsilon_{ik}$$

where, once again, we assume that the error terms are normally distributed with mean 0 and standard deviation σ. That is, the alternative hypothesis also assumes that $\varepsilon \sim \mathrm{Normal}\left(0, \sigma^2\right)$

Okay, now that we've described the statistical models underpinning H_0 and H_1 in more detail, it's now pretty straightforward to say what the mean square values are measuring, and what this means for the interpretation of F. I won't bore you with the proof of this but it turns out that the within-groups mean square, MS_w, can be viewed as an estimator of the error variance σ^2. The between-groups mean square MS_b is also an estimator, but what it estimates is the error variance plus a quantity that depends on the true differences among the group means. If we call this quantity Q, then we can see that the F- statistic is basically [a]

$$F = \frac{\hat{Q} + \hat{\sigma}^2}{\hat{\sigma}^2}$$

where the true value $Q = 0$ if the null hypothesis is true, and $Q > 0$ if the alternative hypothesis is true (e.g., Hays 1994, ch. 10).

Therefore, at a bare minimum the F value must be larger than 1 to have any chance of rejecting the null hypothesis. Note that this doesn't mean that it's impossible to get an F- value less than 1 . Under the null hypothesis, both mean squares estimate the same error variance, so values of F around 1 are expected. Large F values provide evidence against the null hypothesis.

Under the null hypothesis and the ANOVA assumptions, SS_{β}/σ^2 and $SS\diamondsuit/\sigma^2$ have independent chi-square distributions. Dividing each by its degrees of freedom and taking their ratio gives an F distribution with G – 1 and N – G degrees of freedom.

12.2.5 A worked example

The previous discussion was fairly abstract and a little on the technical side, so I think that at this point it might be useful to see a worked example. For that, let's go back to the clinical trial data that I introduced at the start of the chapter. The descriptive statistics that we calculated at the beginning tell us our group means: an average mood gain of 0.45 for the placebo, approximately 0.72 for Anxifree, and approximately 1.48 for Joyzepam. With that in mind, let's party like it's 1899 and start doing some pencil and paper calculations. I'll only show the individual calculations for the first five observations because it's not bloody 1899 and I'm very lazy.

The tables display rounded values for readability. Keep the unrounded values when calculating sums of squares and the final F statistic.

Let's start by calculating SS_w, the within-group sum of squares. First, let's draw up a nice table to help us with our calculations:

First five observations from the clinical trial data.

Group	Outcome
Placebo	0.5
Placebo	0.3
Placebo	0.1
Anxifree	0.6
Anxifree	0.4

At this stage, the only thing I've included in the table is the raw data itself. That is, the grouping variable (drug) and outcome variable (mood.gain) for each person. Note that the outcome variable here corresponds to the Y_{ik} value in our equation previously. The next step in the calculation is to write down, for each person in the study, the corresponding group mean, $\bar{Y}_k$. This is slightly repetitive but not particularly difficult since we already calculated those group means when doing our descriptive statistics:

First five observations with their corresponding group means.

Group	Outcome	Group mean
Placebo	0.5	0.450000
Placebo	0.3	0.450000
Placebo	0.1	0.450000
Anxifree	0.6	0.716667
Anxifree	0.4	0.716667

Now that we've written those down, we need to calculate, again for every person, the deviation from the corresponding group mean. That is, we calculate $Y_{ik} - \bar{Y}_k$. After we've done that, we need to square everything. When we do that, here's what we get:

Deviations and squared deviations for the first five observations.

Group	Outcome	Group mean	Deviation from group mean	Squared deviation
Placebo	0.5	0.450000	0.050000	0.002500
Placebo	0.3	0.450000	−0.150000	0.022500

Group	Outcome	Group mean	Deviation from group mean	Squared deviation
Placebo	0.1	0.450000	−0.350000	0.122500
Anxifree	0.6	0.716667	−0.116667	0.013611
Anxifree	0.4	0.716667	−0.316667	0.100278

The last step is equally straightforward. In order to calculate the within-group sum of squares, we add up the squared deviations across all observations. For the first five observations shown here, the sum is:

$$0.002500 + 0.022500 + 0.122500 + 0.013611 + 0.100278 \approx 0.261389 \tag{12.1.2}$$

Of course, if we actually wanted the complete within-group sum of squares, we'd need to do this for all 18 observations in the data set, not just the first five. We could continue with the pencil and paper calculations if we wanted to, but it's pretty tedious. Alternatively, it's not too hard to do this in a dedicated spreadsheet program such as OpenOffice or Excel. Try and do it yourself. When you do it, you should end up with:

$$\mathrm{SS}_w = \sum_{k=1}^{G} \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2 \approx 1.391667 \tag{12.1.3}$$

Okay. Now that we've calculated the within-group variation, SS_w, it's time to turn our attention to the between-group sum of squares, SS_b. The calculations for this case are very similar. The main difference is that instead of calculating the differences between an observation and its group mean for all observations, we calculate the differences between the group means and the grand mean for all groups.

The grand mean for these data is approximately 0.883333. For the between-group calculation, we multiply each squared deviation by N_k, the number of observations in that group. We do this because every observation in the group is associated with that group mean. For example, the placebo group's squared deviation is approximately 0.187778. Since there are six people in that group, its contribution to the between-group sum of squares is approximately $6(0.187778) = 1.126667$, using unrounded values.

Calculation of the between-group sum of squares.

Group	Group mean	Grand mean	Deviation	Squared deviation	Sample size	Weighted squared deviation
Placebo	0.450000	0.883333	−0.433333	0.187778	6	1.126667
Anxifree	0.716667	0.883333	−0.166667	0.027778	6	0.166667
Joyzepam	1.483333	0.883333	0.600000	0.360000	6	2.160000

And so now our between-group sum of squares is obtained by summing these weighted squared deviations over all three groups in the study. Using unrounded values:

$$\mathrm{SS}_b = \sum_{k=1}^{G} N_k (\bar{Y}_k - \bar{Y})^2 \approx 3.453333 \tag{12.1.4}$$

As you can see, the between-group calculations are a lot shorter. Now that we've calculated our sums of squares, the rest of the ANOVA is pretty painless. The next step is to calculate the degrees of freedom. Since we have $G = 3$ groups and $N = 18$ observations in total, our degrees of freedom can be calculated by simple subtraction:

$$df_b = G - 1 = 3 - 1 = 2 \tag{12.1.5}$$

$$df_w = N - G = 18 - 3 = 15 \tag{12.1.6}$$

Next, we obtain the mean square values by dividing each sum of squares by its corresponding degrees of freedom:

$$\mathrm{MS}_b = \frac{\mathrm{SS}_b}{df_b} \approx \frac{3.453333}{2} \approx 1.726667 \tag{12.1.7}$$

$$\mathrm{MS}_w = \frac{\mathrm{SS}_w}{df_w} \approx \frac{1.391667}{15} \approx 0.092778 \tag{12.1.8}$$

We're almost done. The mean square values can be used to calculate the F statistic. We do this by dividing the between-group mean square by the within-group mean square. Keeping the unrounded values in the calculation gives:

$$F = \frac{\text{MS}_b}{\text{MS}_w} \approx 18.611 \tag{12.1.9}$$

Woohooo! This is terribly exciting, yes? Now that we have our test statistic, the last step is to find out whether the test itself gives us a significant result. Back in the old days, we'd open up a statistics textbook or flick to the back section, which would have a huge lookup table. We would find the critical value corresponding to a particular alpha level and our 2 and 15 degrees of freedom. For an alpha level of .001, the critical value is approximately 11.34. Since our calculated F statistic exceeds this value, we can conclude that $p < .001$.

Nowadays, statistical software calculates the p-value for you. For these data, it is approximately 0.0000865. We therefore reject the null hypothesis at the conventional .05 significance level.

At this point, we're basically done. Having completed our calculations, it's traditional to organize these numbers into an ANOVA table:

One-way ANOVA for mood gain by drug.

Source	df	Sum of squares	Mean square	F	p
Between groups	2	3.453	1.727	18.611	< .001
Within groups	15	1.392	0.093	—	—

Statistical software, including JASP, tends to organize ANOVA output into a table like this, so it's a good idea to get used to reading one. However, you don't usually need to include the entire table in your write-up. A standard way of reporting this result would be:

> A one-way ANOVA showed a significant effect of drug on mood gain, $F(2, 15) = 18.61$, $p < .001$.

Sigh. So much work for one short sentence.

12.3 Running an ANOVA in JASP

I'm pretty sure I know what you're thinking after reading the last section, especially if you followed my advice and did all of that by pencil and paper (i.e., in a spreadsheet) yourself. Doing the ANOVA calculations yourself sucks. There's quite a lot of calculations that we needed to do along the way, and it would be tedious to have to do this over and over again every time you wanted to do an ANOVA.

12.3.1 Using JASP to specify your ANOVA

To make life easier for you, JASP can do ANOVA..hurrah! Go to the 'ANOVA' - 'ANOVA' analysis, and move the mood.gain variable across so it is in the 'Dependent Variable' box, and then move the drug variable across so it is in the 'Fixed Factors' box. This should give the results as shown in Figure 12.3. [4] Note I have also checked the η^2 checkbox, pronounced "eta squared", under the 'Estimates

of effect size' option that can be found in 'Additional Options,' and this is also shown on the results table. We will come back to effect sizes a bit later.

ANOVA

ANOVA - mood.gain						
Cases	Sum of Squares	df	Mean Square	F	p	η^2
drug	3.453	2	1.727	18.611	< . 001	0.713

ANOVA - mood.gain						
Residuals	1.392	15	0.093			
Note. Type III Sum of Squares						

Figure 12.3: JASP results table for ANOVA of mood gain by drug administered

The JASP results table shows you the sums of squares values, the degrees of freedom, and a couple of other quantities that we're not really interested in right now. Notice, however, that JASP doesn't use the names "between-group" and "within-group". Instead, it tries to assign more meaningful names. In our particular example, the between groups variance corresponds to the effect that the drug has on the outcome variable, and the within groups variance corresponds to the "leftover" variability so it calls that the residual error. If we compare these numbers to the numbers that I calculated by hand in Section 12.2.5, you can see that they're more or less the same, apart from rounding errors. The between groups sums of squares is $\mathrm{SS}_b = 3.453$, the within groups sums of squares is $\mathrm{SS}_w = 1.392$, and the degrees of freedom are 2 and 15 respectively. We also get the F- value and the p- value and, again, these are more or less the same, give or take rounding errors, to the numbers that we calculated ourselves when doing it the long and tedious way.

12.4 Effect size

There's a few different ways you could measure the effect size in an ANOVA, but the most commonly used measures are η^2 (eta squared) and partial η^2. For a one way analysis of variance they're identical to each other, so for the moment I'll just explain η^2. The definition of η^2 is actually really simple

$$\eta^2 = \frac{\mathrm{SS}_b}{\mathrm{SS}_{\mathrm{tot}}}$$

That's all it is. So when I look at the ANOVA table in Figure 12.3, I see that $\mathrm{SS}_b = 3.45$ and $\mathrm{SS}_{\mathrm{tot}} = 3.45 + 1.39 = 4.84$. Thus we get an η^2 value of

$$\eta^2 = \frac{3.45}{4.84} = 0.71$$

The interpretation of η^2 is equally straightforward. It refers to the proportion of the variability in the outcome variable (mood.gain) that can be explained in terms of the predictor (drug). A value of $\eta^2 = 0$ means that there is no relationship at all between the two, whereas a value of $\eta^2 = 1$ means that the relationship is perfect. Better yet, the η^2 value is very closely related to R^2, and has an equivalent interpretation.

Although many statistics text books suggest η^2 as the default effect size measure in ANOVA, there's an interesting blog post by Daniel Lakens suggesting that eta-squared is perhaps not the best measure of effect size in real world data analysis, because it can be a biased estimator (http: //daniellakens.blogspot.com.au/2015/06/why-you-should-use-omega-squared.html). Usefully, there is also an option in JASP to specify omega-squared (ω^2), which is less biased, alongside eta-squared.

12.5 Multiple comparisons and post hoc tests

Any time you run an ANOVA with more than two groups and you end up with a significant effect, the first thing you'll probably want to ask is which groups are actually different from one another. In our drugs example, our null hypothesis was that all three drugs (placebo, Anxifree and Joyzepam) have the exact same effect on mood. But if you think about it, the null hypothesis is actually claiming three different things all at once here. Specifically, it claims that:

- -Your competitor's drug (Anxifree) has the same population mean mood gain as the placebo (i.e., $\mu_A = \mu_P$)
- -Your drug (Joyzepam) has the same population mean mood gain as the placebo (i.e., $\mu_J = \mu_P$)
- -Anxifree and Joyzepam are equally effective (i.e., $\mu_J = \mu_A$)

If any one of those three claims is false, then the null hypothesis is also false. So, now that we've rejected our null hypothesis, we're thinking that at least one of those things isn't true. But which ones? All three of these propositions are of interest. Since you certainly want to know if your new drug Joyzepam is better than a placebo, it would be nice to know how well it stacks up against an existing commercial alternative (i.e., Anxifree). It would even be useful to check the performance of Anxifree against

the placebo. Even if Anxifree has already been extensively tested against placebos by other researchers, it can still be very useful to check that your study is producing similar results to earlier work.

When we characterise the null hypothesis in terms of these three distinct propositions, it becomes clear that there are five possible "states of the world" that we need to distinguish between:

possibility:	is $\mu_P = \mu_A$?	is $\mu_P = \mu_J$?	is $\mu_A = \mu_J$?	which hypothesis?
1	✓	✓	✓	null
2	✓			alternative
3		✓		alternative
4			✓	alternative
5				alternative

By rejecting the null hypothesis, we've decided that we don't believe that #1 is the true state of the world. The next question to ask is, which of the other four possibilities do we think is right? When faced with this situation, its usually helps to look at the data. For instance, if we look at the plots in Figure 12.1, it's tempting to conclude that Joyzepam is better than the placebo and better than Anxifree, but there's no real difference between Anxifree and the placebo. However, if we want to get a clearer answer about this, it might help to run some tests.

12.5.1 Running "pairwise" t- tests

How might we go about solving our problem? Given that we've got three separate pairs of means (placebo versus Anxifree, placebo versus Joyzepam, and Anxifree versus Joyzepam) to compare, what we could do is run three separate t- tests and see what happens. This is easy to do in JASP. Go to the ANOVA 'Post Hoc Tests' options, move the 'drug' variable across into the active box on the right. This will immediately produce a neat table showing all the pairwise t- test comparisons amongst the three levels of the drug variable, as in Figure 12.4.

Post Hoc Tests

Standard (LSD)

Post Hoc Comparisons - drug

		Mean Difference	SE	t	p_{bonf}	p_{holm}
placebo	anxifree	-0.267	0.176	-1.516	0.451	0.150
	joyzepam	-1.033	0.176	-5.876	< .001	< .001
anxifree	joyzepam	-0.767	0.176	-4.360	0.002	0.001

Figure 12.4: Pairwise t- tests as post hoc comparisons in JASP

Note. P-value adjusted for comparing a family of 3

12.5.2 Corrections for multiple testing

In the previous section I hinted that there's a problem with just running lots and lots of t- tests. The concern is that, when running these analyses, what we're doing is going on a "fishing expedition". We're running lots and lots of tests without much theoretical guidance in the hope that some of them come up significant. This kind of theory-free search for group differences is referred to as post hoc analysis ("post hoc" being Latin for "after this"). [5]

It's okay to run post hoc analyses, but a lot of care is required. Running these pairwise tests without a multiple-comparison correction increases the chance of making at least one Type I error. Figure 12.4 includes adjusted p-values; the problem

discussed here concerns using the unadjusted results.

The usual solution to this problem is to introduce an adjustment to the p- value, which aims to control the total error rate across the family of tests (see Shaffer 1995). An adjustment of this form, which is usually (but not always) applied because one is doing post hoc analysis, is often referred to as a correction for multiple comparisons, though it is sometimes referred to as "simultaneous inference". In any case, there are quite a few different ways of doing this adjustment. I'll discuss a few of them in this section, but you should be aware that there are many other methods out there (see, e.g., Hsu 1996).

12.5.3 Bonferroni corrections

The simplest of these adjustments is called the Bonferroni correction (Dunn 1961), and it's very very simple indeed. Suppose that my post hoc analysis consists of m separate tests, and I want to ensure that the total probability of making any Type I errors at all is at most α.[6] If so, then the Bonferroni correction just says "multiply all your raw p- values by m ". If we let p denote the original p - value, and let p'_j be the corrected value, then the Bonferroni correction tells that:

$$p' = \min(1, mp) \tag{12.1.10}$$

If multiplication produces a value greater than 1, report the adjusted p-value as 1. And therefore, if you're using the Bonferroni correction, you would reject the null hypothesis if $p' < \alpha$. The logic behind this correction is very straightforward. We're doing m different tests, so if we arrange it so that each test has a Type I error rate of at most α/m, then the total Type I error rate across these tests cannot be larger than α. That's pretty simple, so much so that in the original paper, the author writes,

> The method given here is so simple and so general that I am sure it must have been used before this. I do not find it, however, so can only conclude that perhaps its very simplicity has kept statisticians from realizing that it is a very good method in some situations (Dunn 1961, pp 52-53)

To use the Bonferroni correction in JASP, just click on the 'Bonferroni' checkbox in the 'Correction' options, and you will see another column added to the ANOVA results table showing the adjusted p- values for the Bonferroni correction (12.4).

12.5.4 Holm corrections

Although the Bonferroni correction is the simplest adjustment out there, it's not usually the best one to use. One method that is often used instead is the Holm correction (Holm 1979). The idea behind the Holm correction is to pretend that you're doing the tests sequentially, starting with the smallest (raw) p- value and moving onto the largest one. For the j- th largest of the p- values, the adjustment is either

$$p'_j = j \times p_j$$

(i.e., the biggest p- value remains unchanged, the second biggest p- value is doubled, the third biggest p- value is tripled, and so on), or

$$p'_j = p'_{j+1}$$

whichever one is larger. As with Bonferroni, adjusted p-values cannot exceed 1. This might sound a little confusing, so let's go through it a little more slowly. Here's what the Holm correction does. First, you sort all of your p- values in order, from smallest to largest. For the smallest p- value all you do is multiply it by m, and you're done. However, for all the other ones it's a two-stage process. For instance, when you move to the second smallest p value, you first multiply it by $m - 1$. If this produces a number that is bigger than the adjusted p- value that you got last time, then you keep it. But if it's smaller than the last one, then you copy the last p- value. To illustrate how this works, consider the table below, which shows the calculations of a Holm correction for a collection of five p- values:

raw p	rank j	$p \times j$	Holm p
.001	5	.005	.005
.005	4	.020	.020
.019	3	.057	.057
.022	2	.044	.057

raw p	rank j	$p \times j$	Holm p
.103	1	.103	.103

Hopefully that makes things clear.

Although it's a little harder to calculate, the Holm correction has some very nice properties. It's more powerful than Bonferroni (i.e., it has a lower Type II error rate) but, counter-intuitive as it might seem, it has the same Type I error rate. As a consequence, in practice there's never any reason to use the simpler Bonferroni correction since it is always outperformed by the slightly more elaborate Holm correction. Because of this, the Holm correction should be your go to multiple comparison correction. Figure 12.4 shows both the Bonferroni and the Holm corrected p- values.

12.5.5 Writing up the post hoc test

Finally, having run the post hoc analysis to determine which groups are significantly different to one another, you might write up the result like this:

Post hoc tests (using the Holm correction to adjust p) indicated that Joyzepam produced a significantly larger mood change than both Anxifree $(p = .001)$ and the placebo $(p < 0.001)$. We found no evidence that Anxifree performed better than the placebo $(p = .15)$.

Or, if you don't like the idea of just reporting $p < 0.001$, you can compute exact p- values by going to 'Preferences' - 'Results', and then selecting 'Display exact p-values'. Either way, the key thing is that you indicate that you used Holm's correction to adjust the p- values. And of course, I'm assuming that elsewhere in the write up you've included the relevant descriptive statistics (i.e., the group means and standard deviations), since these p- values on their own aren't terribly informative.

12.6 Assumptions of one-way ANOVA

Like any statistical test, analysis of variance relies on some assumptions about the data, specifically the residuals. There are three key assumptions that you need to be aware of: normality, homogeneity of variance and independence.

Previously, I described the statistical models underpinning ANOVA in this way:

$$\begin{aligned} H_0: &\quad Y_{ik} = \mu + \varepsilon_{ik} \\ H_1: &\quad Y_{ik} = \mu_k + \varepsilon_{ik} \end{aligned}$$

In these equations μ refers to a single grand population mean which is the same for all groups, and μ_k is the population mean for the k- th group. Up to this point we've been mostly interested in whether our data are best described in terms of a single grand mean (the null hypothesis) or in terms of different group-specific means (the alternative hypothesis). This makes sense, of course, as that's actually the important research question! However, all of our testing procedures have,
implicitly, relied on a specific assumption about the residuals, ε_{ik}, namely that

$$\varepsilon_{ik} \sim \text{Normal}\left(0, \sigma^2\right)$$

None of the maths works properly without this bit. Or, to be precise, you can still do all the calculations and you'll end up with an F - statistic, but you have no guarantee that this F- statistic actually measures what you think it's measuring, and so any conclusions that you might draw on the basis of the F test might be wrong.

So, how do we check whether the assumption about the residuals is accurate? Well, as I indicated above, there are three distinct claims buried in this one statement, and we'll consider them separately.

- -Homogeneity of variance. Notice that we've only got the one value for the population standard deviation (i.e., σ), rather than allowing each group to have it's own value (i.e., σ_k). This is referred to as the homogeneity of variance (sometimes called homoscedasticity) assumption. ANOVA assumes that the population standard deviation is the same for all groups. We'll talk about this extensively in Section 12.6.1.
- -Normality. The residuals are assumed to be normally distributed.We can assess this by looking at QQ plots (or running a Shapiro-Wilk test. I'll talk about this more in an ANOVA context.
- -Independence. The independence assumption is a little trickier. What it basically means is that, knowing one residual

tells you nothing about any other residual. All of the ε_{ik} values are assumed to have been generated without any "regard for" or "relationship to" any of the other ones. There's not an obvious or simple way to test for this, but there are some situations that are clear violations of this. For instance, if you have a repeated-measures design, where each participant in your study appears in more than one condition, then independence doesn't hold. There's a special relationship between some observations, namely those that correspond to the same person! When that happens, you need to use something like repeated measures ANOVA (see Section 12.8).

12.6.1 Checking the homogeneity of variance assumption

> To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port! - George Box (Box 1953)

There's more than one way to skin a cat, as the saying goes, and more than one way to test the homogeneity of variance assumption, too (though for some reason no-one made a saying out of that). The most commonly used test for this that I've seen in the literature is the Levene test (Levene 1960), and the closely related Brown-Forsythe test (Brown and Forsythe 1974).

Regardless of whether you're doing the standard Levene test or the Brown-Forsythe test, the test statistic, which is sometimes denoted F but also sometimes written as W, is calculated in exactly the same way that the F- statistic for the regular ANOVA is calculated, just using a Z_{ik} rather than Y_{ik}. With that in mind, we can go on to look at how to run the test in JASP.

The Levene test is shockingly simple. Suppose we have our outcome variable Y_{ik}. All we do is define a new variable, which I'll call Z_{ik}, corresponding to the absolute deviation from the group mean

$$Z_{ik} = \left|Y_{ik} - \bar{Y}_k\right|$$

Okay, what good does this do us? Well, let's take a moment to think about what Z_{ik} actually is and what we're trying to test. The value of Z_{ik} is a measure of how the i- th observation in the k- th group deviates from its group mean. And our null hypothesis is that all groups have the same variance, i.e., the same overall deviations from the group means! So the null hypothesis in a Levene test is that the population means of Z are identical for all groups. Hmm. So what we need now is a statistical test of the null hypothesis that all group means are identical. Where have we seen that before? Oh right, that's what ANOVA is, and so all that the Levene test does is run an ANOVA on the new variable Z_{ik} .

What about the Brown-Forsythe test? Does that do anything particularly different? Nope. The only change from the Levene test is that it constructs the transformed variable Z in a slightly different way, using deviations from the group medians rather than deviations from the group means. That is, for the Brown-Forsythe test

$$Z_{ik} =\mid Y_{ik} - \text{ median }_k(Y) \mid$$

where $\text{median}_k(Y)$ is the median for group k .

12.6.2 Running the Levene test in JASP

Okay, so how do we run the Levene test? Simple really - under the ANOVA 'Assumption Checks' option, just click on the 'Homogeneity tests' checkbox. If we look at the output, shown in Figure 12.5, we see that the test is non-significant ($F_{2,15} = 1.45, p = .266$), so it looks like the homogeneity of variance assumption is fine. However, looks can be deceptive! If your sample size is pretty big, then the Levene test could show up a significant effect (i.e. $p < .05$) even when the homogeneity of variance assumption is not violated to an extent which troubles the robustness of ANOVA. This was the point George Box was making in the quote above. Similarly, if your sample size is quite small, then the homogeneity of variance assumption might not be satisfied and yet a Levene test could be non-significant (i.e. $p > .05$). What this means is that, alongside any statistical test of the assumption being met, you should always plot the standard deviation around the means for each group / category in the analysis...just to see if they look fairly similar (i.e. homogeneity of variance) or not.

Assumption Checks

Test for Equality of Variances (Levene's)

F	df1	df2	p
1.450	2.000	15.000	0.266

Figure 12.5: Levene test output for one-way ANOVA in JASP

12.6.3 Removing the homogeneity of variance assumption

In our example, the homogeneity of variance assumption turned out to be a pretty safe one: the Levene test came back non-significant (notwithstanding that we should also look at the plot of standard deviations), so we probably don't need to worry. However, in real life we aren't always that lucky. How do we save our ANOVA when the homogeneity of variance assumption is violated? If you recall from our discussion of t- tests, we've seen this problem before. The Student t- test assumes equal variances, so the solution was to use the Welch t- test, which does not. In fact, Welch (1951) also showed how we can solve this problem for ANOVA too (the Welch one-way test). It's implemented in JASP using the One-Way ANOVA analysis. To implement the Welch correction, you simply need to choose the option 'Welch' under 'Assumption Checks' - 'Homogeneity corrections'. The result is displayed in Figure 12.6.

ANOVA - mood.gain

Homogeneity Correction	Cases	Sum of Squares	df	Mean Square	F	p	η^2
None	drug	3.453	2.000	1.727	18.611	< . 001	0.713
	Residuals	1.392	15.000	0.093			
Welch	drug	3.453	2.000	1.727	26.322	< . 001	0.713
	Residuals	1.392	9.493	0.147			

Figure 12.6: Welch's homogeneity correction as part of the One Way ANOVA analysis in JASP

Note. Type III Sum of Squares

To understand what's happening here, let's compare these numbers to what we got earlier in Section 12.3 when we ran our original ANOVA. To save you the trouble of flicking back, this is what we got last time: $F(2, 15) = 18.611, p < 0.001$, also shown as None under Homogeneity Correction in
the One-Way ANOVA shown in Figure 12.6.
Okay, so originally our ANOVA gave us the result $F(2, 15) = 18.6$, whereas the Welch correction gave us $F(2, 9.49) = 26.32$. Welch's ANOVA uses a different weighted test statistic and approximate denominator degrees of freedom. Here it gives F(2, 9.49) = 26.32, p < .001.

12.6.4 Checking the normality assumption

Testing the normality assumption is relatively straightforward. The only thing we really need to do is draw a QQ plot [7]. The QQ plot can be produced in JASP by going to 'Assumption Checks' and checking 'Q-Q plot of residuals'. The result is shown in Figure 12.7 and it looks pretty normal to me.

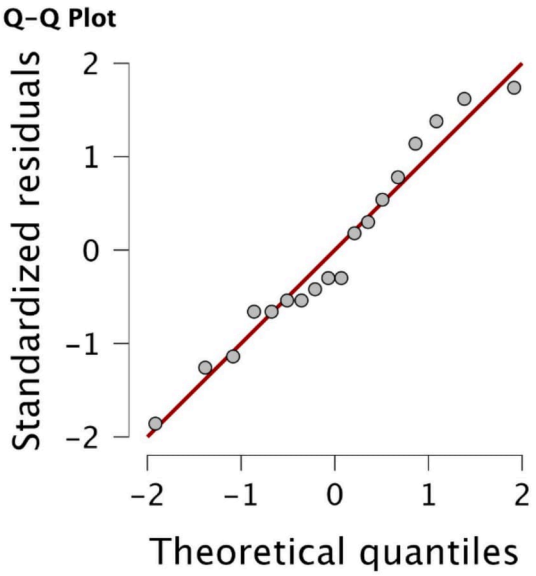


Figure 12.7: QQ plot produced from JASP

12.7 Removing the normality assumption

Now that we've seen how to check for normality, we are led naturally to ask what we can do to address violations of normality. In the context of a one-way ANOVA, the easiest solution is probably to switch to a non-parametric test (i.e., one that does not require normally distributed errors). When you only have two groups, the Mann-Whitney or the Wilcoxon test provides the non-parametric alternative that you need. When you've got three or more groups, you can use the Kruskal-Wallis rank sum test (Kruskal and Wallis 1952). So that's the test we'll talk about next.

12.7.1 The logic behind the Kruskal-Wallis test

The Kruskal-Wallis test is surprisingly similar to ANOVA, in some ways. In ANOVA we started with Y_{ik}, the value of the outcome variable for the i th person in the k th group. For the Kruskal-Wallis test what we'll do is rank order all of these Y_{ik} values and conduct our analysis on the ranked data.

So let's let R_{ik} refer to the ranking given to the i th member of the k th group. Now, let's calculate

$\bar{R}_k$, the average rank given to observations in the k th group

$$\bar{R}_k = \frac{1}{N_K} \sum_i R_{ik}$$

and let's also calculate $\bar{R}$, the grand mean rank

$$\bar{R} = \frac{1}{N} \sum_i \sum_k R_{ik}$$

Now that we've done this, we can calculate the squared deviations from the grand mean rank $\bar{R}$. When we do this for the individual scores, i.e., if we calculate $\left(R_{ik} - \bar{R}\right)^2$, what we have is a "nonparametric" measure of how far the ik- th observation deviates from the grand mean rank. When we calculate the squared deviation of the group means from the grand means, i.e., if we calculate $\left(\bar{R}_k - \bar{R}\right)^2$, then what we have is a nonparametric measure of how much the group deviates from the grand mean rank. With this in mind, we'll follow the same logic that we did with ANOVA and define our ranked sums of squares measures, much like we did earlier. First, we have our "total ranked sums of squares"

$$\mathrm{RSS}_{\mathrm{tot}} = \sum_k \sum_i \left(R_{ik} - \bar{R}\right)^2$$

and we can define the "between groups ranked sums of squares" like this

$$\begin{aligned} \mathrm{RSS}_b &= \sum_k \sum_i \left(\bar{R}_k - \bar{R}\right)^2 \\ &= \sum_k N_k \left(\bar{R}_k - \bar{R}\right)^2 \end{aligned}$$

So, if the null hypothesis is true and there are no true group differences at all, you'd expect the between group rank sums RSS_b to be very small, much smaller than the total rank sums $\mathrm{RSS}_{\mathrm{tot}}$. Qualitatively this is very much the same as what we found when we went about constructing the ANOVA F- statistic, but for technical reasons the Kruskal-Wallis test statistic, usually denoted K, is constructed in a slightly different way,

$$K = (N-1) \times \frac{\mathrm{RSS}_b}{\mathrm{RSS}_{\mathrm{tot}}}$$

and if the null hypothesis is true, then the sampling distribution of K is approximately chi-square with $G-1$ degrees of freedom (where G is the number of groups). The larger the value of K, the less consistent the data are with the null hypothesis, so this is a one-sided test. We reject H_0 when K is sufficiently large.

12.7.2 Additional details

The description in the previous section illustrates the logic behind the Kruskal-Wallis test. At a conceptual level, this is the right way to think about how the test works. However, from a purely mathematical perspective it's needlessly complicated. I won't show you the derivation, but you can use a bit of algebraic jiggery-pokery [a] to show that the equation for K can be rewritten as

$$K = \frac{12}{N(N-1)} \sum_k N_k \bar{R}_k^2 - 3(N+1)$$

It's this last equation that you sometimes see given for K. This is way easier to calculate than the version I described in the previous section, but it's just that it's totally meaningless to actual humans. It's probably best to think of K the way I described it earlier, as an analogue of ANOVA based on ranks. But keep in mind that the test statistic that gets calculated ends up with a rather different look to it than the one we used for our original ANOVA.

But wait, there's more! Dear lord, why is there always more? The story I've told so far is only actually true when there are no ties in the raw data. That is, if there are no two observations that have exactly the same value. If there are ties, then we have to introduce a correction factor to these calculations. At this point I'm assuming that even the most diligent reader has stopped caring (or at least formed the opinion that the tie-correction factor is something that doesn't require their immediate attention). So I'll very quickly tell you how it's calculated, and omit the tedious details about why it's done this way. Suppose we construct a frequency table for the raw data, and let f_j be the number of observations that have the j- th unique value. This might sound a bit abstract, so here's a concrete example from the frequency table of mood.gain from the clinicaltrials.csv data set:

0.1	0.2	0.3	0.4	0.5	0.6	0.8	0.9	1.1	1.2	1.3	1.4	1.7	1.8
1	1	2	1	1	2	1	1	1	1	2	2	1	1

Looking at this table, notice that the third entry in the frequency table has a value of 2. Since this corresponds to a mood.gain of 0.3, this table is telling us that two people's mood increased by 0.3. More to the point, in the mathematical notation I introduced above, this is telling us that $f_3 = 2$. Yay. So, now that we know this, the tie correction factor (TCF) is:

$$\mathrm{TCF} = 1 - \frac{\sum_j f_j^3 - f_j}{N^3 - N}$$

The tie-corrected value of the Kruskal-Wallis statistic is obtained by dividing the value of K by this quantity. It is this tie-corrected version that JASP calculates. And at long last, we're actually finished with the theory of the Kruskal-Wallis test. I'm sure you're all terribly relieved that I've cured
you of the existential anxiety that naturally arises when you realise that you don't know how to calculate the tie-correction factor for the Kruskal-Wallis test. Right?

[a] A technical term.

12.7.3 How to run the Kruskal-Wallis test in JASP

Despite the horror that we've gone through in trying to understand what the Kruskal-Wallis test actually does, it turns out that running the test is pretty painless, since JASP has an analysis as part of the ANOVA analysis set called 'Nonparametrics'. All you need to do is move the grouping variable drug into the active box. What this gives us is a Kruskal-Wallis $\chi^2 = 12.076, \mathrm{df} = 2, p$-value $= 0.002$, as in Figure 12.8

Kruskal-Wallis Test

Kruskal-Wallis Test

Factor	Statistic	df	p
drug	12.076	2	0.002

Figure 12.8: Kruskall-Wallis one-way non-parametric ANOVA in JASP

12.8 Repeated measures one-way ANOVA

The one-way repeated measures ANOVA test is a statistical method of testing for significant differences between three or more groups where the same participants are used in each group (or each participant is closely matched with participants in other experimental groups). For this reason, there should always be an equal number of scores (data points) in each experimental group. This type of design and analysis can also be called a 'related ANOVA' or a 'within-subjects ANOVA'.

The logic behind a repeated measures ANOVA is very similar to that of an independent ANOVA (sometimes called a 'between-subjects' ANOVA). You'll remember that earlier we showed that in a between-subjects ANOVA total variability is partitioned into between-groups variability (SS_b) and within-groups variability (SS_w), and after each is divided by the respective degrees of freedom to give MS_b and MS_w (see Table 12.1) the F-ratio is calculated as:

$$F = \frac{\mathrm{MS}_b}{\mathrm{MS}_w}$$

In a repeated measures ANOVA, the F-ratio is calculated in a similar way, but whereas in an independent ANOVA the within-group variability (SS_w) is used as the basis for the MS_w denominator, in a repeated measures ANOVA the SS_w is partioned into two parts. As we are using the same subjects in each group, we can remove the variability due to the individual differences between subjects (referred to as $\mathrm{SS}_{\text{subjects}}$) from the within-groups variability. We won't go into too much technical detail about how this is done, but essentially each subject becomes a level of a factor called subjects. The variability in this within-subjects factor is then calculated in the same way as any between-subjects factor. And then we can subtract $\mathrm{SS}_{\text{subjects}}$ from SS_w to provide a smaller $\mathrm{SS}_{\text{error}}$ term:

$$\begin{aligned}\text{Independent ANOVA: } \mathrm{SS}_{\text{error}} &= \mathrm{SS}_w \\ \text{Repeated Measures ANOVA: } \mathrm{SS}_{\text{error}} &= \mathrm{SS}_w - \mathrm{SS}_{\text{subjects}}\end{aligned}$$

This change in $\mathrm{SS}_{\text{error}}$ term often leads to a more powerful statistical test, but this does depend on whether the reduction in the $\mathrm{SS}_{\text{error}}$ more than compensates for the reduction in degrees of freedom for the error term (as degrees of freedom go from $(\mathrm{n}-\mathrm{k})$[8] to $(\mathrm{n}-1)(\mathrm{k}-1)$ (remembering that there are more subjects in the independent ANOVA design).

12.8.1 Repeated measures ANOVA in JASP

First, we need some data. Geschwind (1972) has suggested that the exact nature of a patient's language deficit following a stroke can be used to diagnose the specific region of the brain that has been damaged. A researcher is concerned with identifying the specific communication difficulties experienced by six patients suffering from Broca's Aphasia (a language deficit commonly experienced following a stroke).

The patients were required to complete three word recognition tasks. On the first (speech production) task, patients were required to repeat single words read out aloud by the researcher. On the second (conceptual) task, designed to test word comprehension, patients were required to match a series of pictures with their correct name. On the third (syntax) task, designed to test knowledge

of correct word order, patients were asked to reorder syntactically incorrect sentences. Each patient completed all three tasks. The order in which patients attempted the tasks was counterbalanced between participants. Each task consisted of a series of 10 attempts. The number of attempts successfully completed by each patient are shown in Table 12.2. You'll find these data in the broca.csv file, which you can load into JASP.

Table 12.2: Number of attempts successfully completed on three experimental tasks.

Participant	Speech	Conceptual	Syntax
1	8	7	6
2	7	8	6
3	9	5	3
4	5	4	5
5	6	6	2
6	8	7	4

To perform a one-way repeated measures ANOVA in JASP, click 'ANOVA' - 'Repeated Measures ANOVA' and then perform the following (see Figure 12.9),

- -Enter a Repeated Measures Factor Name. This should be a label that you choose to describe the conditions repeated by all participants. For example, to describe the speech, conceptual and syntax tasks completed by all participants a suitable label would be 'Task'. Note that this new factor name represents the independent variable in the analysis. To accomplish this in JASP, simply click on the 'RM Factor 1' name; it will become highlighted and await its new name - simply type it!
- -Next you'll want to name each level of the 'Task' factor. Note that you'll need to add a third level in the Repeated Measures Factors text box, as there are three levels representing the three tasks: speech, conceptual, and syntax. Change the labels of the levels accordingly; simply click on each and type the new name.
- -Then move each of the variables in the left box across to the 'Repeated Measures Cells' text box. Make sure the variable names match the level names you specified earlier.
- -Finally, under the Assumption Checks option, tick the "Sphericity checks" text box (trust me for now!).

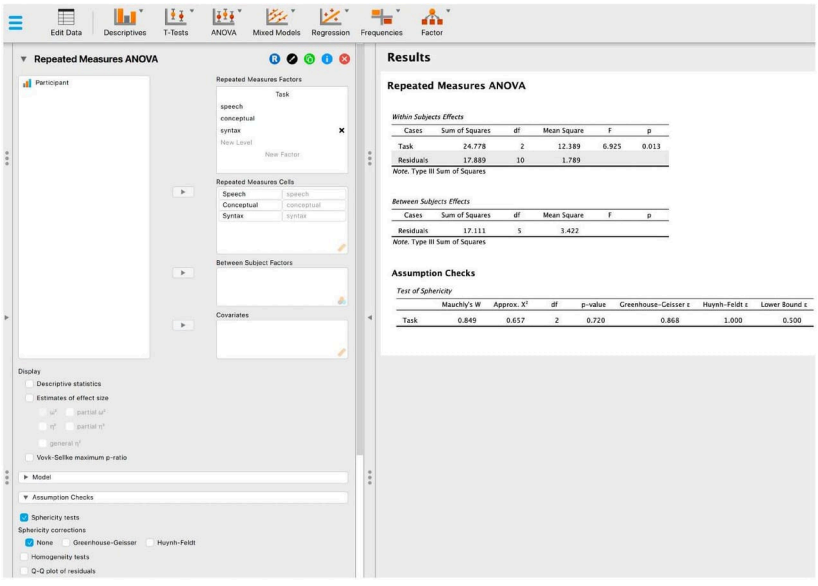


Figure 12.9: Repeated measures ANOVA in JASP

Test of Sphericity

Assumption Checks

	Mauchly's W	Approx. X^2	df	p	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
Task	0.849	0.657	2	0.720	0.868	1.000	0.500

Figure 12.10: One-way repeated measures ANOVA output: Mauchly's Test of Sphericity

The JASP output for the repeated measures ANOVA is produced as shown in Figures 12.10. Before looking at the results, we should take a look at Mauchly's Test of Sphericity, which tests the hypothesis that the variances of the differences between the conditions are equal (meaning that the spread of difference scores between the study conditions is approximately the same). In Figure 12.10, Mauchly's test significance level is $p = .720$. If Mauchly's test is non-significant (i.e. $p > .05$, as is the case in this analysis) then it is reasonable to conclude that the variances of the differences are not significantly different (i.e. they are roughly equal and sphericity can be assumed.).

A non-significant Mauchly's test means that we have not detected a violation of sphericity. It does not prove that sphericity holds, especially with a small sample. In this case, we should adjust the degrees of freedom used to calculate the p-value obtained in the one-way related ANOVA analysis:

- -If the Greenhouse-Geisser value in the "Tests of Sphericity" table is > . 75 then you should use the Huynh-Feldt correction.
- -But if the Greenhouse-Geisser value is < .75, then you should use the Greenhouse-Geisser correction.

Both corrections can be selected under Assumption Checks → Sphericity Corrections. They adjust the degrees of freedom and p-values while leaving the F statistic unchanged and the corrected F- values are then shown in the results table, as in Figure 12.11.

Within Subjects Effects

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p
Task	None	24.778	2.000	12.389	6.925	0.013
	Greenhouse-Geisser	24.778	1.737	14.265	6.925	0.018
	Huynh-Feldt	24.778	2.000	12.389	6.925	0.013
Residuals	None	17.889	10.000	1.789		
	Greenhouse-Geisser	17.889	8.685	2.060		
	Huynh-Feldt	17.889	10.000	1.789		
Note. Type III Sum of Squares						

Figure 12.11: One-way repeated measures ANOVA output: Tests of Within-Subjects Effects

Mauchly's test was not significant, p = .720, so we use the uncorrected results here. Therefore, we can use the 'None' Sphericity Correction output values for the repeated measure 'Task': $F = 6.93, df = 2, p = .013$, and we can conclude that the number of tests successfully completed on each language task did vary significantly depending on whether the task was speech, comprehension or syntax based $(F(2, 10) = 6.93, p = .013)$.

As usual, descriptive statistics should be reviewed to help interpret the results. One particularly nice way to produce them in the JASP output is to go to the 'Additional Options' menu and move 'Task' to the action box under 'Marginal Means'. This will produce the table shown in Figure 12.12, which includes not only means for each condition, but also 95% Cls for each mean. Comparison of the mean number of trials successfully completed by participants shows that Broca's Aphasics perform reasonably well on speech production (mean $= 7.17$) and language comprehension (mean $= 6.17$) tasks. However, their performance was considerably worse on the syntax task (mean $= 4.33$), with a significant difference in post-hoc tests between Speech and Syntax task performance.

Marginal Means

Marginal Means - Task		95% CI for Mean Difference		
Task	Marginal Mean	Lower	Upper	SE
speech	7.167	5.622	8.711	0.601
conceptual	6.167	4.622	7.711	0.601
syntax	4.333	2.620	6.047	0.667

Figure 12.12: Repeated measures ANOVA output: Descriptive statistics via the 'Marginal Means' dialogue.
the same participants are in each group, or each participant is closely matched with participants in other conditions. If the dependent variable is ordinal, or if the assumption of normality is not met, then the Friedman test can be used.

12.9 The Friedman non-parametric repeated measures ANOVA test

The Friedman test is a nonparametric alternative to a repeated measures ANOVA. It is used to compare three or more related conditions, where the same participants complete each condition or participants are matched across conditions. Unlike the Kruskal–Wallis test, it accounts for the relationship among observations within each participant or matched set. It can be useful with ordinal outcomes or when the normality assumption for repeated measures ANOVA is not appropriate.

The Friedman test in JASP

Friedman Test

Factor	X^(2)_(F)	df	p	Kendall's W
Task	6.636	2	0.036	0.553

Conover Test

Conover's Post Hoc Comparisons- Task

		T-Stat	df	W_(i)	W_(j)	r_(rb)	p	p_(bonf)	Pholm
speech	conceptual	1.441	10	16	12.5	0.667	0.18	0.541	0.18
	syntax	3.499	10	16	7.5	1	0.006	0.017	0.017
conceptual	syntax	2.058	10	12.5	7.5	0.857	0.067	0.2	0.133

Note: Grouped by subject

Note: Rank-biserial correlation based on individual signed-rank tests

Figure 12.13: The Friedman nonparametric repeated measures ANOVA in JASP

As with the Kruskall-Wallis test, the underlying mathematics is complicated, and won't be presented here. For the purpose of this book, it is sufficient to note that JASP calculates the tie-corrected version of the Friedman test, and in Figure 12.13 there is an example using the Broca's Aphasia data we have already looked at.

It's pretty straightforward to run a Friedman test in JASP. Just go to the 'Nonparametrics' menu and move 'Task' into the 'RM Factor' box. You should also click the 'Conover's post hoc tests' option.

The JASP results in Figure 12.13 show the chi-square statistic, degrees of freedom, p-value, and Kendall's W effect size. The Friedman test indicated a significant difference among the three tasks, $\chi^2(2) = 6.636$, $p = .036$, Kendall's $W = .553$.

Conover's post hoc comparisons showed that performance was higher on the Speech task than on the Syntax task. This difference

remained significant after either the Bonferroni or Holm correction, with an adjusted p-value of .017. The Speech–Conceptual and Conceptual–Syntax comparisons were not significant after correction.

12.10 On the relationship between ANOVA and the Student t- test

There's one last thing I want to point out before finishing. It's something that a lot of people find kind of surprising, but it's worth knowing about. An ANOVA with two groups is identical to the Student t -test. No, really. It's not just that they are similar, but they are actually equivalent in every meaningful way. I won't try to prove that this is always true, but I will show you a single concrete demonstration. Suppose that, instead of running an ANOVA on our mood.gain ~ drug model, let's instead do it using therapy as the predictor. If we run this ANOVA we get an F- statistic of $F(1,16) = 1.71$, and a p- value $= 0.21$. Since we only have two groups, I didn't actually need to resort to an ANOVA, I could have just decided to run a Student t- test. So let's see what happens when I do that: I get a t- statistic of $t(16) = -1.3068$ and a p- value $= 0.21$. Curiously, the p- values are identical. Once again we obtain a value of $p = .21$. But what about the test statistic? Having run a t- test instead of an ANOVA, we get a somewhat different answer, namely $t(16) = -1.3068$. However, there is a fairly straightforward relationship here. If we square the t- statistic then we get the F - statistic from before: \)- 1.3068^{2}=1.7077\)

12.11 Summary

There's a fair bit covered in this chapter, but there's still a lot missing. In terms of what we have discussed, the key topics were:

- The basic logic behind how ANOVA works and how to run one in JASP
- How to compute an effect size for an ANOVA

- Post-hoc analysis and corrections for multiple testing
- The assumptions made by ANOVA
- How to check the homogeneity of variance assumption and what to do if it is violated
- How to check the normality assumption and what to do if it is violated
- Repeated measures ANOVA and the non-parametric equivalent, the Friedman test

As with all of the chapters in this book, there are quite a few different sources that I've relied upon, but the one stand-out text that I've been most heavily influenced by is Sahai and Ageel (2000). It's not a good book for beginners, but it's an excellent book for more advanced readers who are interested in understanding the mathematics behind ANOVA.

---

1. [1] When all groups have the same number of observations, the experimental design is said to be "balanced". Balance isn't such a big deal for one-way ANOVA, which is the topic of this chapter. It becomes more important when you start doing more complicated ANOVAs.
2. [2] SS_w is also referred to in an independent ANOVA as the error variance, or $\mathrm{SS}_{\mathrm{error}}$
3. [3] Or , to be precise, party like "it's 1899 and we've got no friends and nothing better to do with our time than do some calculations that wouldn't have made any sense in 1899 because ANOVA didn't exist until about the 1920s".
4. [4] The JASP results are more accurate than the ones in the text above, due to rounding errors.

5. [5] If you do have some theoretical basis for wanting to investigate some comparisons but not others, it's a different story. In those circumstances you're not really running "post hoc" analyses at all, you're making "planned comparisons". I do talk about this situation later in the book, but for now I want to keep things simple.
   [6] It's worth noting in passing that not all adjustment methods try to do this. What I've described here is an approach for controlling "family wise Type I error rate". However, there are other post hoc tests that seek to control the "false discovery rate", which is a somewhat different thing.

6. [7] A Shapiro–Wilk test can supplement a Q–Q plot. A non-significant result does not establish normality, particularly with a small sample. With a large sample, the test may detect departures that are too small to materially affect the analysis. Consider the plot, sample size, and nature of any departures together.
7. [8] $(n - k)$: (number of subjects - number of groups)

This page titled 12.1: Comparing several means (one-way ANOVA) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 12: Comparing several means (one-way ANOVA) by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

13: FACTORIAL ANOVA - MAIN EFFECTS AND INTERACTIONS

13.1: Factorial ANOVA
13.2: Worked Example - Calculating Sums of Squares for a Two-Factor ANOVA
13.3: Worked Example - Completing and Reporting a Two-Factor ANOVA

This page titled 13: Factorial ANOVA - Main Effects and Interactions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

13.1: Factorial ANOVA

Over the course of the last few chapters we have done quite a lot. We have looked at statistical tests you can use when you have one nominal predictor variable with two groups (e.g. the t-test) or with three or more groups (e.g. one-way ANOVA). The chapter on regression introduced a powerful new idea, that is building statistical models with multiple continuous predictor variables used to explain a single outcome variable. For instance, a regression model could be used to predict the number of errors a student makes in a reading comprehension test based on the number of hours they studied for the test and their score on a standardised IQ test.

The goal in this chapter is to extend the idea of using multiple predictors into the ANOVA framework. For instance, suppose we were interested in using the reading comprehension test to measure student achievements in three different schools, and we suspect that girls and boys are developing at different rates (and so would be expected to have different performance on average). Each student is classified in two different ways: on the basis of their gender and on the basis of their school. What we'd like to do is analyse the reading comprehension scores in terms of both of these grouping variables. The tool for doing so is generically referred to as factorial ANOVA. However, since we have two grouping variables, we sometimes refer to the analysis as a two-way ANOVA, in contrast to the one-way ANOVA.

13.1 Factorial ANOVA 1: balanced designs, no interactions

When we discussed analysis of variance, we assumed a fairly simple experimental design. Each person is in one of several groups and we want to know whether these groups have different mean scores on some outcome variable. In this section, I'll discuss a broader class of experimental designs known as factorial designs, in which we have more than one grouping variable. I gave one example of how this kind of design might arise above. In a previous chapter, we had an example in which we were looking at the effect of different drugs on the mood.gain experienced by each person. In that chapter we did find a significant effect of drug, but at the end of the chapter we also ran an analysis to see if there was an effect of therapy. We didn't find one, but there's something a bit
worrying about trying to run two separate analyses trying to predict the same outcome. Maybe there actually is an effect of therapy on mood gain, but we couldn't find it because it was being "hidden" by the effect of drug? In other words, we're going to want to run a single analysis that includes both drug and therapy as predictors. For this analysis each person is cross-classified by the drug they were given (a factor with 3 levels) and what therapy they received (a factor with 2 levels). We refer to this as a 3 × 2 factorial design.

If we cross-tabulate drug by therapy, using the 'Frequencies' - 'Contingency Tables' analysis in JASP (see Section 9.2), we get the table shown in Figure 13.1

Contingency Tables

drug	therapy	Total	
no.therapy	CBT		
placebo	3	3	6
anxifree	3	3	6
joyzepam	3	3	6
Total	9	9	18

Figure 13.1: JASP contingency table of drug by therapy

As you can see, not only do we have participants corresponding to all possible combinations of the two factors, indicating that our design is completely crossed, it turns out that there are an equal number of people in each group. In other words, we have a balanced design. In this section I'll talk about how to analyse data from balanced designs, since this is the simplest case. The story for unbalanced designs is quite tedious, so we'll put it to one side for the moment.

13.1.1 What hypotheses are we testing?

Like one-way ANOVA, factorial ANOVA is a tool for testing certain types of hypotheses about population means. So a sensible place to start would be to be explicit about what our hypotheses actually are. However, before we can even get to that point, it's really useful to have some clean and simple notation to describe the population means. Because of the fact that observations are cross-classified in terms of two different factors, there are quite a lot of different means that one might be interested in. To see this, let's start by thinking about all the different sample means that we can calculate for this kind of design. Firstly, there's the obvious idea that we might be interested in this list of group means:

drug	therapy	mood.gain
placebo	no.therapy	0.300000
anxifree	no.therapy	0.400000
joyzepam	no.therapy	1.466667
placebo	CBT	0.600000
anxifree	CBT	1.033333
joyzepam	CBT	1.500000

Now, this output shows a list of the group means for all possible combinations of the two factors (e.g., people who received the placebo and no therapy, people who received the placebo while getting CBT, etc.). It is helpful to organise all these numbers, plus the row and column means and the overall mean, into a single table which looks like this:

	no therapy	CBT	total
placebo	0.30	0.60	0.45
anxifree	0.40	1.03	0.72
joyzepam	1.47	1.50	1.48
total	0.72	1.04	0.88

Now, each of these different means is of course a sample statistic. It's a quantity that pertains to the specific observations that we've made during our study. What we want to make inferences about are the corresponding population parameters. That is, the true means as they exist within some broader population. Those population means can also be organised into a similar table, but we'll need a little mathematical notation to do so. As usual, I'll use the symbol μ to denote a population mean. However, because there are lots of different means, I'll need to use subscripts to distinguish between them.

Here's how the notation works. Our table is defined in terms of two factors. Each row corresponds to a different level of Factor A (in this case drug), and each column corresponds to a different level of Factor B (in this case therapy). If we let R denote the number of rows in the table, and C denote the number of columns, we can refer to this as an $R \times C$ factorial ANOVA. In this case $R = 3$ and $C = 2$. We'll use lowercase letters to refer to specific rows and columns, so μ_{rc} refers to the population mean associated with the r th level of Factor A (i.e. row number r) and the c th level of Factor B (column number c). [1] So the population means are now written like this:

	no therapy	CBT	total
placebo	μ_{11}	μ_{12}	
anxifree	μ_{21}	μ_{22}	
joyzepam	μ_{31}	μ_{32}	
total			

Okay, what about the remaining entries? For instance, how should we describe the average mood gain across the entire (hypothetical) population of people who might be given Joyzepam in an experiment like this, regardless of whether they were in CBT? We use the "dot" notation to express this. In the case of Joyzepam, notice that we're talking about the mean associated with

the third row in the table. That is, we're averaging across two cell means (i.e., μ_{31} and μ_{32}). The result of this averaging is referred to as a marginal mean, and would be denoted $\mu_{3.}$ in this case. The marginal mean for CBT corresponds to the population mean associated with the second column in the table, so we use the notation $\mu_{.2}$ to describe it. The grand mean is denoted $\mu_{..}$ because it is the mean obtained by averaging (marginalising [2]) over both. So our full table of population means can be written down like this:

	no therapy	CBT	total
placebo	μ_{11}	μ_{12}	$\mu_{1.}$
anxifree	μ_{21}	μ_{22}	$\mu_{2.}$
joyzepam	μ_{31}	μ_{32}	$\mu_{3.}$
total	$\mu_{.1}$	$\mu_{.2}$	$\mu_{..}$

Now that we have this notation, it is straightforward to formulate and express some hypotheses. Let's suppose that the goal is to find out two things. First, does the choice of drug have any effect on mood? And second, does CBT have any effect on mood? These aren't the only hypotheses that we could formulate of course, and we'll see a really important example of a different kind of hypothesis in Section 13.2, but these are the two simplest hypotheses to test, and so we'll start there. Consider the first test. If the drug has no effect then we would expect all of the row means to be identical, right? So that's our null hypothesis. On the other hand, if the drug does matter then we should expect these row means to be different. Formally, we write down our null and alternative hypotheses in terms of the equality of marginal means:

$$\begin{array}{ll} \text{Null hypothesis, } H_0: & \text{row means are the same, i.e., } \mu_{1.} = \mu_{2.} = \mu_{3.} \\ \text{Alternative hypothesis, } H_1: & \text{at least one row mean is different.} \end{array}$$

It's worth noting that these are exactly the same statistical hypotheses that we formed when we ran a one-way ANOVA on these data back in a previous chapter. Back then I used the notation μ_P to refer to the mean mood gain for the placebo group, with μ_A and μ_J corresponding to the group means for the two drugs, and the null hypothesis was $\mu_P = \mu_A = \mu_J$. So we're actually talking about the same hypothesis, it's just that the more complicated ANOVA requires more careful notation due to the presence of multiple grouping variables, so we're now referring to this hypothesis as $\mu_{1.} = \mu_{2.} = \mu_{3.}$. However, as we'll see shortly, although the hypothesis is identical the test of that hypothesis is subtly different due to the fact that we're now acknowledging the existence of the second grouping variable.

Speaking of the other grouping variable, you won't be surprised to discover that our second hypothesis test is formulated the same way. However, since we're talking about the psychological therapy rather than drugs our null hypothesis now corresponds to the equality of the column means:

$$\begin{array}{ll} \text{Null hypothesis, } H_0: & \text{column means are the same, i.e., } \mu_{.1} = \mu_{.2} \\ \text{Alternative hypothesis, } H_1: & \text{column means are different, i.e., } \mu_{.1} \neq \mu_{.2} \end{array}$$

13.1.2 Running the analysis in JASP

The null and alternative hypotheses that I described in the last section should seem awfully familiar. They're basically the same as the hypotheses that we were testing in our simpler one-way ANOVAs in a previous chapter. So you're probably expecting that the hypothesis tests that are used in factorial ANOVA will be essentially the same as the F-test from the previous chapter. You're expecting to see references to sums of squares (SS), mean squares (MS), degrees of freedom (df), and finally an F-statistic that we can convert into a p-value, right? Well, you're absolutely and completely right. So much so that I'm going to depart from my usual approach. Throughout this book, I've generally taken the approach of describing the logic (and to an extent the mathematics) that underpins a particular analysis first and only then introducing the analysis in JASP. This time I'm going to do it the other way around and show you how to do it in JASP first. The reason for doing this is that I want to highlight the similarities between the simple one-way ANOVA tool that we discussed in the previous chapter, and the more complicated approach that we're going to use in this chapter.

If the data you're trying to analyse correspond to a balanced factorial design then running your analysis of variance is easy. To see how easy it is, let's start by reproducing the original analysis from the previous chapter. In case you've forgotten, for that analysis we were using only a single factor (i.e., drug) to predict our outcome variable (i.e., mood.gain), and we got the results shown in Figure 13.2.

ANOVA - mood.gain

Cases	Sum of Squares	df	Mean Square	F	p
drug	3.453	2	1.727	18.611	< . 001
Residuals	1.392	15	0.093		
Note. Type III Sum of Squares					

Figure 13.2: JASP one way ANOVA of mood.gain by drug

Now, suppose I'm also curious to find out if therapy has a relationship to mood.gain. In light of what we've seen from our discussion of multiple regression, you probably won't be surprised that all we have to do is add therapy as a second 'Fixed Factor' in the analysis, see Figure 13.3. For this first analysis, open Model and remove drug × therapy from the Model Terms box. Keep drug and therapy as main effects. We will add the interaction later.

ANOVA - mood.gain

Cases	Sum of Squares	df	Mean Square	F	p
drug	3.453	2.000	1.727	26.149	< . 001
therapy	0.467	1.000	0.467	7.076	0.019
Residual	0.924	14.000	0.066		

Figure 13.3: JASP two way ANOVA of mood.gain by drug and therapy

Note. Type III Sum of Squares

This output is pretty simple to read too. The first row of the table reports a between-group sum of squares (SS) value associated with the drug factor, along with a corresponding between-group df value. It also calculates a mean square value (MS), an F -statistic and a p-value. There is also a row corresponding to the therapy factor and a row corresponding to the residuals (i.e., the within groups variation).

Not only are all of the individual quantities pretty familiar, the relationships between these different quantities has remained unchanged, just like we saw with the original one-way ANOVA. Note that the mean square value is calculated by dividing SS by the corresponding df. That is, it's still true that

$$MS = \frac{SS}{df}$$

regardless of whether we're talking about drug, therapy or the residuals. To see this, let's not worry about how the sums of squares values are calculated. Instead, let's take it on faith that JASP has calculated the SS values correctly, and try to verify that all the rest of the numbers make sense. First, note that for the drug factor, we divide 3.453 by 2 and end up with a mean square value of 1.727. For the therapy factor, there's only 1 degree of freedom, so our calculations are even simpler: dividing 0.467 (the SS value) by 1 gives us an answer of 0.467 (the MS value).

Turning to the F statistics and the p values, notice that we have two of each; one corresponding to the drug factor and the other corresponding to the therapy factor. Regardless of which one we're talking about, the F statistic is calculated by dividing the mean square value associated with the factor by the mean square value associated with the residuals. If we use "A" as shorthand notation to refer to the first factor (factor A; in this case drug) and "R" as shorthand notation to refer to the residuals, then the F statistic associated with factor A is denoted F_A, and is calculated as follows:

$$F_A = \frac{\mathrm{MS}_A}{\mathrm{MS}_R}$$

and an equivalent formula exists for factor B (i.e., therapy). Note that this use of " R " to refer to residuals is a bit awkward, since we also used the letter R to refer to the number of rows in the table,
but I'm only going to use "R" to mean residuals in the context of SS_R and MS_R, so hopefully this shouldn't be confusing. Anyway, to

apply this formula to the drugs factor we take the mean square of 1.727 and divide it by the residual mean square value of 0.066 , which gives us an F-statistic of 26.149. The corresponding calculation for the therapy variable would be to divide 0.467 by 0.066 which gives 7.076 as the F-statistic. Not surprisingly, of course, these are the same values that JASP has reported in the ANOVA table above.

Also in the ANOVA table is the calculation of the p values. Once again, there is nothing new here. For each of our two factors what we're trying to do is test the null hypothesis that there is no relationship between the factor and the outcome variable (I'll be a bit more precise about this later on). To that end, we've (apparently) followed a similar strategy to what we did in the one way ANOVA and have calculated an F-statistic for each of these hypotheses. To convert these to p values, all we need to do is note that the sampling distribution for the F statistic under the null hypothesis (that the factor in question is irrelevant) is an F distribution. Also note that the two degrees of freedom values are those corresponding to the factor and those corresponding to the residuals. For the drug factor we're talking about an F distribution with 2 and 14 degrees of freedom (I'll discuss degrees of freedom in more detail later). In contrast, for the therapy factor the sampling distribution is F with 1 and 14 degrees of freedom.

At this point, I hope you can see that the ANOVA table for this more complicated factorial analysis should be read in much the same way as the ANOVA table for the simpler one way analysis. In short, it's telling us that the factorial ANOVA for our 3×2 design found a significant effect of drug ($F_{2,14} = 26.15, p < .001$) as well as a significant effect of therapy ($F_{1,14} = 7.08, p = .02$). Or, to use the more technically correct terminology, we would say that there are two main effects of drug and therapy. At the moment, it probably seems a bit redundant to refer to these as "main" effects, but it actually does make sense. Later on, we're going to want to talk about the possibility of "interactions" between the two factors, and so we generally make a distinction between main effects and interaction effects.

13.1.3 How are the sum of squares calculated?

In the previous section I had two goals. Firstly, to show you that the JASP method needed to do factorial ANOVA is pretty much the same as what we used for a one way ANOVA. The only difference is the addition of a second factor. Secondly, I wanted to show you what the ANOVA table looks like in this case, so that you can see from the outset that the basic logic and structure behind factorial ANOVA is the same as that which underpins one way ANOVA. Try to hold onto that feeling. It's genuinely true, insofar as factorial ANOVA is built in more or less the same way as the simpler one-way ANOVA model. It's just that this feeling of familiarity starts to evaporate once you start digging into the details. Traditionally, this comforting sensation is replaced by an urge to hurl abuse at the authors of statistics textbooks.

Okay, let's start by looking at some of those details. The explanation that I gave in the last section illustrates the fact that the hypothesis tests for the main effects (of drug and therapy in this case)
are F-tests, but what it doesn't do is show you how the sum of squares (SS) values are calculated. Nor does it tell you explicitly how to calculate degrees of freedom (df values) though that's a simple thing by comparison. Let's assume for now that we have only two predictor variables, Factor A and Factor B. If we use Y to refer to the outcome variable, then we would use Y_{rci} to refer to the outcome associated with the i-th member of group rc (i.e., level/row r for Factor A and level/column c for Factor B). Thus, if we use $\bar{Y}$ to refer to a sample mean, we can use the same notation as before to refer to group means, marginal means and grand means. That is, $\bar{Y}_{rc}$ is the sample mean associated with the r th level of Factor A and the cth level of Factor B, $\bar{Y}_{r}$.would be the marginal mean for the r th level of Factor A, $\bar{Y}_{.c}$ would be the marginal mean for the cth level of Factor B, and $\bar{Y}_{..}$ is the grand mean. In other words, our sample means can be organised into the same table as the population means. For our clinical trial data, that table looks like this:

	no therapy	CBT	total
placebo	$\bar{Y}_{11}$	$\bar{Y}_{12}$	$\bar{Y}_{1.}$
anxifree	$\bar{Y}_{21}$	$\bar{Y}_{22}$	$\bar{Y}_{2.}$
joyzepam	$\bar{Y}_{31}$	$\bar{Y}_{32}$	$\bar{Y}_{3.}$
total	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	$\bar{Y}_{..}$

And if we look at the sample means that I showed earlier, we have $\bar{Y}_{11} = 0.30, \bar{Y}_{12} = 0.60$ etc. In our clinical trial example, the drugs factor has 3 levels and the therapy factor has 2 levels, and so what we're trying to run is a 3 × 2 factorial ANOVA. However, we'll be a little more general and say that Factor A (the row factor) has R levels and Factor B (the column factor) has C levels, and so what we're runnning here is an $R \times C$ factorial ANOVA.

Now that we've got our notation straight, we can compute the sum of squares values for each of the two factors in a relatively familiar way. For Factor A, our between group sum of squares is calculated by assessing the extent to which the (row) marginal means $\bar{Y}_{1.}, \bar{Y}_{2.}$ etc, are different from the grand mean $\bar{\gamma}_{..}$. We do this in the same way that we did for one-way ANOVA: calculate the sum of squared difference between the $\bar{Y}_{i.}$ values and the $\bar{Y}_{..}$ values. Specifically, if there are N people in each group, then we calculate this

$$\mathrm{SS}_A = (N \times C) \sum_{r=1}^{R} \left(\bar{Y}_{r.} - \bar{Y}_{..} \right)^2$$

As with one-way ANOVA, the most interesting [a] part of this formula is the $\left(\bar{Y}_{r.} - \bar{Y}_{..} \right)^2$ bit, which corresponds to the squared deviation associated with level r. All that this formula does is calculate this squared deviation for all R levels of the factor, add them up, and then multiply the result by $N \times C$. The reason for this last part is that there are multiple cells in our design that have level r on Factor A. In fact, there are C of them, one corresponding to each possible level of Factor B! For
instance, in our example there are two different cells in the design corresponding to the anxifree drug: one for people with no.therapy and one for the CBT group. Not only that, within each of these cells there are N observations. So, if we want to convert our SS value into a quantity that calculates the between-groups sum of squares on a "per observation" basis, we have to multiply by $N \times C$. The formula for factor B is of course the same thing, just with some subscripts shuffled around

$$\mathrm{SS}_B = (N \times R) \sum_{c=1}^{C} \left(\bar{Y}_{.c} - \bar{Y}_{..} \right)^2$$

Now that we have these formulas we can check them against the JASP output from the earlier section. Once again, a dedicated spreadsheet programme is helpful for these sorts of calculations.

First, let's calculate the sum of squares associated with the main effect of drug. There are a total of $N = 3$ people in each group and $C = 2$ different types of therapy. Or, to put it another way, there are $3 \times 2 = 6$ people who received any particular drug. When we do these calculations in a spreadsheet programme, we get a value of 3.45 for the sum of squares associated with the main effect of drug. Not surprisingly, this is the same number that you get when you look up the SS value for the drugs factor in the ANOVA table that I presented earlier, in Figure 13.3.

We can repeat the same kind of calculation for the effect of therapy. Again there are $N = 3$ people in each group, but since there are $R = 3$ different drugs, this time around we note that there are $3 \times 3 = 9$ people who received CBT and an additional 9 people who received no therapy. So our calculation in this case gives us a value of 0.47 for the sum of squares associated with the main effect of therapy. Once again, we are not surprised to see that our calculations are identical to the ANOVA output in Figure 13.3.

So that's how you calculate the SS values for the two main effects. These SS values are analogous to the between-group sum of squares values that we calculated when doing one-way ANOVA. However, it's not a good idea to think of them as between-groups SS values anymore, just because we have two different grouping variables and it's easy to get confused. In order to construct an F test, however, we also need to calculate the within-groups sum of squares. In keeping with the terminology that we used in the regression chapter and the terminology that JASP uses when printing out the ANOVA table, I'll start referring to the within-groups SS value as the residual sum of squares SS_R.

The easiest way to think about the residual SS values in this context, I think, is to think of it as the leftover variation in the outcome variable after you take into account the differences in the marginal means (i.e., after you remove SS_A and SS_B). What I mean by that is we can start by calculating the total sum of squares, which I'll label SS_T. The formula for this is pretty much the same as it was for one-way ANOVA. We take the difference between each observation Y_{rci} and the grand mean $\bar{Y}_{...}$, square the differences, and add them all up

$$\mathrm{SS}_T = \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{i=1}^{N} \left(Y_{rci} - \bar{Y}_{..} \right)^2$$

The "triple summation" here looks more complicated than it is. In the first two summations, we're summing across all levels of Factor A (i.e., over all possible rows r in our table) and across all levels of Factor B (i.e., all possible columns c). Each rc combination corresponds to a single group and each group contains N people, so we have to sum across all those people (i.e., all i values) too. In other words, all we're doing here is summing across all observations in the data set (i.e., all possible rci combinations).

At this point, we know the total variability of the outcome variable SS_T, and we know how much of that variability can be attributed

to Factor A (SS_A) and how much of it can be attributed to Factor $\mathrm{B}\,(\mathrm{SS}_B)$. The residual sum of squares is thus defined to be the variability in Y that can't be attributed to either of our two factors. In other words

$$\mathrm{SS}_R = \mathrm{SS}_T - (\mathrm{SS}_A + \mathrm{SS}_B)$$

Of course, there is a formula that you can use to calculate the residual SS directly, but I think that it makes more conceptual sense to think of it like this. The whole point of calling it a residual is that it's the leftover variation, and the formula above makes that clear. I should also note that, in keeping with the terminology used in the regression chapter, it is commonplace to refer to $\mathrm{SS}_A + \mathrm{SS}_B$ as the variance attributable to the "ANOVA model", denoted SS_M, and so we often say that the total sum of squares is equal to the model sum of squares plus the residual sum of squares. Later on in this chapter we'll see that this isn't just a surface similarity: ANOVA and regression are actually the same thing under the hood.

In any case, it's probably worth taking a moment to check that we can calculate SS_R using this formula and verify that we do obtain the same answer that JASP produces in its ANOVA table. Again, the calculations are pretty straightforward when done in a spreadsheet. We can calculate the total SS using the formulas above (getting an answer of total SS $= 4.85$) and then the residual SS $(= 0.92)$. Yet again, we get the same answer.

13.1.4 What are our degrees of freedom?

The degrees of freedom are calculated in much the same way as for one-way ANOVA. For any given factor, the degrees of freedom is equal to the number of levels minus 1 (i.e., $R - 1$ for the row variable Factor A, and $C - 1$ for the column variable Factor B). So, for the drugs factor we obtain $df = 2$, and for the therapy factor we obtain $df = 1$. Later on, when we discuss the interpretation of ANOVA as a regression model (see Section 13.6), I'll give a clearer statement of how we arrive at this number. But for the moment we can use the simple definition of degrees of freedom, namely that the degrees of freedom equals the number of quantities that are observed, minus the number of constraints. So, for the drugs factor, we observe 3 separate group means, but these are constrained by 1 grand mean, and therefore the degrees of freedom is 2 . For the residuals, the logic is similar, but not quite the same. The total number of observations in our experiment is 18. The constraints
correspond to 1 grand mean, the 2 additional group means that the drug factor introduces, and the 1 additional group mean that the the therapy factor introduces, and so our degrees of freedom is 14. As a formula, this is $N - 1 - (R - 1) - (C - 1)$, which simplifies to $N - R - C + 1$.

13.1.5 Factorial ANOVA versus one-way ANOVAs

Now that we've seen how a factorial ANOVA works, it's worth taking a moment to compare it to the results of the one way analyses, because this will give us a really good sense of why it's a good idea to run the factorial ANOVA. In a previous chapter, I ran a one-way ANOVA that looked to see if there are any differences between drugs, and a second one-way ANOVA to see if there were any differences between therapies. As we saw in Section 13.1.1, the null and alternative hypotheses tested by the one-way ANOVAs are in fact identical to the hypotheses tested by the factorial ANOVA. Looking even more carefully at the ANOVA tables, we can see that the sum of squares associated with the factors are identical in the two different analyses (3.45 for drug and 0.47 for therapy), as are the degrees of freedom (2 for drug, 1 for therapy). But they don't give the same answers! Most notably, when we ran the one-way ANOVA for therapy in Section 12.10 we didn't find a significant effect (the p-value was .21). However, when we look at the main effect of therapy within the context of the two-way ANOVA, we do get a significant effect $(p = .019)$. The two analyses are clearly not the same.

Why does that happen? The answer lies in understanding how the residuals are calculated. Recall that the whole idea behind an F-test is to compare the variability that can be attributed to a particular factor with the variability that cannot be accounted for (the residuals). If you run a one-way ANOVA for therapy, and therefore ignore the effect of drug, the ANOVA will end up dumping all of the drug-induced variability into the residuals! This has the effect of making the data look more noisy than they really are, and the effect of therapy which is correctly found to be significant in the two-way ANOVA now becomes non-significant. If we ignore something that actually matters (e.g., drug) when trying to assess the contribution of something else (e.g., therapy) then our analysis will be distorted. Of course, it's perfectly okay to ignore variables that are genuinely irrelevant to the phenomenon of interest. If we had recorded the colour of the walls, and that turned out to be a non-significant factor in a three-way ANOVA, it would be perfectly okay to disregard it and just report the simpler two-way ANOVA that doesn't include this irrelevant factor. What you shouldn't do is drop variables that actually make a difference!

13.1.6 What kinds of outcomes does this analysis capture?

The ANOVA model that we've been talking about so far covers a range of different patterns that we might observe in our data. For instance, in a two-way ANOVA design there are four possibilities: (a) only Factor A matters, (b) only Factor B matters, (c) both A and B matter, and (d) neither A nor B matters. An example of each of these four possibilities is plotted in Figure 13.4.

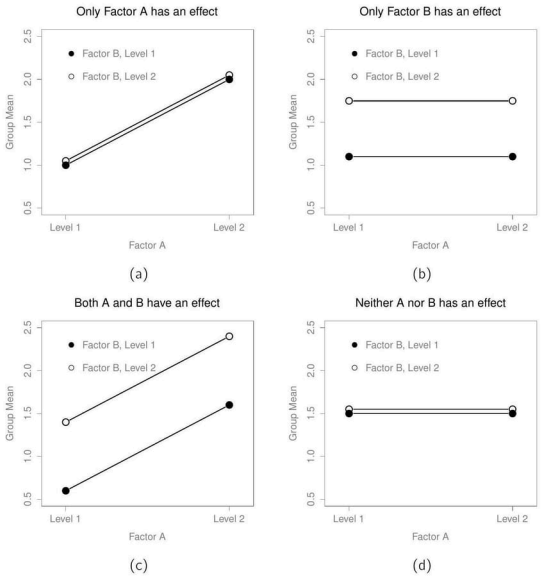


Figure 13.4: The four different outcomes for a 2 × 2 ANOVA when no interactions are present. In panel (a) we see a main effect of Factor A and no effect of Factor B. Panel (b) shows a main effect of Factor B but no effect of Factor A. Panel (c) shows main effects of both Factor A and Factor B. Finally, panel (d) shows no effect of either factor.

13.2 Factorial ANOVA 2: balanced designs, interactions allowed

The four patterns of data shown in Figure 13.4 are all quite realistic. There are a great many data sets that produce exactly those patterns. However, they are not the whole story and the ANOVA model that we have been talking about up to this point is not sufficient to fully account for a table of group means. Why not? Well, so far we have the ability to talk about the idea that drugs can influence mood, and therapy can influence mood, but no way of talking about the possibility of an interaction between the two. An interaction between A and B is said to occur whenever the effect of Factor A is different, depending on which level of Factor B we're talking about. Several examples of an interaction effect with the context of a 2×2 ANOVA are shown in Figure 13.5. To give a more concrete example, suppose that the operation of Anxifree and Joyzepam is governed by quite different physiological mechanisms. One consequence of this is that while Joyzepam has more or less the same effect on mood regardless of whether one is in therapy, Anxifree is actually much more effective when administered in conjunction with CBT. The ANOVA that we developed in the previous section does not capture this idea. To get some idea of whether an interaction is actually happening here, it helps to plot the various group means. In JASP this is done via the ANOVA 'Descriptives Plots' option - just move drug across into the 'Horizontal axis' box, and move therapy across into the 'Separate Lines' box. This should look something like Figure 13.6. Our main concern relates to the fact that the two lines aren't parallel. The effect of CBT (difference between line with solid dots and line with open dots) when the drug is Joyzepam (middle) appears to be near zero, even smaller than the effect of CBT when a placebo is used (right side). However, when Anxifree is administered, the effect of CBT is larger than the placebo (left side). Is this effect real, or is this just random variation due to chance? Our original ANOVA cannot answer this question, because we make no allowances for the idea that interactions even exist! In this section, we'll fix this problem.

13.2.1 What exactly is an interaction effect?

The key idea that we're going to introduce in this section is that of an interaction effect. In the ANOVA model we have looked at so far there are only two factors involved in our model (i.e., drug and therapy). But when we add an interaction we add a new component to the model: the combination of drug and therapy. Intuitively, the idea behind an interaction effect is fairly simple. It just means that the effect of Factor A is different, depending on which level of Factor B we're talking about. But what does that actually mean in terms of our data? The plot in Figure 13.5 depicts several different patterns that, although quite different to each

other, would all count as an interaction effect. So it's not entirely straightforward to translate this qualitative idea into something mathematical that a statistician can work with.

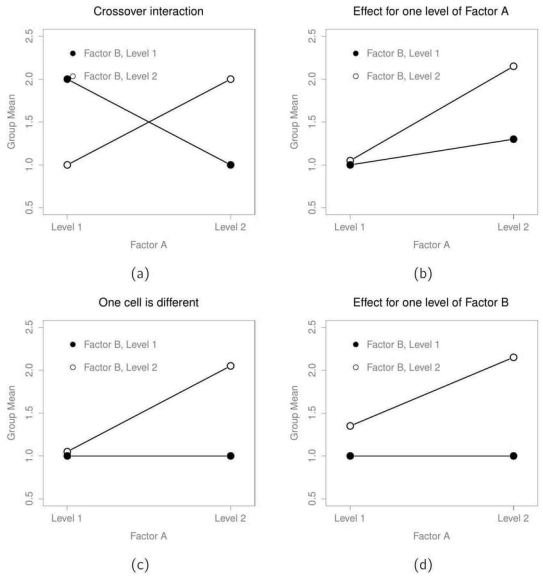


Figure 13.5: Qualitatively different interactions for a 2 × 2 ANOVA

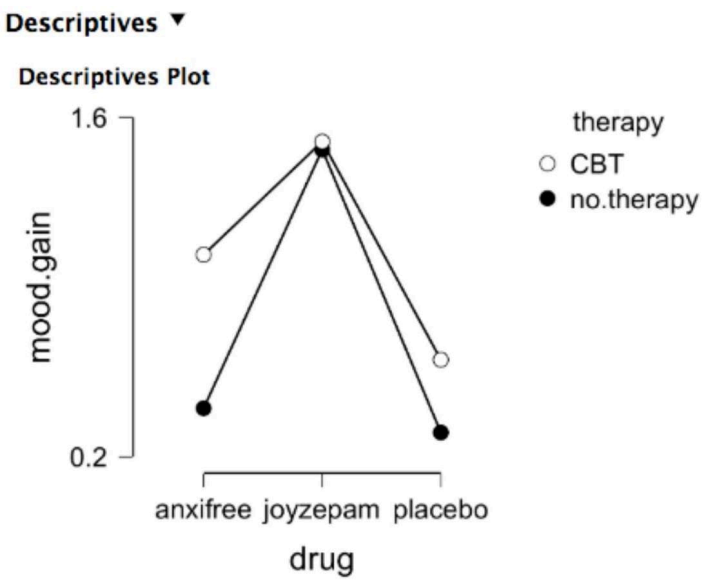


Figure 13.6: JASP screen showing a 'Descriptives Plot' in ANOVA using the clinical trial data

As a consequence, the way that the idea of an interaction effect is formalised in terms of null and alternative hypotheses is slightly difficult, and I'm guessing that a lot of readers of this book probably won't be all that interested. Even so, I'll try to give the basic idea here.

To start with, we need to be a little more explicit about our main effects. Consider the main effect of Factor A (drug in our running example). We originally formulated this in terms of the null hypothesis that the marginal means $\mu_{r.}$ are all equal to each other. Obviously, if all of these are equal to each other, then they must also be equal to the grand mean $\mu_{..}$ as well, right? So what we can do is define the effect of Factor A at level r to be equal to the difference between the marginal mean $\mu_{r.}$ and the grand mean $\mu_{...}$. Let's denote this effect by α_r, and note that

$$\alpha_r = \mu_{r.} - \mu_{..}$$

Now, by definition all of the α_r values must sum to zero, for the same reason that the average of the marginal means μ_r. must be the grand mean $\mu_{..}$. We can similarly define the effect of Factor B at level i to be the difference between the column marginal mean $\mu_{.c}$ and the grand mean $\mu_{..}$

$$\beta_c = \mu_{.c} - \mu_{..}$$

and once again, these β_c values must sum to zero. The reason that statisticians sometimes like to talk about the main effects in terms of these α_r and β_c values is that it allows them to be precise about what it means to say that there is no interaction effect. If there is no interaction at all, then these α_r and β_c values will perfectly describe the group means μ_{rc}. Specifically, it means that

$$\mu_{rc} = \mu_{..} + \alpha_r + \beta_c$$

That is, there's nothing special about the group means that you couldn't predict perfectly by knowing all the marginal means. And that's our null hypothesis, right there. The alternative hypothesis is that

$$\mu_{rc} \neq \mu_{..} + \alpha_r + \beta_c$$

for at least one group rc in our table. However, statisticians often like to write this slightly differently. They'll usually define the specific interaction associated with group rc to be some number, awkwardly referred to as $(\alpha\beta)_{rc}$, and then they will say that the alternative hypothesis is that

$$\mu_{rc} = \mu_{...} + \alpha_r + \beta_c + (\alpha\beta)_{rc}$$

where $(\alpha\beta)_{rc}$ is non-zero for at least one group. This notation is kind of ugly to look at, but it is handy as we'll see in the next section when discussing how to calculate the sum of squares.

13.2.2 Calculating sums of squares for the interaction

How should we calculate the sum of squares for the interaction terms, $\mathrm{SS}_{A:B}$? Well, first off, it helps to notice how the previous section defined the interaction effect in terms of the extent to which the actual group means differ from what you'd expect by just looking at the marginal means. Of course, all of those formulas refer to population parameters rather than sample statistics, so we don't actually know what they are. However, we can estimate them by using sample means in place of population means. So for Factor A, a good way to estimate the main effect at level r is as the difference between the sample marginal mean $\bar{Y}_{rc}$ and the sample grand mean $\bar{Y}_{..}$. That is, we would use this as our estimate of the effect

$$\hat{\alpha}_r = \bar{Y}_{r.} - \bar{Y}_{..}$$

Similarly, our estimate of the main effect of Factor B at level c can be defined as follows

$$\hat{\beta}_c = \bar{Y}_{.c} - \bar{Y}_{..}$$

Now, if you go back to the formulas that I used to describe the SS values for the two main effects, you'll notice that these effect terms are exactly the quantities that we were squaring and summing! So, what's the analog of this for interaction terms? The answer to this can be found by first rearranging the formula for the group means μ_{rc} under the alternative hypothesis, so that we get
this

$$\begin{aligned}(\alpha\beta)_{rc} &= \mu_{rc} - \mu_{..} - \alpha_r - \beta_c \\ &= \mu_{rc} - \mu_{..} - (\mu_{r.} - \mu_{..}) - (\mu_{.c} - \mu_{..}) \\ &= \mu_{rc} - \mu_{r.} - \mu_{.c} + \mu_{..}\end{aligned}$$

So, once again if we substitute our sample statistics in place of the population means, we get the following as our estimate of the interaction effect for group rc, which is

$$\hat{(\alpha\beta)}_{rc} = \bar{Y}_{rc} - \bar{Y}_{r.} - \bar{Y}_{.c} + \bar{Y}_{..}$$

Now we square each estimated interaction effect, sum across all combinations of Factor A and Factor B, and multiply by the number of observations in each cell. This gives us the following formula for the interaction sum of squares:

$$\text{SS}_{A:B} = N \sum_{r=1}^{R} \sum_{c=1}^{C} \left(\bar{Y}_{rc} - \bar{Y}_{r.} - \bar{Y}_{.c} + \bar{Y}_{..} \right)^2$$

where we multiply by N because there are N observations in each of the groups, and we want our SS values to reflect the variation among observations accounted for by the interaction, not the variation among groups.

Now that we have a formula for calculating $\text{SS}_{A:B}$, it's important to recognise that the interaction term is part of the model (of course), so the total sum of squares associated with the model, SS_M, is now equal to the sum of the three relevant SS values, $\text{SS}_A + \text{SS}_B + \text{SS}_{A:B}$. The residual sum of squares SS_R is still defined as the leftover variation, namely $\text{SS}_T - \text{SS}_M$, but now that we have the interaction term this becomes

$$\text{SS}_R = \text{SS}_T - (\text{SS}_A + \text{SS}_B + \text{SS}_{A:B})$$

As a consequence, the residual sum of squares SS_R will be smaller than in our original ANOVA that didn't include interactions.

13.2.3 Degrees of freedom for the interaction

Calculating the degrees of freedom for the interaction is, once again, slightly trickier than the corresponding calculation for the main effects. To start with, let's think about the ANOVA model as a whole. Once we include interaction effects in the model we're allowing every single group to have a unique mean, μ_{rc}. For an $R \times C$ factorial ANOVA, this means that there are $R \times C$ quantities of interest in the model and only the one constraint: all of the group means need to average out to the grand mean. So the model as a whole needs to have $(R \times C) - 1$ degrees of freedom. But the main effect of Factor A has $R - 1$ degrees of freedom, and the main effect of Factor B has $C - 1$ degrees
of freedom. This means that the degrees of freedom associated with the interaction is

$$\begin{aligned} df_{A:B} &= (R \times C - 1) - (R - 1) - (C - 1) \\ &= RC - R - C + 1 \\ &= (R - 1)(C - 1) \end{aligned}$$

which is just the product of the degrees of freedom associated with the row factor and the column factor.

What about the residual degrees of freedom? Because we've added interaction terms which absorb some degrees of freedom, there are fewer residual degrees of freedom left over. Specifically, note that if the model with interaction has a total of $(R \times C) - 1$, and there are N observations in your data set that are constrained to satisfy 1 grand mean, your residual degrees of freedom now become $N - (R \times C) - 1 + 1$, or just $N - (R \times C)$.

13.2.4 Running the ANOVA in JASP

Adding interaction terms to the ANOVA model in JASP is straightforward. In fact it is more than straightforward because it is the default option for ANOVA. This means that when you specify an ANOVA with two factors, e.g. drug and therapy then the interaction component - drug*therapy - is added automatically to the model [3]. When we run the ANOVA with the interaction term included, then we get the results shown in Figure 13.7.

ANOVA - mood.gain

Cases	Sum of Squares	df	Mean Square	F	p
drug	3.453	2	1.727	31.714	< . 001
therapy	0.467	1	0.467	8.582	0.013
drug * therapy	0.271	2	0.136	2.490	0.125
Residuals	0.653	12	0.054		
Note. Type III Sum of Squares					

Figure 13.7: Results for the full factorial model, including the interaction component drug*therapy

As it turns out, while we do have a significant main effect of drug $(F_{2,12} = 31.7, p < .001)$ and therapy type $(F_{1,12} = 8.6, p = .013)$, there is no significant interaction between the two $(F_{2,12} = 2.5, p = 0.125)$.

13.2.5 Interpreting the results

There's a couple of very important things to consider when interpreting the results of factorial ANOVA. First, there's the same issue that we had with one-way ANOVA, which is that if you obtain a significant main effect of (say) drug, it doesn't tell you anything about which drugs are different to one another. To find that out, you need to run additional analyses. We'll talk about some analyses that you can run in Sections 13.7 and 13.8. The same is true for interaction effects. Knowing that there's a significant interaction doesn't tell you anything about what kind of interaction exists. Again, you'll need to run additional analyses.

Secondly, there's a very peculiar interpretation issue that arises when you obtain a significant interaction effect but no corresponding main effect. This happens sometimes. For instance, in the crossover interaction shown in Figure 13.5a, this is exactly what you'd find. In this case, neither of the main effects would be significant, but the interaction effect would be. This is a difficult situation to interpret, and people often get a bit confused about it. The general advice that statisticians like to give in this situation is that you shouldn't pay much attention to the main effects when an interaction is present. The reason they say this is that, although the tests of the main effects are perfectly valid from a mathematical point of view, when there is a significant interaction effect the main effects rarely test interesting hypotheses. Recall from Section 13.1.1 that the null hypothesis for a main effect is that the marginal means are equal to each other, and that a marginal mean is formed by averaging across several different groups. But if you have a significant interaction effect then you know that the groups that comprise the marginal mean aren't homogeneous, so it's not really obvious why you would even care about those marginal means.

Here's what I mean. Again, let's stick with a clinical example. Suppose that we had a 2 × 2 design comparing two different treatments for phobias (e.g., systematic desensitisation vs flooding), and two different anxiety reducing drugs (e.g., Anxifree vs Joyzepam). Now, suppose what we found was that Anxifree had no effect when desensitisation was the treatment, and Joyzepam had no effect when flooding was the treatment. But both were pretty effective for the other treatment. This is a classic crossover interaction, and what we'd find when running the ANOVA is that there is no main effect of drug, but a significant interaction. Now, what does it actually mean to say that there's no main effect? Well, it means that if we average over the two different psychological treatments, then the average effect of Anxifree and Joyzepam is the same. But why would anyone care about that? When treating someone for phobias it is never the case that a person can be treated using an "average" of flooding and desensitisation. That doesn't make a lot of sense. You either get one or the other. For one treatment one drug is effective, and for the other treatment the other drug is effective. The interaction is the important thing and the main effect is kind of irrelevant.

This sort of thing happens a lot. The main effect are tests of marginal means, and when an interaction is present we often find ourselves not being terribly interested in marginal means because they imply averaging over things that the interaction tells us shouldn't be averaged! Of course, it's not always the case that a main effect is meaningless when an interaction is present. Often you can get a big main effect and a very small interaction, in which case you can still say things like "drug A is generally more effective than drug B" (because there was a big effect of drug), but you'd need to
modify it a bit by adding that "the difference in effectiveness was different for different psychological treatments". In any case, the main point here is that whenever you get a significant interaction you should stop and think about what the main effect actually means in this context. Don't automatically assume that the main effect is interesting.

13.3 Effect size

The effect size calculation for a factorial ANOVA is pretty similar to those used in one way ANOVA (see Section 12.4). Specifically, we can use η^2 (eta-squared) as a simple way to measure how big the overall effect is for any particular term. As before, η^2 is defined by dividing the sum of squares associated with that term by the total sum of squares. For instance, to determine the size of the main effect of Factor A, we would use the following formula:

$$\eta_A^2 = \frac{\text{SS}_A}{\text{SS}_T}$$

As before, this can be interpreted in much the same way as R^2 in regression. [4] It tells you the proportion of variance in the outcome variable that can be accounted for by the main effect of Factor A. It is therefore a number that ranges from 0 (no effect at all) to 1 (accounts for all of the variability in the outcome). Moreover, the sum of all the η^2 values, taken across all the terms in the model, will sum to the the total R^2 for the ANOVA model. If, for instance, the ANOVA model fits perfectly (i.e., there is no within-groups variability at all!), the η^2 values will sum to 1. Of course, that rarely if ever happens in real life.

However, when doing a factorial ANOVA, there is a second measure of effect size that people like to report, known as partial η^2. The idea behind partial η^2 (which is sometimes denoted ${}_p\eta^2$ or η_p^2) is that, when measuring the effect size for a particular term (say, the main effect of Factor A), you want to deliberately ignore the other effects in the model (e.g., the main effect of Factor B).

That is, you would pretend that the effect of all these other terms is zero, and then calculate what the η^2 value would have been. This is actually pretty easy to calculate. All you have to do is remove the sum of squares associated with the other terms from the denominator. In other words, if you want the partial η^2 for the main effect of Factor A, the denominator is just the sum of the SS values for Factor A and the residuals

$$\text{partial } \eta^2_A = \frac{\text{SS}_A}{\text{SS}_A + \text{SS}_R}$$

Partial η² is at least as large as η² because its denominator excludes the sums of squares for the other effects. And once again you get a number between 0 and 1 , where 0 represents no effect. However, it's slightly trickier to interpret what a large partial η^2 value means. Partial eta squared describes an effect relative to that effect plus the residual variation. Because the other effects are excluded from its denominator, partial eta-squared values do not generally add up to the model's R². Comparisons across different studies or designs require care because the residual variation and included factors may differ.

To see what I mean by this, it's useful to see a concrete example. First, let's have a look at the effect sizes for the original ANOVA without the interaction term, from Figure 13.3:

	η²	Partial η²
drug	0.71	0.79
therapy	0.10	0.34

Looking at the η^2 values first, we see that drug accounts for 71% of the variance (i.e. $\eta^2 = 0.71$) in mood.gain, whereas therapy only accounts for 10%. This leaves a total of 19% of the variation unaccounted for (i.e., the residuals constitute 19% of the variation in the outcome). Overall, this implies that we have a very large effect [5] of drug and a modest effect of therapy.

Now let's look at the partial η^2 values, shown in Figure 13.3. Because the effect of therapy isn't all that large, controlling for it doesn't make much of a difference, so the partial η^2 for drug doesn't increase very much, and we obtain a value of ${}_p\eta^2 = 0.79$). In contrast, because the effect of drug was very large, controlling for it makes a big difference, and so when we calculate the partial η^2 for therapy you can see that it rises to ${}_p\eta^2 = 0.34$. For therapy, partial η² = .34 means that therapy accounts for approximately 34% of the sum of the therapy and residual sums of squares. It does not mean that therapy accounts for 34% of the total variation in mood gain, nor does it automatically predict the effect size in a different experiment.

Speaking of interaction effects, here's what we get when we calculate the effect sizes for the model that includes the interaction term, as in Figure 13.7. To get these in JASP, simply select 'Additional Options' - 'Estimates of effect size', and choose the measures you want. As you can see, the η^2 values for the main effects don't change, but the partial η^2 values do:

	η²	Partial η²
drug	0.71	0.84
therapy	0.10	0.42
drug*therapy	0.06	0.29

13.3.1 Estimated group means

In many situations you will find yourself wanting to report estimates of all the group means based on the results of your ANOVA, as well as confidence intervals associated with them. You can use the 'Additional Options' - 'Marginal Means' section in the JASP ANOVA analysis to do this, as in Figure 13.8. Just move the interaction term drug*therapy into the action box. If the ANOVA that you have run is a full factorial model (i.e., contains all possible main effects and all possible interaction effects) then the estimates of the group means are actually identical to the sample means, but importantly, the confidence intervals will use a pooled estimate of the standard errors rather than use a separate one for each group.

*Marginal Means Marginal Means - drug * therapy*

drug	therapy	Marginal Mean	95% CI for Mean	SE
Lower	Upper			

drug	therapy	Marginal Mean	95% CI for Mean		SE
placebo	no.therapy	0.300	0.006	0.594	0.135
anxifree		0.400	0.106	0.694	0.135
joyzepam		1.467	1.173	1.760	0.135
placebo	CBT	0.600	0.306	0.894	0.135
anxifree		1.033	0.740	1.327	0.135
joyzepam		1.500	1.206	1.794	0.135

Figure 13.8: JASP screenshot showing the marginal means for the full factorial model, i.e. including the interaction component, with the clinical trial data set

In the output we see that the estimated mean mood gain for the placebo group with no therapy was 0.300, with a 95% confidence interval from 0.006 to 0.594. Note that these are NOT the same confidence intervals that you would get if you calculated them separately for each group, because of the fact that the ANOVA model assumes homogeneity of variance and therefore uses a pooled estimate of the standard deviation.

13.4 Assumption checking

As with one-way ANOVA, the key assumptions of factorial ANOVA are homogeneity of variance (all groups have the same standard deviation), normality of the residuals, and independence of the observations. The first two are things we can check for. The third is something that you need to assess yourself by asking if there are any special relationships between different observations, for example repeated measures where the independent variable is time so there is a relationship between the observations at time one and time two: observations at different time points are from the same people. Additionally, if you aren't using a saturated model (e.g., if you've omitted the interaction terms) then you're also assuming that the omitted terms aren't important. Of course, you can check this last one by running an ANOVA with the omitted terms included and see if they're significant, so that's pretty easy. What about homogeneity of variance and normality of the residuals? As it turns out, these are pretty easy to check. It's no different to the checks we did for a one-way ANOVA.

13.4.1 Homogeneity of variance

As mentioned in Section 12.6.1 it's a good idea to visually inspect a plot of the standard deviations compared across different groups / categories, and also see if the Levene test is consistent with the visual inspection. The theory behind the Levene test was discussed in Section 12.6.1, so I won't discuss it again. This test expects that you have a saturated model (i.e., including all of the relevant terms), because the test is primarily concerned with the within-group variance, and it doesn't really make a lot of sense to calculate this any way other than with respect to the full model. The Levene test can be specified under the ANOVA 'Assumption Checks' - 'Homogeneity tests' option in JASP, with the result shown as in Figure 13.9. A non-significant Levene test provides no evidence against homogeneity of variance, but does not establish that the variances are equal. Consider it alongside the group spreads and sample sizes.

13.4.2 Normality of residuals

As with one-way ANOVA we can test for the normality of residuals in a straightforward fashion (see Section 12.6.4). Primarily though, it's generally a good idea to examine the residuals graphically using a QQ plot. See Figure 13.9.

Assumption Checks

Test for Homogeneity of Variances (Levene's)

F	df1	df2	p
0.21	5	12	0.95384

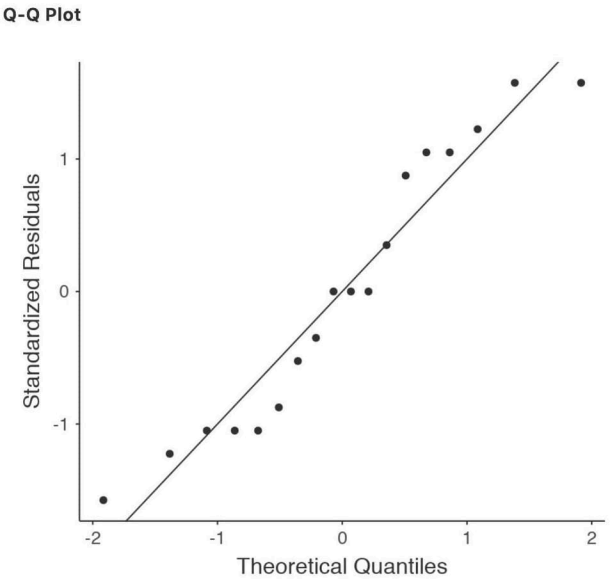


Figure 13.9: Checking assumptions in an ANOVA model

13.5 Analysis of Covariance (ANCOVA)

A variation in ANOVA is when you have an additional continuous variable that you think might be related to the dependent variable. This additional variable can be added to the analysis as a covariate, in the aptly named analysis of covariance (ANCOVA).

In ANCOVA the values of the dependent variable are "adjusted" for the influence of the covariate,
and then the "adjusted" score means are tested between groups in the usual way. This technique can increase the precision of an experiment, and therefore provide a more "powerful" test of the equality of group means in the dependent variable. How does ANCOVA do this? Well, although the covariate itself is typically not of any experimental interest, adjustment for the covariate can decrease the estimate of experimental error and thus, by reducing error variance, precision is increased. This means that an inappropriate failure to reject the null hypothesis (false negative or type II error) is less likely.

Despite this advantage, ANCOVA runs the risk of undoing real differences between groups, and this should be avoided. Look at Figure 13.10, for example, which shows a plot of Statistics anxiety against age and shows two distinct groups - students who have either an Arts or Science background or preference. ANCOVA with age as a covariate might lead to the conclusion that statistics anxiety does not differ in the two groups. Would this conclusion be reasonable - probably not because the ages of the two groups do not overlap and analysis of variance has essentially "extrapolated into a region with no data" (Everitt 1996, p. 68).

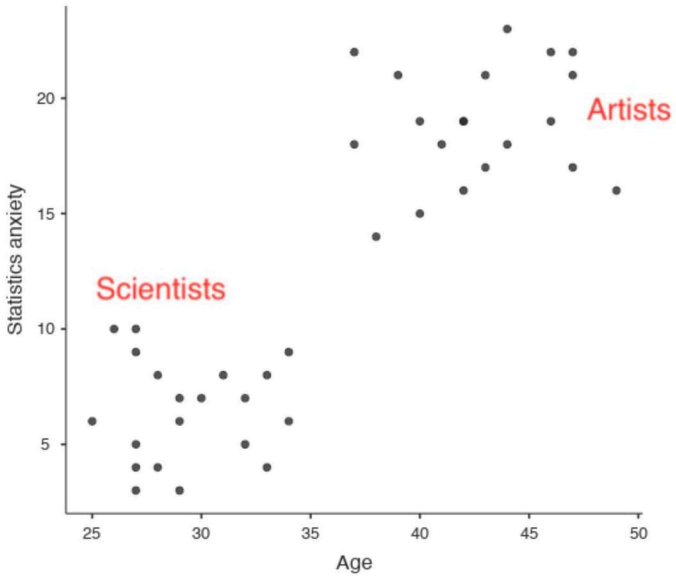


Figure 13.10: Plot of Statistics anxiety against age for two distinct groups

Clearly, careful thought needs to be given to an analysis of covariance with distinct groups. This applies to both one-way and factorial designs, as ANCOVA can be used with both.

13.5.1 Running ANCOVA in JASP

A health psychologist was interested in the effect of routine cycling and stress on happiness levels, with age as a covariate. You can find the dataset in the file ancova.csv. Open this file in JASP and then, to undertake an ANCOVA, select 'ANOVA' - 'ANCOVA' to open the ANCOVA analysis window (Figure 13.11). Highlight the dependent variable 'happiness' and transfer it into the 'Dependent Variable' text box. Highlight the independent variables 'stress' and 'commute' and transfer them into the 'Fixed Factors' text box. Highlight the covariate 'age' and transfer it into the 'Covariates' text box. Then click on 'Additional Options' - 'Marginal Means' and transfer the interaction term stress*commute into the action box.

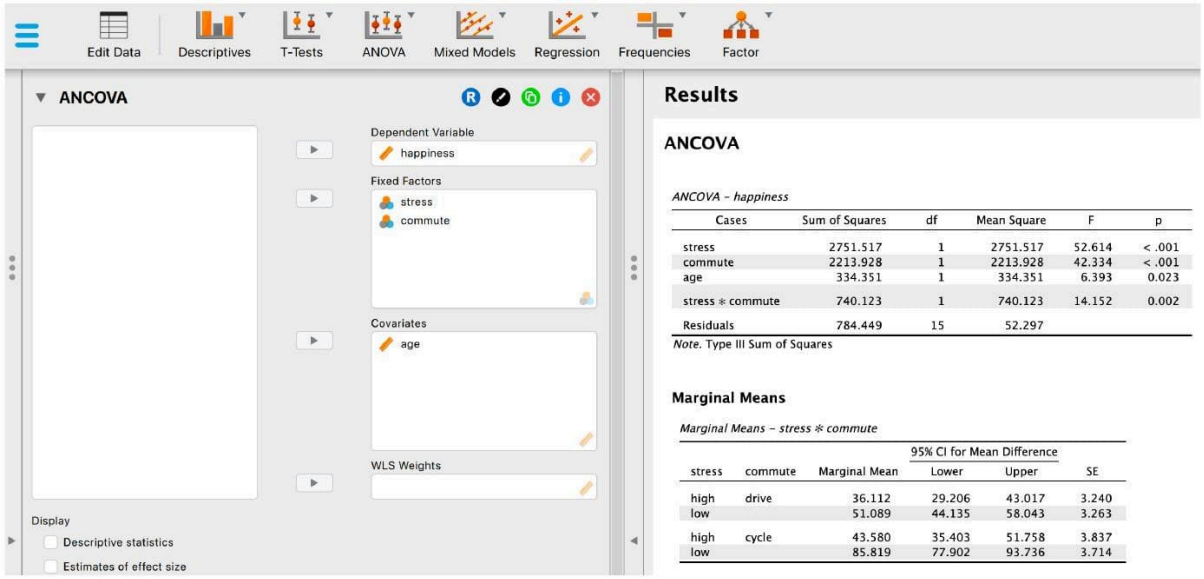


Figure 13.11: The JASP ANCOVA analysis window

An ANCOVA table is produced in the JASP results window (Figure 13.11). The F value for the covariate 'age' is significant at $p = .023$, suggesting that age is an important predictor of the dependent variable happiness. When we look at the estimated marginal mean

scores, adjustments have been made (compared to an analysis without the covariate) because of the inclusion of the covariate 'age' in this ANCOVA. A plot (Figure 13.12) is a good way of visualising and interpreting the significant effects.

The F value for the main effect 'stress' (52.61) has an associated probability of $p < .001$. The F value for the main effect 'commute' (42.33) has an associated probability of $p < .001$. Since both
of these are less than the probability that is typically used to decide if a statistical result is significant ($p < .05$) we can conclude that there was a significant main effect of stress ($F(1, 15) = 52.61, p < .001$) and a significant main effect of commuting method $(F(1, 15) = 42.33, p < .001)$. A significant interaction between stress and commuting method was also found $(F(1, 15) = 14.15, p = .002)$.

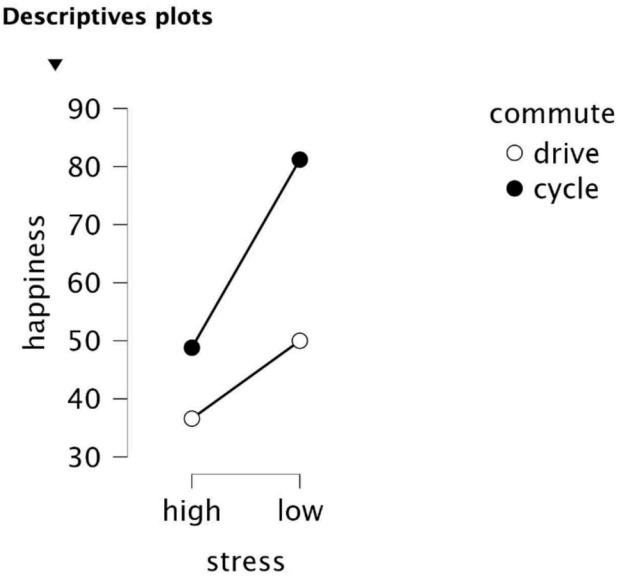


Figure 13.12: Plot of mean happiness level as a function of stress and commuting method

In Figure 13.12 we can see the adjusted, marginal, mean happiness scores when age is a covariate in an ANCOVA. The plot shows the highest adjusted mean happiness in the low-stress cycling group. The significant interaction indicates that the association between commuting method and happiness differs by stress level. Specific claims about which groups differ significantly require follow-up comparisons.

13.6 ANOVA as a linear model

One of the most important things to understand about ANOVA and regression is that they're basically the same thing. On the surface of it, you maybe wouldn't think this is true. After all, the way that I've described them so far suggests that ANOVA is primarily concerned with testing for group differences, and regression is primarily concerned with understanding the correlations between variables. And, as far as it goes that's perfectly true. But when you look under the hood, so to speak, the underlying mechanics of ANOVA and regression are awfully similar. In fact, if you think about it, you've already seen evidence of this. ANOVA and regression both rely heavily on sums of squares (SS), both make use of F tests, and so on.

The reason for this is that ANOVA and regression are both kinds of linear models. In the case of regression, this is kind of obvious. The regression equation that we use to define the relationship between predictors and outcomes is the equation for a straight line, so it's quite obviously a linear model, with the equation

$$Y_p = b_0 + b_1 X_{1p} + b_2 X_{2p} + \varepsilon_p$$

where Y_p is the outcome value for the p-th observation (e.g., p-th person), X_{1p} is the value of the first predictor for the p -th observation, X_{2p} is the value of the second predictor for the p-th observation, the b_0, b_1, and b_2 terms are our regression coefficients, and ε_p is the p-th residual. If we ignore the residuals ε_p and just focus on the regression line itself, we get the following formula:

$$\hat{Y}_p = b_0 + b_1 X_{1p} + b_2 X_{2p}$$

where $\hat{Y}_p$ is the value of Y that the regression line predicts for person p, as opposed to the actually-observed value Y_p. The thing that isn't immediately obvious is that we can write ANOVA as a linear model as well. However, it's actually pretty straightforward to do this. Let's start with a really simple example, rewriting a 2 × 2 factorial ANOVA as a linear model.

13.6.1 Some data

To make things concrete, let's suppose that our outcome variable is the grade that a student receives in my class, a ratio-scale variable corresponding to a mark from 0% to 100%. There are two predictor variables of interest: whether or not the student turned up to lectures (the attend variable) and whether or not the student actually read the textbook (the reading variable). We'll say that attend $= 1$ if the student attended class, and attend $= 0$ if they did not. Similarly, we'll say that reading $= 1$ if the student read the textbook, and reading $= 0$ if they did not.

Okay, so far that's simple enough. The next thing we need to do is to wrap some maths around this (sorry!). For the purposes of this example, let Y_p denote the grade of the p-th student in the
class. This is not quite the same notation that we used earlier in this chapter. Previously, we've used the notation Y_{rci} to refer to the i-th person in the r-th group for predictor 1 (the row factor) and the c-th group for predictor 2 (the column factor). This extended notation was really handy for describing how the SS values are calculated, but it's a pain in the current context, so I'll switch notation here. Now, the Y_p notation is visually simpler than Y_{rci}, but it has the shortcoming that it doesn't actually keep track of the group memberships! That is, if I told you that $Y_{0,0,3} = 35$, you'd immediately know that we're talking about a student (the 3rd such student, in fact) who didn't attend the lectures (i.e., attend $= 0$) and didn't read the textbook (i.e. reading $= 0$), and who ended up failing the class (grade $= 35$). But if I tell you that $Y_p = 35$, all you know is that the p-th student didn't get a good grade. We've lost some key information here. Of course, it doesn't take a lot of thought to figure out how to fix this. What we'll do instead is introduce two new variables X_{1p} and X_{2p} that keep track of this information. In the case of our hypothetical student, we know that $X_{1p} = 0$ (i.e., attend $= 0$) and $X_{2p} = 0$ (i.e., reading $= 0$). So the data might look like this:

person, p	grade, Y_p	attendance, X_{1p}	reading, X_{2p}
1	90	1	1
2	87	1	1
3	75	0	1
4	60	1	0
5	35	0	0
6	50	0	0
7	65	1	0
8	70	0	1

This isn't anything particularly special, of course. It's exactly the format in which we expect to see our data! See the data file rtfm.csv.

13.6.2 ANOVA with binary factors as a regression model

Okay, let's get back to talking about the mathematics. We now have our data expressed in terms of three numeric variables: the continuous variable Y and the two binary variables X_1 and X_2. What I want you to recognize is that our 2 × 2 factorial ANOVA without an interaction is exactly equivalent to the regression model:

$$Y_p = b_0 + b_1 X_{1p} + b_2 X_{2p} + \varepsilon_p$$

This is, of course, the exact same equation that I used earlier to describe a two-predictor regression model! The only difference is that X_1 and X_2 are now binary variables (i.e., values can only be 0 or 1), whereas in a regression analysis we expect that X_1 and X_2 will be continuous. There's a couple of ways I could try to convince you of this. One possibility would be to do a lengthy mathematical exercise proving that the two are identical. However, I'm going to go out on a limb and guess that most of the readership of this book will find that annoying rather than helpful. Instead, I'll explain the basic ideas and then rely on JASP to show that ANOVA analyses and regression analyses aren't just
similar, they're identical for all intents and purposes. Let's start by running this as an ANOVA. To do this, we'll use the rtfm data

set, and Figure 13.13 shows what we get when we run the analysis in JASP.

ANOVA

ANOVA - grade

Cases	Sum of Squares	df	Mean Square	F	p
attend	648.000	1	648.000	21.600	0.006
reading	1568.000	1	1568.000	52.267	< . 001
Residuals	150.000	5	30.000		
Note. Type III Sum of Squares					

Figure 13.13: ANOVA of the rtfm.csv data set in JASP, without the interaction term

So, by reading the key numbers off the ANOVA table and the mean scores that we presented earlier, we can see that the students obtained a higher grade if they attended class ($F_{1,5} = 21.6, p = .006$) and if they read the textbook ($F_{1,5} = 52.3, p < 0.001$). Let's make a note of those p-values and those F statistics.

Now let's think about the same analysis from a linear regression perspective. In the rtfm data set, we have encoded attend and reading as if they were numeric predictors. In this case, this is perfectly acceptable. There really is a sense in which a student who turns up to class (i.e. attend = 1) has in fact done "more attendance" than a student who does not (i.e. attend $= 0$). So it's not at all unreasonable to include it as a predictor in a regression model. It's a little unusual, because the predictor only takes on two possible values, but it doesn't violate any of the assumptions of linear regression. And it's easy to interpret. If the regression coefficient for attend is greater than 0 it means that students that attend lectures get higher grades. If it's less than zero then students attending lectures get lower grades. The same is true for our reading variable.

Wait a second though. Why is this true? It's something that is intuitively obvious to everyone who has taken a few stats classes and is comfortable with the maths, but it isn't clear to everyone else at first pass. To see why this is true, it helps to look closely at a few specific students. Let's start by considering the 6th and 7th students in our data set (i.e. $p = 6$ and $p = 7$). Neither one has read the textbook, so in both cases we can set reading $= 0$. Or, to say the same thing in our mathematical notation, we observe $X_{2,6} = 0$ and $X_{2,7} = 0$. However, student number 7 did turn up to lectures (i.e., attend $= 1, X_{1,7} = 1$) whereas student number 6 did not (i.e., attend $= 0, X_{1,6} = 0$). Now let's look at what happens when we insert these numbers into the general formula for our regression line. For
student number 6, the regression predicts that

$$\begin{aligned}\hat{Y}_6 &= b_0 + b_1 X_{1,6} + b_2 X_{2,6} \\ &= b_0 + (b_1 \times 0) + (b_2 \times 0) \\ &= b_0\end{aligned}$$

So we're expecting that this student will obtain a grade corresponding to the value of the intercept term b_0. What about student 7? This time when we insert the numbers into the formula for the regression line, we obtain the following

$$\begin{aligned}\hat{Y}_7 &= b_0 + b_1 X_{1,7} + b_2 X_{2,7} \\ &= b_0 + (b_1 \times 1) + (b_2 \times 0) \\ &= b_0 + b_1\end{aligned}$$

Because this student attended class, the predicted grade is equal to the intercept term b_0 plus the coefficient associated with the attend variable, b_1. So, if b_1 is greater than zero, we're expecting that the students who turn up to lectures will get higher grades than those students who don't. If this coefficient is negative we're expecting the opposite: students who turn up at class end up performing much worse. In fact, we can push this a little bit further. What about student number 1, who turned up to class $(X_{1,1} = 1)$ and read the textbook $(X_{2,1} = 1)$? If we plug these numbers into the regression we get

$$\begin{aligned}\hat{\gamma}_1 &= b_0 + b_1 x_{1,1} + b_2 x_{2,1} \\ &= b_0 + (b_1 \times 1) + (b_2 \times 1) \\ &= b_0 + b_1 + b_2\end{aligned}$$

So if we assume that attending class helps you get a good grade (i.e., $b_1 > 0$) and if we assume that reading the textbook also helps you get a good grade (i.e., $b_2 > 0$), then our expectation is that student 1 will get a grade that that is higher than student 6 and student 7 .

And at this point you won't be at all suprised to learn that the regression model predicts that student 3 , who read the book but didn't attend lectures, will obtain a grade of $b_2 + b_0$. I won't bore you with yet another regression formula. Instead, what I'll do is show you the following table of expected grades:

Attended lectures?	Did not read textbook	Read textbook
No	b_0	$b_0 + b_2$
Yes	$b_0 + b_1$	$b_0 + b_1 + b_2$

As you can see, the intercept term b_0 acts like a kind of "baseline" grade that you would expect from those students who don't take the time to attend class or read the textbook. Similarly, b_1 represents the boost that you're expected to get if you come to class, and b_2 represents the boost that comes
from reading the textbook. In fact, if this were an ANOVA you might very well want to characterise b_1 as the main effect of attendance, and b_2 as the main effect of reading! In fact, for a simple 2 × 2 ANOVA that's exactly how it plays out.

Okay, now that we're really starting to see why ANOVA and regression are basically the same thing, let's actually run our regression using the rtfm data and the JASP regression analysis to convince ourselves that this is really true. Running the regression in the usual way gives the results shown in Figure 13.14.

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
M_0	(Intercept)	66.500	6.500		10.231	< . 001
M_1	(Intercept)	43.500	3.354		12.969	< . 001
	attend	18.000	3.873	0.523	4.648	0.006
	reading	28.000	3.873	0.814	7.230	< . 001

Figure 13.14: Regression analysis of the rtfm.csv data set in JASP, without the interaction term

There's a few interesting things to note here. First, notice that the intercept term is 43.5 which is close to the "group" mean of 42.5 observed for those two students who didn't read the text or attend class. Second, notice that we have the regression coefficient of $b_1 = 18.0$ for the attendance variable, suggesting that those students that attended class scored 18 points higher than those who didn't. So our expectation would be that those students who turned up to class but didn't read the textbook would obtain a grade of $b_0 + b_1$, which is equal to $43.5 + 18.0 = 61.5$. You can verify for yourself that the same thing happens when we look at the students that read the textbook.

Actually, we can push a little further in establishing the equivalence of our ANOVA and our regression. Look at the p-values associated with the attend variable and the reading variable in the regression output. They're identical to the ones we encountered earlier when running the ANOVA. This might seem a little surprising, since the test used when running our regression model calculates a t-statistic and the ANOVA calculates an F-statistic. However, if you can remember I mentioned that there's a relationship between the t-distribution and the F-distribution. If you have some quantity that is distributed according to a t -distribution with k degrees of freedom and you square it, then this new squared quantity follows an F-distribution whose degrees of freedom are 1 and k. We can check this with respect to the t statistics in our regression model. For the attend variable we get a t value of 4.65. If we square this number we end up with 21.6, which matches the corresponding F statistic in our ANOVA.

13.6.3 Using contrasts to encode non binary factors

At this point, I've shown you how we can view a 2 × 2 ANOVA into a linear model. And it's pretty easy to see how this generalises to a 2 × 2 × 2 ANOVA or a 2 × 2 × 2 × 2 ANOVA. It's the same thing, really. You just add a new binary variable for each of your factors. Where it begins to get trickier is when we consider factors that have more than two levels. Consider, for instance, the 3×2 ANOVA that we ran earlier in this chapter using the clinicaltrial.csv data. How can we convert the three-level drug factor into a numerical form that is appropriate for a regression?

The answer to this question is pretty simple, actually. All we have to do is realise that a three-level factor can be redescribed as two binary variables. Suppose, for instance, I were to create a new binary variable called druganxifree. Whenever the drug variable is equal to "anxifree" we set druganxifree = 1. Otherwise, we set druganxifree = 0. This variable sets up a contrast, in this case between anxifree and the other two drugs. By itself, of course, the druganxifree contrast isn't enough to fully capture all of the information in our drug variable. We need a second contrast, one that allows us to distinguish between joyzepam and the placebo. To do this, we can create a second binary contrast, called drugjoyzepam, which equals 1 if the drug is joyzepam and 0 if it is not. Taken together, these two contrasts allows us to perfectly discriminate between all three possible drugs. The table below illustrates this:

drug	druganxifree	drugjoyzepam
"placebo"	0	0
"anxifree"	1	0
"joyzepam"	0	1

If the drug administered to a patient is a placebo then both of the two contrast variables will equal 0 . If the drug is Anxifree then the druganxifree variable will equal 1, and drugjoyzepam will be 0. The reverse is true for Joyzepam: drugjoyzepam is 1 and druganxifree is 0 .

Creating contrast variables is not too difficult to do using JASP's 'Computed Column' functionality. For example, to create the druganxifree variable, you can use some simple R code to create the binary variable in the formula box: ifelse(drug == 'anxifree', 1, 0)'. Similarly, to create the new variable drugjoyzepam use this code: ifelse(drug == 'joyzepam', 1, 0). Likewise for therapyCBT: ifelse(therapy == 'CBT', 1, 0). You can see these new variables in the JASP data file clinicaltrial2.jasp. As before, if you click on the " f_x " symbol, you will also be able to see the associated R code.

We have now recoded our three-level factor in terms of two binary variables, and we've already seen that ANOVA and regression behave the same way for binary variables. However, there are some additional complexities that arise in this case, which we'll discuss in the next section.

13.6.4 The equivalence between ANOVA and regression for non-binary factors

Now we have two different versions of the same data set. Our original data in which the drug variable from the clinicaltrial.csv file is expressed as a single three-level factor, and the expanded
data clinicaltrial2.jasp in which it is expanded into two binary contrasts. Once again, the thing that we want to demonstrate is that our original 3 × 2 factorial ANOVA is equivalent to a regression model applied to the contrast variables. Let's start by re-running the ANOVA, with results shown in Figure 13.15. Let's remove the interaction component; remember to go to 'Model' and move the drug*therapy term out of the 'Model Terms' box.

ANOVA

ANOVA - mood.gain

Cases	Sum of Squares	df	Mean Square	F	p
drug	3.453	2	1.727	26.149	< . 001
therapy	0.467	1	0.467	7.076	0.019
Residuals	0.924	14	0.066		
Note. Type III Sum of Squares					

Figure 13.15: JASP ANOVA results, without interaction component

Obviously, there are no surprises here. That's the exact same ANOVA that we ran earlier. Next, let's run a regression using druganxifree, drugjoyzepam and therapyCBT as the predictors. The results are shown in Figure 13.16.

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
M_0	(Intercept)	0.883	0.126		7.020	< . 001
M_1	(Intercept)	0.289	0.121		2.385	0.032
	druganxifree	0.267	0.148	0.242	1.797	0.094
	drugjoyzepam	1.033	0.148	0.939	6.965	< . 001
	therapyCBT	0.322	0.121	0.311	2.660	0.019

Figure 13.16: JASP regression results, with contrast variables druganxifree and drugjoyzepam

Hmm. This will look quite different than what we got last time. Not surprisingly, the regression output prints out the results for each of the three predictors separately, just like it did every other time we conducted a regression analysis. On the one hand we can see that the p-value for the therapyCBT variable is exactly the same as the one for the therapy factor in our original ANOVA, so we can be reassured that the regression model is doing the same thing as the ANOVA did. On the other hand, this regression model is testing the druganxifree contrast and the drugjoyzepam contrast separately,
as if they were two completely unrelated variables. It's not surprising of course, because the poor regression analysis has no way of knowing that drugjoyzepam and druganxifree are actually the two different contrasts that we used to encode our three-level drug factor. As far as it knows, drugjoyzepam and druganxifree are no more related to one another than drugjoyzepam and therapyCBT. However, you and I know better.

The individual drug coefficients test specific comparisons with the placebo, rather than the overall effect of drug. An overall drug test would compare a model containing therapyCBT alone with a model containing therapyCBT and both drug contrasts.

Figure 13.17 illustrates the reverse comparison: testing whether therapy adds information after accounting for drug. To reproduce this output, add both druganxifree and drugjoyzepam to the null model. The full model also includes therapyCBT. The therapy coefficient gives t = 2.660, p = .019, matching the therapy result in the main-effects ANOVA.

13.7 Different ways to specify contrasts

In the previous section, I showed you a method for converting a factor into a collection of contrasts. In the method I showed you we specify a set of binary variables in which we defined a table like this one:

drug	druganxifree	drugjoyzepam
"placebo"	0	0
"anxifree"	1	0
"joyzepam"	0	1

Each row in the table corresponds to one of the factor levels, and each column corresponds to one of the contrasts. This table, which always has one more row than columns, has a special name. It is called a contrast matrix. However, there are lots of different ways to specify a contrast matrix. In this section I discuss a few of the standard contrast matrices that statisticians use and how you can use them in JASP. If you're planning to read the section on unbalanced ANOVA later on (Section 13.10), it's worth reading this section carefully. If not, you can get away with skimming it, because the choice of contrasts doesn't matter much for balanced designs.

13.7.1 Treatment contrasts

In the particular kind of contrasts that I've described above, one level of the factor is special, and acts as a kind of "baseline" category (i.e., placebo in our example), against which the other two are defined. The name for these kinds of contrasts is treatment contrasts, also known as "dummy

Linear Regression

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
0	0.844	0.713	0.674	0.305
1	0.900	0.809	0.768	0.257

Note. Null model includes druganxifree, drugjoyzepam

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
0	Regression	3.453	2	1.727	18.611	< . 001
	Residual	1.392	15	0.093		
	Total	4.845	17			
1	Regression	3.921	3	1.307	19.791	< . 001
	Residual	0.924	14	0.066		
	Total	4.845	17			

Note. Null model includes druganxifree, drugjoyzepam

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
0	(Intercept)	0.450	0.124		3.619	0.003
	druganxifree	0.267	0.176	0.242	1.516	0.150
	drugjoyzepam	1.033	0.176	0.939	5.876	< . 001
1	(Intercept)	0.289	0.121		2.385	0.032
	druganxifree	0.267	0.148	0.242	1.797	0.094
	drugjoyzepam	1.033	0.148	0.939	6.965	< . 001
	therapyCBT	0.322	0.121	0.311	2.660	0.019

Figure 13.17: Model comparison in JASP regression, null model 0 vs. contrasts model 1

coding". In this contrast each level of the factor is compared to a base reference level, and the base reference level is the value of the intercept.

The name reflects the fact that these contrasts are quite natural and sensible when one of the categories in your factor really is special because it actually does represent a baseline. That makes sense in our clinical trial example. The placebo condition corresponds to the situation where you don't give people any real drugs, and so it's special. The other two conditions are defined in relation to the placebo. In one case you replace the placebo with Anxifree, and in the other case your replace it with Joyzepam.

The table shown above is a matrix of treatment contrasts for a factor that has 3 levels. But suppose I want a matrix of treatment contrasts for a factor with 5 levels? You would set this out like this:

Level	2	3	4	5
1	0	0	0	0
2	1	0	0	0
3	0	1	0	0

Level	2	3	4	5
4	0	0	1	0
5	0	0	0	1

In this example, the first contrast is level 2 compared with level 1 , the second contrast is level 3 compared with level 1, and so on. Notice that, by default, the first level of the factor is always treated as the baseline category (i.e., it's the one that has all zeros and doesn't have an explicit contrast associated with it). In JASP you can change which category is the first level of the factor by manipulating the order of the levels of the variable shown in the 'Data Variable' window (double click on the name of the variable in the spreadsheet column to bring up the 'Data Variable' view.

13.7.2 Helmert contrasts

Treatment contrasts are useful for a lot of situations. However, they make most sense in the situation when there really is a baseline category, and you want to assess all the other groups in relation to that one. In other situations, however, no such baseline category exists, and it may make more sense to compare each group to the mean of the other groups. This is where we meet Helmert-style contrasts. The version shown here compares each level with the mean of the previous levels and is available in JASP through ANOVA → Contrasts → Difference. JASP's Helmert option instead compares each level with the mean of subsequent levels. The idea behind Helmert contrasts is to compare each group to the mean of the "previous" ones. That is, the first contrast represents the difference between group 2 and group 1, the second contrast represents the difference between group 3 and the mean of groups 1 and 2, and so on. This translates to a contrast matrix that looks like this for a factor with five levels:

1	-1	-1	-1	-1
2	1	-1	-1	-1
3	0	2	-1	-1
4	0	0	3	-1
5	0	0	0	4

One useful thing about Helmert contrasts is that every contrast sums to zero (i.e., all the columns sum to zero). This has the consequence that, when we interpret the ANOVA as a regression, the intercept term corresponds to the grand mean $\mu_{..}$ if we are using Helmert contrasts. Compare this to treatment contrasts, in which the intercept term corresponds to the group mean for the baseline category. This property can be very useful in some situations. It doesn't matter very much if you have a balanced design, which we've been assuming so far, but it will turn out to be important later when we consider unbalanced designs in Section 13.10. In fact, the main reason why I've even bothered to include this section is that contrasts become important if you want to understand unbalanced ANOVA.

13.7.3 Sum to zero contrasts

Another coding scheme is sum-to-zero, or effect, coding. With the matrix below, the intercept represents the unweighted mean of the level means. Each displayed coefficient represents one level's deviation from that mean.

1	-1	-1	-1	-1
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1

Each column sums to zero. With this coding, the first coefficient represents the mean of group 2 minus the grand mean, the second represents the mean of group 3 minus the grand mean, and so on. These coefficients are not the pairwise differences between group 1 and the other groups.

13.7.4 Optional contrasts in JASP

JASP also comes with a variety of options that can generate different kinds of contrasts in ANOVA. These can be found in the 'Contrasts' option in the main ANOVA analysis window, where the following contrast types are listed:

Contrast type	
Deviation	Compares the mean of each level (except a reference cate- gory) to the mean of all of the levels (grand mean)
Simple	Like the treatment contrasts, the simple contrast compares the mean of each level to the mean of a specified level. This type of contrast is useful when there is a control group. By default the first category is the reference. However, with a simple contrast the intercept is the grand mean of all the levels of the factors.
Difference	Compares the mean of each level (except the first) to the mean of previous levels. (Sometimes called reverse Helmert contrasts)
Helmert	Compares the mean of each level of the factor (except the last) to the mean of subsequent levels
Repeated	Compares the mean of each level (except the last) to the mean of the subsequent level
Polynomial	Compares the linear effect and quadratic effect. The first degree of freedom contains the linear effect across all cate- gories; the second degree of freedom, the quadratic effect. These contrasts are often used to estimate polynomial trends

13.8 Post hoc tests

Time to switch to a different topic. Rather than pre-planned comparisons that you have tested using contrasts, let's suppose you've done your ANOVA and it turns out that you obtained some significant effects. Because of the fact that the F-tests are "omnibus" tests that only really test the null hypothesis that there are no differences among groups, obtaining a significant effect doesn't tell you which groups are different to which other ones. We discussed this issue and in that chapter our solution was to run t-tests for all possible pairs of groups, making corrections for multiple comparisons (e.g., Bonferroni, Holm) to control the Type I error rate across all comparisons. The methods that we used back in that chapter have the advantage of being relatively simple and being the kind of tools that you can use in a lot of different situations where you're testing multiple hypotheses, but they're not necessarily the best choices if you're interested in doing efficient post hoc testing in an ANOVA context. There are actually quite a lot of different methods for performing multiple comparisons in the statistics literature (Hsu 1996), and it would be beyond the scope of an introductory text like this one to discuss all of them in any detail.

That being said, there's one tool that I do want to draw your attention to, namely Tukey's "Honestly Significant Difference", or Tukey's HSD for short. For once, I'll spare you the formulas and just stick to the qualitative ideas. The basic idea in Tukey's HSD is to examine all relevant pairwise comparisons between groups, and it's only really appropriate to use Tukey's HSD if it is pairwise differences that you're interested in. [7] For instance, earlier we conducted a factorial ANOVA using the clinicaltrial.csv data set, and where we specified a main effect for drug and a main effect of therapy we would be interested in the following four comparisons:

- -The difference in mood gain for people given Anxifree versus people given the placebo.
- -The difference in mood gain for people given Joyzepam versus people given the placebo.
- -The difference in mood gain for people given Anxifree versus people given Joyzepam.
- -The difference in mood gain for people treated with CBT and people given no therapy.

For any one of these comparisons, we're interested in the true difference between (population) group means. Tukey's HSD controls comparisons within the specified family. Here, the three drug comparisons form one family, while therapy has a single comparison. Within a family, 95% simultaneous confidence intervals are constructed so that, across repeated studies, approximately 95% of the complete sets of intervals contain all their corresponding population differences. This does not automatically provide 95% simultaneous coverage across both families combined. Moreover, we can use these confidence intervals to calculate an adjusted p value for any specific comparison.

The TukeyHSD function in JASP is pretty easy to use. You simply specify the ANOVA model term that you want to run the post hoc tests for. For example, if we were looking to run post hoc tests for the main effects but not the interaction, we would open up the 'Post Hoc Tests' option in the ANOVA analysis screen, move the drug and therapy variables across to the box on the right, and then select the 'Tukey' checkbox in the list of possible post hoc corrections that could be applied. This, along with the corresponding

results table, is shown in Figure 13.18

Figure 13.18 uses the full model, including the drug × therapy interaction. The output shown in the 'Post Hoc Tests' results table is (I hope) pretty straightforward. The first comparison, for example, is the Anxifree versus placebo difference, and the first part of the output indicates that the observed difference in group means is .27. The next number is the standard error for the difference, from which we could calculate the 95% confidence interval (you can actually display this 95% CI if you select the 'Confidence intervals' box). Then there is a column with the degrees of freedom, a column with the t-value, and finally a column with the p -value. For the first comparison the adjusted p-value is .160. In contrast, if you look at the next line, we see that the observed difference between joyzepam and the placebo is 1.03, and this result is significant $(p < .001)$.

Post Hoc Tests

Standard (LSD)

Post Hoc Comparisons - drug

		Mean Difference	SE	t	p_{tukey}
placebo	anxifree	-0.267	0.135	-1.979	0.160
	joyzepam	-1.033	0.135	-7.671	< . 001
anxifree	joyzepam	-0.767	0.135	-5.691	< . 001

Note. P-value adjusted for comparing a family of 3
Note. Results are averaged over the levels of: therapy

Post Hoc Comparisons - therapy

		Mean Difference	SE	t	p_{tukey}
no.therapy	CBT	-0.322	0.110	-2.929	0.013

Figure 13.18: Tukey HSD post hoc test in JASP

Note. Results are averaged over the levels of: drug

So far, so good. What about the situation where your model includes interaction terms? For instance, the default option in JASP is to allow for the possibility that there is an interaction between drug and therapy. If that's the case, the number of pairwise comparisons that we need to consider starts to increase. As before, we need to consider the three comparisons that are relevant to the main effect of drug and the one comparison that is relevant to the main effect of therapy. But, if we want to consider the possibility of a significant interaction (and try to find the group differences that underpin that significant interaction), we need to include comparisons such as the following:

- -The difference in mood gain for people given Anxifree and treated with CBT, versus people given the placebo and treated with CBT
- -The difference in mood gain for people given Anxifree and given no therapy, versus people given the placebo and given no therapy.
- -etc

There are quite a lot of these comparisons that you need to consider. Figure 13.19 shows the 15 pairwise comparisons among the six drug-by-therapy cells. The four marginal comparisons are shown separately in Figure 13.18.

*Post Hoc Comparisons - drug * therapy*

		Mean Difference	SE	t	p_{tukey}
placebo no.therapy	anxifree no.therapy	-0.100	0.191	-0.525	0.994
	joyzepam no.therapy	-1.167	0.191	-6.124	< . 001

		Mean Difference	SE	t	p_{tukey}
	placebo CBT	-0.300	0.191	-1.575	0.628
	anxifree CBT	-0.733	0.191	-3.849	0.022
	joyzepam CBT	-1.200	0.191	-6.299	< .001
anxifree no.therapy	joyzepam no.therapy	-1.067	0.191	-5.599	0.001
	placebo CBT	-0.200	0.191	-1.050	0.892
	anxifree CBT	-0.633	0.191	-3.324	0.053
	joyzepam CBT	-1.100	0.191	-5.774	< .001
joyzepam no.therapy	placebo CBT	0.867	0.191	4.549	0.007
	anxifree CBT	0.433	0.191	2.275	0.275
	joyzepam CBT	-0.033	0.191	-0.175	1.000
placebo CBT	anxifree CBT	-0.433	0.191	-2.275	0.275
	joyzepam CBT	-0.900	0.191	-4.724	0.005
anxifree CBT	joyzepam CBT	-0.467	0.191	-2.449	0.214
Note. P-value adjusted for comparing a family of 6					

Figure 13.19: Tukey HSD post hoc test in JASP factorial ANOVA with an interaction term

13.9 The method of planned comparisons

Following on from the previous sections on contrasts and post hoc tests in ANOVA, I think the method of planned comparisons is important enough to deserve a quick discussion. In our discussions of multiple comparisons, in the previous section I've been assuming that the tests you want to run are genuinely post hoc. For instance, in our drugs example above, maybe you thought that the drugs would all have different effects on mood (i.e., you hypothesised a main effect of drug), but you didn't have any specific hypothesis about how they would be different, nor did you have any real idea about which pairwise comparisons would be worth looking at. If that is the case, then you really have to resort to something like Tukey's HSD to do your pairwise comparisons.

The situation is rather different, however, if you genuinely did have real, specific hypotheses about which comparisons are of interest, and you never ever have any intention to look at any other comparisons besides the ones that you specified ahead of time. When this is true, and if you honestly and rigorously stick to your noble intentions to not run any other comparisons (even when the data look like they're showing you deliciously significant effects for stuff you didn't have a hypothesis test for), then it doesn't really make a lot of sense to run something like Tukey's HSD, because it makes corrections for a whole bunch of comparisons that you never cared about and never had any intention of looking at. Planning comparisons in advance does not automatically remove the multiple-testing issue. Decide which comparisons form a family and whether to control the familywise error rate. A single prespecified comparison does not require a multiple-comparison adjustment; several planned comparisons may still require one. This situation is known as the method of planned comparisons, and it is sometimes used in clinical trials. However, further consideration is out of scope for this introductory book, but at least you know that this method exists!

13.10 Factorial ANOVA 3: unbalanced designs

Factorial ANOVA is a very handy thing to know about. It's been one of the standard tools used to analyse experimental data for many decades, and you'll find that you can't read more than two or three papers in psychology without running into an ANOVA in there somewhere. However, there's one huge difference between the ANOVAs that you'll see in a lot of real scientific articles and the ANOVAs that I've described so far. In in real life we're rarely lucky enough to have perfectly balanced designs. For one reason or another, it's typical to end up with more observations in some cells than in others. Or, to put it another way, we have an unbalanced design.

Unbalanced designs need to be treated with a lot more care than balanced designs, and the statistical theory that underpins them is a lot messier. It might be a consequence of this messiness, or it might be a shortage of time, but my experience has

been that undergraduate research methods classes in psychology have a nasty tendency to ignore this issue completely. A lot of stats textbooks tend to gloss over it too. The net result of this, I think, is that a lot of active researchers in the field don't actually know that there's several different "types" of unbalanced ANOVAs, and they produce quite different answers. In fact, reading the psychological literature, I'm kind of amazed at the fact that most people who report the results of an unbalanced factorial ANOVA don't actually give you enough details to reproduce the analysis. I secretly suspect that most people don't even realise that their statistical software package is making a whole lot of substantive data analysis decisions on their behalf. It's actually a little terrifying when you think about it. So, if you want to avoid handing control of your data analysis to stupid software, read on.

13.10.1 The coffee data

As usual, it will help us to work with some data. The coffee.csv file contains a hypothetical data set that produces an unbalanced 3×2 ANOVA. Suppose we were interested in finding out whether or not the tendency of people to babble when they have too much coffee is purely an effect of the coffee itself, or whether there's some effect of the milk and sugar that people add to the coffee. Suppose we took 18 people and gave them some coffee to drink. The amount of coffee / caffeine was held constant, and we varied whether or not milk was added, so milk is a binary factor with two levels, "yes" and "no". We also varied the kind of sugar involved. The coffee might contain "real" sugar or it might contain "fake" sugar (i.e., artificial sweetener) or it might contain "none" at all, so the sugar variable is a three level factor. Our outcome variable is a continuous variable that presumably refers to some psychologically sensible measure of the extent to which someone is "babbling". The details don't really matter for our purpose. Take a look at the data in the JASP spreadsheet view, as in Figure 13.20.

Looking at the table of means in Figure 13.20 we get a strong impression that there are differences between the groups. This is especially true when we look at the standard deviations. Across groups, this standard deviation varies quite a lot. [8] Whilst this at first may seem like a straightforward factorial ANOVA, a problem arises when we look at how many observations we have in each group. See the different Ns for different groups shown in Figure 13.20. This makes the design unbalanced. Equal cell sizes simplified our earlier calculations, but they are not a general requirement of factorial ANOVA. We haven't really discussed how to handle this situation.

13.10.2 "Standard ANOVA" does not exist for unbalanced designs

Unbalanced designs lead us to the somewhat unsettling discovery that there isn't really any one thing that we might refer to as a standard ANOVA. In fact, it turns out that there are three fundamentally different ways [9] in which you might want to run an ANOVA in an unbalanced design. If you have a balanced design all three versions produce identical results, with the sums of squares, F-values, etc., all conforming to the formulas that I gave at the start of the chapter. However, when your design is unbalanced they don't give the same answers. Furthermore, they are not all equally appropriate to every situation. Some methods will be more appropriate to your situation than others. Given all this, it's important to understand what the different types of ANOVA are and how they differ from one another.

The first kind of ANOVA is conventionally referred to as Type I sum of squares. I'm sure you can guess what the other two are called. The "sum of squares" part of the name was introduced by the SAS statistical software package and has become standard nomenclature, but it's a bit misleading in some ways. I think the logic for referring to them as different types of sum of squares is that, when you look at the ANOVA tables that they produce, the key difference in the numbers is the SS values. The degrees of freedom don't change, the MS values are still defined as SS divided by df,

Results

Descriptive Statistics

Descriptive Statistics

	babble	
	no	yes
Valid	10	8
Missing	0	0
Mean	5.320	4.750

Std. Deviation	0.796	0.962
Minimum	3.900	3.500
Maximum	6.600	5.900

Descriptive Statistics

Descriptive Statistics

	babble		
	fake	none	real
Valid	6	5	7
Missing	0	0	0
Mean	5.033	4.440	5.543
Std. Deviation	0.814	1.038	0.637
Minimum	3.900	3.500	4.600
Maximum	5.900	5.800	6.600

Figure 13.20: Descriptives for the coffee.csv data set, separately split by milk and sugar, respectively.

etc. However, what the terminology gets wrong is that it hides the reason why the SS values are different from one another. To that end, it's a lot more helpful to think of the three different kinds of ANOVA as three different hypothesis testing strategies. These different strategies lead to different SS values, to be sure, but it's the strategy that is the important thing here, not the SS values themselves. Recall from Section 13.6 that any particular F-test is best thought of as a comparison between two linear models. So, when you're looking at an ANOVA table, it helps to remember that each of those F-tests corresponds to a pair of models that are being compared. Of course, this leads naturally to the question of which pair of models is being compared. This is the fundamental difference between ANOVA Types I, II and III: each one corresponds to a different way of choosing the model pairs for the tests.

13.10.3 Type I sum of squares

The Type I method is sometimes referred to as the "sequential" sum of squares, because it involves a process of adding terms to the model one at a time. Consider the coffee data, for instance. Suppose we want to run the full 3 × 2 factorial ANOVA, including interaction terms. The full model contains the outcome variable babble, the predictor variables sugar and milk, and the interaction term sugar*milk. This can be written as babble ~ sugar + milk + sugar*milk. The Type I strategy builds this model up sequentially, starting from the simplest possible model and gradually adding terms.

The simplest possible model for the data would be one in which neither milk nor sugar is assumed to have any effect on babbling. The only term that would be included in such a model is the intercept, written as babble ~ 1. This is our initial null hypothesis. The next simplest model for the data would be one in which only one of the two main effects is included. In the coffee data, there are two different possible choices here, because we could choose to add milk first or to add sugar first. The order actually turns out to matter, as we'll see later, but for now let's just make a choice arbitrarily and pick sugar. So, the second model in our sequence of models is babble ~ sugar, and it forms the alternative hypothesis for our first test. We now have our first hypothesis test:

```
Null model: babble ~ 1
Alternative model: babble ~ sugar
```

This comparison forms our hypothesis test of the main effect of sugar. The next step in our model building exercise is to add the other main effect term, so the next model in our sequence is babble ~ sugar + milk. The second hypothesis test is then formed by comparing the following pair of models:

```
Null model: babble ~ sugar
Alternative model: babble ~ sugar + milk
```

This comparison forms our hypothesis test of the main effect of milk. In one sense, this approach is very elegant: the alternative hypothesis from the first test forms the null hypothesis for the second one. It is in this sense that the Type I method is strictly sequential. Every test builds directly on the results of the last one. However, in another sense it's very inelegant, because there's a strong asymmetry between the two tests. The test of the main effect of sugar (the first test) completely ignores milk, whereas the test of the main effect of milk (the second test) does take sugar into account. In any case, the fourth model in our sequence is now the full model, babble ~ sugar + milk + sugar*milk, and the corresponding hypothesis test is:

Null model: babble ~ sugar + milk
Alternative model: babble ~ sugar + milk + sugar*milk
Type III sum of squares is the default hypothesis testing method used by JASP ANOVA, so to run a Type I sum of squares analysis we have to select 'Type 1' in the 'Sum of squares' selection box in the JASP 'ANOVA' - 'Model' options. This gives us the ANOVA table shown in Figure 13.21.

ANOVA

ANOVA - babble

Cases	Sum of Squares	df	Mean Square	F	p
sugar	3.558	2	1.779	6.749	0.011
milk	0.956	1	0.956	3.628	0.081
sugar * milk	5.944	2	2.972	11.277	0.002
Residuals	3.162	12	0.264		
Note. Type I Sum of Squares					

Figure 13.21: ANOVA results table using Type I sum of squares in JASP

The big problem with using Type I sum of squares is the fact that it really does depend on the order in which you enter the variables. Yet, in many situations the researcher has no reason to prefer one ordering over another. This is presumably the case for our milk and sugar problem. Should we add milk first or sugar first? It feels exactly as arbitrary as a data analysis question as it does as a coffee-making question. There may in fact be some people with firm opinions about ordering, but it's hard to imagine a principled answer to the question. Yet, look what happens when we change the ordering, as in Figure 13.22.

The p-values for both main effect terms have changed, and fairly dramatically. Among other things, the effect of milk has become significant (though one should avoid drawing any strong conclusions about this, as I've mentioned previously). Which of these two ANOVAs should one report? It's not immediately obvious.

When you look at the hypothesis tests that are used to define the "first" main effect and the "second" one, it's clear that they're qualitatively different from one another. In our initial example, we saw that the test for the main effect of sugar completely ignores milk, whereas the test of the main effect of milk does take sugar into account. As such, the Type I testing strategy really does treat the first main effect as if it had a kind of theoretical primacy over the second one. In my experience there is very rarely if ever any theoretically primacy of this kind that would justify treating any two main effects asymmetrically.

ANOVA

ANOVA - babble

Cases	Sum of Squares	df	Mean Square	F	p
milk	1.444	1	1.444	5.479	0.037
sugar	3.070	2	1.535	5.824	0.017

Cases	Sum of Squares	df	Mean Square	F	p
milk * sugar	5.944	2	2.972	11.277	0.002
Residuals	3.162	12	0.264		
Note. Type I Sum of Squares					

Figure 13.22: ANOVA results table using Type I sum of squares in JASP, but with factors entered in a different order (milk first)

The consequence of all this is that Type I tests are very rarely of much interest, and so we should move on to discuss Type II tests and Type III tests.

13.10.4 Type III sum of squares

Having just finished talking about Type I tests, you might think that the natural thing to do next would be to talk about Type II tests. However, I think it's actually a bit more natural to discuss Type III tests (which are simple and the default in JASP) before talking about Type II tests (which are trickier). The basic idea behind Type III tests is extremely simple. Regardless of which term you're trying to evaluate, run the F-test in which the alternative hypothesis corresponds to the full ANOVA model as specified by the user, and the null model just deletes that one term that you're testing. The model labels below describe the terms being tested conceptually. Type III comparisons hold the chosen contrast coding fixed and test the columns associated with a term; these labels should not be treated as directly executable R formulas. For instance, in the coffee example, in which our full model was babble ~ sugar + milk + sugar*milk, the test for a main effect of sugar would correspond to a comparison between the following two models:

Null model: Intercept + milk + sugar × milk
Alternative model: Intercept + sugar + milk + sugar × milk
Similarly the main effect of milk is evaluated by testing the full model against a null model that removes the milk term, like so:

$$\begin{array}{ll} \text{Null model:} & \text{babble } \sim \text{ sugar } + \text{ sugar } * \text{ milk} \\ \text{Alternative model:} & \text{babble } \sim \text{ sugar } + \text{ milk } + \text{ sugar*milk} \end{array}$$

Finally, the interaction term sugar*milk is evaluated in exactly the same way. Once again, we test the full model against a null model that removes the sugar*milk interaction term, like so:

$$\begin{array}{ll} \text{Null model:} & \text{babble } \sim \text{ sugar } + \text{ milk} \\ \text{Alternative model:} & \text{babble } \sim \text{ sugar } + \text{ milk } + \text{ sugar*milk} \end{array}$$

The basic idea generalises to higher order ANOVAs. For instance, suppose that we were trying to run an ANOVA with three factors, A, B and C, and we wanted to consider all possible main effects and all possible interactions, including the three way interaction $\mathrm{A} * \mathrm{B} * \mathrm{C}$. The table below shows you what the Type III tests look like for this situation:

Term being tested is	Null model is outcome ~ ...	Alternative model is outcome ~ ...
A	$\mathrm{B}+\mathrm{C}+\mathrm{A}*\mathrm{B}+\mathrm{A}*\mathrm{C}+\mathrm{B}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$	$\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{A}*\mathrm{B}+\mathrm{A}*\mathrm{C}+\mathrm{B}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$
B	$\mathrm{A}+\mathrm{C}+\mathrm{A}*\mathrm{B}+\mathrm{A}*\mathrm{C}+\mathrm{B}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$	$\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{A}*\mathrm{B}+\mathrm{A}*\mathrm{C}+\mathrm{B}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$
C	$\mathrm{A}+\mathrm{B}+\mathrm{A}*\mathrm{B}+\mathrm{A}*\mathrm{C}+\mathrm{B}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$	$\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{A}*\mathrm{B}+\mathrm{A}*\mathrm{C}+\mathrm{B}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$
A*B	$\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{A}*\mathrm{C}+\mathrm{B}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$	$\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{A}*\mathrm{B}+\mathrm{A}*\mathrm{C}+\mathrm{B}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$
A*C	$\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{A}*\mathrm{B}+\mathrm{B}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$	$\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{A}*\mathrm{B}+\mathrm{A}*\mathrm{C}+\mathrm{B}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$
B*C	$\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{A}*\mathrm{B}+\mathrm{A}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$	$\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{A}*\mathrm{B}+\mathrm{A}*\mathrm{C}+\mathrm{B}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$
A*B*C	$\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{A}*\mathrm{B}+\mathrm{A}*\mathrm{C}+\mathrm{B}*\mathrm{C}$	$\mathrm{A}+\mathrm{B}+\mathrm{C}+\mathrm{A}*\mathrm{B}+\mathrm{A}*\mathrm{C}+\mathrm{B}*\mathrm{C}+\mathrm{A}*\mathrm{B}*\mathrm{C}$

As ugly as that table looks, it's pretty simple. In all cases, the alternative hypothesis corresponds to the full model which contains three main effect terms (e.g. A), three two-way interactions (e.g. A*B) and one three-way interaction (i.e., $\mathrm{A} * \mathrm{B} * \mathrm{C}$). The null

model always contains 6 of these 7 terms, and the missing one is the one whose significance we're trying to test.

At first pass, Type III tests seem like a nice idea. Firstly, we've removed the asymmetry that caused us to have problems when running Type I tests. And because we're now treating all terms the same way, the results of the hypothesis tests do not depend on the order in which we specify them. This is definitely a good thing. However, there is a big problem when interpreting the results of the tests, especially for main effect terms. Consider the coffee data. Suppose it turns out that the main effect of milk is not significant according to the Type III tests. This means that the Type III test did not find sufficient evidence for the milk main effect after accounting for the other terms. A non-significant result does not establish that the reduced model is better. *But what does that even mean? If the interaction term sugar*milk was also non-significant, we'd be tempted to conclude that the data are telling us that the only thing that matters is sugar. But suppose we have a significant interaction term, but a non-significant main effect of milk. In this case, are we to assume that there really is an "effect of sugar", an "interaction between milk and sugar", but no "effect of milk"? That seems crazy. The right answer simply must be that it's meaningless [10] to talk about the main effect if the interaction is significant. In general, this seems to be what most statisticians advise us to do, and

I think that's the right advice. But if it really is meaningless to talk about non-significant main effects in the presence of a significant interaction, then it's not at all obvious why Type III tests should allow the null hypothesis to rely on a model that includes the interaction but omits one of the main effects that make it up. When characterised in this fashion, the null hypotheses really don't make much sense at all.

Later on, we'll see that Type III tests can be redeemed in some contexts, but first let's take a look at the ANOVA results table using Type III sum of squares, see Figure 13.23.

ANOVA

ANOVA - babble

Cases	Sum of Squares	df	Mean Square	F	p
milk	1.004	1	1.004	3.810	0.075
sugar	2.132	2	1.066	4.045	0.045
milk * sugar	5.944	2	2.972	11.277	0.002
Residuals	3.162	12	0.264		
Note. Type III Sum of Squares					

Figure 13.23: ANOVA results table using Type III sum of squares in JASP

But be aware, one of the perverse features of the Type III testing strategy is that typically the results turn out to depend on the contrasts that you use to encode your factors (see Section 13.7 if you've forgotten what the different types of contrasts are). [11]

Okay, so if the p-values that typically come out of Type III analyses are so sensitive to the choice of contrasts, does that mean that Type III tests are essentially arbitrary and not to be trusted? To some extent that's true, and when we turn to a discussion of Type II tests we'll see that Type II analyses avoid this arbitrariness entirely, but I think that's too strong a conclusion. Firstly, it's important to recognise that some choices of contrasts will always produce the same answers (ah, so this is what is happening in JASP). Of particular importance is the fact that if the columns of our contrast matrix are all constrained to sum to zero, then the Type III analysis will always give the same answers.

13.10.5 Type II sum of squares

Okay, so we've seen Type I and III tests now, and both are pretty straightforward. Type I tests are

performed by gradually adding terms one at a time, whereas Type III tests are performed by taking the full model and looking to see what happens when you remove each term. However, both can have some limitations. Type I tests are dependent on the order in which you enter the terms, and Type III tests are dependent on how you code up your contrasts. Type II tests are a little harder to describe, but they avoid both of these problems, and as a result they are a little easier to interpret.

Type II tests are broadly similar to Type III tests. Start with a "full" model, and test a particular term by deleting it from that model.

However, Type II tests are based on the marginality principle which states that you should not omit a lower order term from your model if there are any higher order ones that depend on it. So, for instance, if your model contains the two-way interaction A*B (a 2nd order term), then it really ought to contain the main effects A and B (1st order terms). Similarly, if it contains a three-way interaction term $\mathrm{A} * \mathrm{B} * \mathrm{C}$, then the model must also include the main effects A, B and C as well as the simpler interactions A*B, A*C and B*C. Type III tests routinely violate the marginality principle. For instance, consider the test of the main effect of A in the context of a three-way ANOVA that includes all possible interaction terms. According to Type III tests, our null and alternative models are:

$$\begin{array}{ll} \text{Null model:} & \text{outcome } \sim \mathrm{B} + \mathrm{C} + \mathrm{A} * \mathrm{B} + \mathrm{A} * \mathrm{C} + \mathrm{B} * \mathrm{C} + \mathrm{A} * \mathrm{B} * \mathrm{C} \\ \text{Alternative model:} & \text{outcome } \sim \mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{A} * \mathrm{B} + \mathrm{A} * \mathrm{C} + \mathrm{B} * \mathrm{C} + \mathrm{A} * \mathrm{B} * \mathrm{C} \end{array}$$

Notice that the null hypothesis omits A, but includes A*B*, *A*C and A*B*C as part of the model. This, according to the Type II tests, is not a good choice of null hypothesis. What we should do instead, if we want to test the null hypothesis that A is not relevant to our outcome, is to specify the null hypothesis that is the most complicated model that does not rely on A in any form, even as an interaction. The alternative hypothesis corresponds to this null model plus a main effect term of A. This is a lot closer to what most people would intuitively think of as a "main effect of A", and it yields the following as our Type II test of the main effect of A: [12]

$$\begin{array}{ll} \text{Null model:} & \text{outcome } \sim \mathrm{B} + \mathrm{C} + \mathrm{B} * \mathrm{C} \\ \text{Alternative model:} & \text{outcome } \sim \mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{B} * \mathrm{C} \end{array}$$

Anyway, just to give you a sense of how the Type II tests play out, here's the full table of tests that would be applied in a three-way factorial ANOVA:

Term being tested is	Null model is outcome ~ ...	Alternative model is outcome ~...
A	B + C + B*C	$\mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{B} * \mathrm{C}$
B	$\mathrm{A} + \mathrm{C} + \mathrm{A} * \mathrm{C}$	$\mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{A} * \mathrm{C}$
C	$\mathrm{A} + \mathrm{B} + \mathrm{A} * \mathrm{B}$	$\mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{A} * \mathrm{B}$
A*B	$\mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{A} * \mathrm{C} + \mathrm{B} * \mathrm{C}$	$\mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{A} * \mathrm{B} + \mathrm{A} * \mathrm{C} + \mathrm{B} * \mathrm{C}$
A*C	$\mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{A} * \mathrm{B} + \mathrm{B} * \mathrm{C}$	$\mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{A} * \mathrm{B} + \mathrm{A} * \mathrm{C} + \mathrm{B} * \mathrm{C}$
B*C	$\mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{A} * \mathrm{B} + \mathrm{A} * \mathrm{C}$	$\mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{A} * \mathrm{B} + \mathrm{A} * \mathrm{C} + \mathrm{B} * \mathrm{C}$
A*B*C	$\mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{A} * \mathrm{B} + \mathrm{A} * \mathrm{C} + \mathrm{B} * \mathrm{C}$	$\mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{A} * \mathrm{B} + \mathrm{A} * \mathrm{C} + \mathrm{B} * \mathrm{C} + \mathrm{A} * \mathrm{B} * \mathrm{C}$

In the context of the two way ANOVA that we've been using in the coffee data, the hypothesis tests are even simpler. The main effect of sugar corresponds to an F-test comparing these two models:

```
Null model: babble ~ milk
Alternative model: babble ~ sugar + milk
```

The test for the main effect of milk is

```
Null model: babble ~ sugar
Alternative model: babble ~ sugar + milk
```

Finally, the test for the interaction sugar*milk is:

```
Null model: babble ~ sugar + milk
Alternative model: babble ~ sugar + milk + sugar*milk
```

Running the tests are again straightforward. Just select 'Type 2' in the 'Sum of squares' selection box in the JASP 'ANOVA' - 'Model' options, This gives us the ANOVA table shown in Figure 13.24.

ANOVA

ANOVA - babble

Cases	Sum of Squares	df	Mean Square	F	p
milk	0.956	1	0.956	3.628	0.081
sugar	3.070	2	1.535	5.824	0.017
milk * sugar	5.944	2	2.972	11.277	0.002
Residuals	3.162	12	0.264		

Figure 13.24: ANOVA results table using Type II sum of squares in JASP

Note. Type II Sum of Squares

Type II tests have some clear advantages over Type I and Type III tests. They don't depend on the order in which you specify factors (unlike Type I), and they don't depend on the contrasts that you use to specify your factors (unlike Type III). And although opinions may differ on this last point, and it will definitely depend on what you're trying to do with your data, I do think that the hypothesis tests that they specify are more likely to correspond to something that you actually care about. As a consequence, I find that it's usually easier to interpret the results of a Type II test than the results of a Type I or Type III test. Type II tests can be useful when interactions involving the main effects are absent or not substantively important. When interactions matter, choose tests or contrasts that address the research question rather than treating Type II as an automatic default.

13.10.6 Effect sizes (and non-additive sums of squares)

JASP also provides the effect sizes η^2 and partial η^2 when you select these options. However, when you've got an unbalanced design there's a bit of extra complexity involved.

If you remember back to our very early discussions of ANOVA, one of the key ideas behind the sums of squares calculations is that if we add up all the SS terms associated with the effects in the model, and add that to the residual SS, they're supposed to add up to the total sum of squares. And, on top of that, the whole idea behind η^2 is that, because you're dividing one of the SS terms by the total SS value, an η^2 value can be interpreted as the proportion of variance accounted for by a particular term. But this is not so straightforward in unbalanced designs because some of the variance goes "missing".

This seems a bit odd at first, but here's why. In unbalanced designs, your factors can become associated with one another and it becomes difficult to tell the difference between the effect of Factor A and the effect of Factor B. In the extreme case, suppose that we'd run a 2 × 2 design in which the number of participants in each group had been as follows:

	sugar	no sugar
milk	100	0
no milk	0	100

Here we have a spectacularly unbalanced design: 100 people have milk and sugar, 100 people have no milk and no sugar, and that's all. There are 0 people with milk and no sugar, and 0 people with sugar but no milk. Now suppose that, when we collected the data, it turned out there is a large (and statistically significant) difference between the "milk and sugar" group and the "no-milk and no-sugar" group. Is this a main effect of sugar? A main effect of milk? Or an interaction? It's impossible to tell, because the presence of sugar has a perfect association with the presence of milk. Now suppose the design had been a little more balanced:

	sugar	no sugar
milk	100	5
no milk	5	100

This time around, it's technically possible to distinguish between the effect of milk and the effect of sugar, because we have a few people that have one but not the other. However, it will still be pretty difficult to do so, because the association between sugar and

milk is still extremely strong, and there are so few observations in two of the groups. Again, we're very likely to be in the situation where we know that the predictor variables (milk and sugar) are related to the outcome (babbling), but we don't know if the nature of that relationship is a main effect of one or the other predictor, or the interaction.

With unbalanced data, Type II and Type III sums of squares are adjusted quantities from different model comparisons. They do not generally form an additive partition of the total sum of squares. Their sum can be smaller or larger than the model sum of squares; variance has not literally disappeared. Type I sums of squares provide a sequential partition, but the allocation depends on the order of the terms.

13.11 Summary

- -Factorial ANOVA with balanced designs, without interactions and with interactions included
- -Effect size, estimated means, and confidence intervals in a factorial ANOVA
- -Checking assumptions in ANOVA
- -Analysis of Covariance (ANCOVA)
- -Understanding the linear model underlying ANOVA, including different contrasts
- -Post hoc testing using Tukey's HSD and a brief commentary on planned comparisons
- -Factorial ANOVA with unbalanced designs

---

1. \){ }^{1}\) The nice thing about the subscript notation is that it generalises nicely. If our experiment had involved a third factor, then we could just add a third subscript. In principle, the notation extends to as many factors as you might care to include, but in this book we'll rarely consider analyses involving more than two factors, and never more than three.
2. \){ }^{2}\) Technically, marginalising isn't quite identical to a regular mean. It's a weighted average where you take into account the frequency of the different events that you're averaging over. However, in a balanced design, all of our cell frequencies are equal by definition so the two are equivalent. We'll discuss unbalanced designs later, and when we do so you'll see that all of our calculations become a real headache. But let's ignore this for now.
3. \){ }^{a}\) English translation: "least tedious".
4. \){ }^{3}\) You may have spotted this already when trying to re-create the main effects analysis in JASP that we described earlier. For the purpose of the explanation in this book I removed the interaction component from the earlier model to keep things clean and simple
5. \){ }^{4}\) This chapter seems to be setting a new record for the number of different things that the letter R can stand for. So far we have R referring to the software package, the number of rows in our table of means, the residuals in the model, and now the correlation coefficient in a regression. Sorry. We clearly don't have enough letters in the alphabet. However, I've tried pretty hard to be clear on which thing R is referring to in each case.
6. \){ }^{5}\) Implausibly large, I would think. The artificiality of this data set is really starting to show!
7. \){ }^{6}\) What's the difference between treatment and simple contrasts, I hear you ask? Well, as a basic example consider a gender main effect, with $\mathrm{m} = 0$ and $\mathrm{f} = 1$. The coefficient corresponding to the treatment contrast will measure the difference in mean between females and males, and the intercept would be the mean of the males. However, with a simple contrast, i.e., $\mathrm{m} = -1$ and $\mathrm{f} = 1$, the intercept is the average of the means and the main effect is the difference of each group mean from the intercept.
8. \){ }^{7}\) If, for instance, you actually find yourself interested to know if Group A is significantly different from the mean of Group B and Group C, then you need to use a different tool (e.g., Scheffe's method, which is more conservative, and beyond the scope of this book). However, in most cases you probably are interested in pairwise group differences, so Tukey's HSD is a pretty useful thing to know about.

9. \){ }^{8}\) This discrepancy in standard deviations might (and should) make you wonder if we have a violation of the homogeneity of variance assumption. I'll leave it as an exercise for the reader to double check this using the Levene test option.
   \){ }^{9}\) Actually, this is a bit of a lie. ANOVAs can vary in other ways besides the ones I've discussed in this book. For instance, I've completely ignored the difference between fixed-effect models in which the levels of a factor are "fixed" by the experimenter or the world, and random-effect models in which the levels are random samples from a larger population of possible levels (this book only covers fixed-effect models). Don't make the mistake of thinking that this book, or any other one, will tell you "everything you need to know" about statistics, any more than a single book could possibly tell you everything you need to know about psychology, physics or philosophy. Life is too complicated for that to ever be true. This isn't a cause for despair, though. Most researchers get by with a basic working knowledge of ANOVA that doesn't go any further than this book does. I just want you to keep in mind that this book is only the beginning of a very long story, not the whole story.

10. \){ }^{10}\) Or, at the very least, rarely of interest.
11. \){ }^{11}\) However, in JASP the results for Type III sum of squares ANOVA are the same regardless of the contrast selected, so JASP is obviously doing something different!
12. \){ }^{12}\) Note, of course, that this does depend on the model that the user specified. If the original ANOVA model doesn't contain an interaction term for B*C, then obviously it won't appear in either the null or the alternative. But that's true for Types I, II and III. They never include any terms that you didn't include, but they make different choices about how to construct tests for the ones that you did include.
13. \){ }^{13}\) I find it amusing to note that the default in R is Type I and the default in SPSS, JASP, and jamovi is Type III. Neither of these appeals to me all that much. Relatedly, I find it depressing that almost nobody in the psychological literature ever bothers to report which Type of tests they ran, much less the order of variables (for Type I) or the contrasts used (for Type III). Often they don't report what software they used either. The only way I can ever make any sense of what people typically report is to try to guess from auxiliary cues which software they were using, and to assume that they never changed the default settings. Please don't do this! Now that you know about these issues make sure you indicate what software you used, and if you're reporting ANOVA results for unbalanced data, then specify what Type of tests you ran, specify order information if you've done Type I tests and specify contrasts if you've done Type III tests. Or, even better, do hypotheses tests that correspond to things you really care about and then report those!

This page titled 13.1: Factorial ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 13: Factorial ANOVA by Danielle J. Navarro, David R. Foxcroft, and Thomas J. Faulkenberry is licensed CC BY-SA 4.0.

13.2: Worked Example - Calculating Sums of Squares for a Two-Factor ANOVA

We are now going to work through some examples of calculating the ANOVA table for 2x2 designs. We will start with the between-subjects ANOVA for 2x2 designs. We do essentially the same thing that we did before (in the other ANOVAs), and the only new thing is to show how to compute the interaction effect.

Remember the logic of the ANOVA is to partition the variance into different parts. The SS formula for the between-subjects 2x2 ANOVA looks like this:

$$SS_{\text{Total}} = SS_{\text{Effect IV1}} + SS_{\text{Effect IV2}} + SS_{\text{Effect IV1xIV2}} + SS_{\text{Error}}$$

SS Total

We calculate the grand mean (mean of all of the score). Then, we calculate the differences between each score and the grand mean. We square the difference scores, and sum them up. That is SS_{Total}, reported in the bottom yellow row.

	All Conditions				Difference from Grand Mean				Squared Differences			
	No Reward		Reward		No Reward		Reward		No Reward		Reward	
	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction
	A	B	C	D	A-GrandM	B-GrandM	C-GrandM	D-GrandM	(A-GrandM)^2	(B-GrandM)^2	(C-GrandM)^2	(D-GrandM)^2
	10	5	12	9	1.05	-3.95	3.05	0.05	1.1025	15.6025	9.3025	0.0025
	8	4	13	8	-0.95	-4.95	4.05	-0.95	0.9025	24.5025	16.4025	0.9025
	11	3	14	10	2.05	-5.95	5.05	1.05	4.2025	35.4025	25.5025	1.1025
	9	4	11	11	0.05	-4.95	2.05	2.05	0.0025	24.5025	4.2025	4.2025
	10	2	13	12	1.05	-6.95	4.05	3.05	1.1025	48.3025	16.4025	9.3025
Means	9.6	3.6	12.6	10								
Grand Mean	8.95											
sums								Sums	7.3125	148.3125	71.8125	15.5125
SS Total								SS Total	242.95			

SS Distraction

We need to compute the SS for the main effect for distraction. We calculate the grand mean (mean of all of the scores). Then, we calculate the means for the two distraction conditions. Then we treat each score as if it was the mean for it's respective distraction condition. We find the differences between each distraction condition mean and the grand mean. Then we square the differences and sum them up. That is $SS_{\text{Distraction}}$, reported in the bottom yellow row.

	All Conditions				Distraction Mean - GM				Squared Differences			
	No Reward		Reward		No Reward		Reward		No Reward		Reward	
	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction
	A	B	C	D	NDM-GM A	DM-GM B	NDM-GM C	DM-GM D	(NDM-GM)^2 A	(DM-GM)^2 B	(NDM-GM)^2 C	(DM-GM)^2 D
	10	5	12	9	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
	8	4	13	8	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
	11	3	14	10	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
	9	4	11	11	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
	10	2	13	12	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
Means	9.6	3.6	12.6	10								
Grand Mean	8.95	No Distraction	11.1	Distraction	6.8							
sums								Sums	23.1125	23.1125	23.1125	23.1125
SS Distraction								SS Distraction	92.45			

These tables are a lot to look at! Notice here, that we first found the grand mean (8.95). Then we found the mean for all the scores in the no-distraction condition (columns A and C), that was 11.1. All of the difference scores for the no-distraction condition are 11.1-8.95 = 2.15. We also found the mean for the scores in the distraction condition (columns B and D), that was 6.8. So, all of the difference scores are 6.8-8.95 = -2.15. Remember, means are the balancing point in the data, this is why the difference scores are +2.15 and -2.15. The grand mean 8.95 is in between the two condition means (11.1 and 6.8), by a difference of 2.15.

SS Reward

We need to compute the SS for the main effect for reward. We calculate the grand mean (mean of all of the scores). Then, we calculate the means for the two reward conditions. Then we treat each score as if it was the mean for it's respective reward condition. We find the differences between each reward condition mean and the grand mean. Then we square the differences and sum them up. That is SS_{Reward}, reported in the bottom yellow row.

	All Conditions				Reward Mean - GM				Squared Differences			
	No Reward		Reward		No Reward		Reward		No Reward		Reward	
	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction
	A	B	C	D	NRM-GM A	NRM-GM B	RM-GM C	RM-GM D	(NRM-GM)^2 A	(NRM-GM)^2 B	(RM-GM)^2 C	(RM-GM)^2 D
	10	5	12	9	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
	8	4	13	8	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
	11	3	14	10	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
	9	4	11	11	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
	10	2	13	12	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
Means	9.6	3.6	12.6	10								
Grand Mean	8.95	No Reward	6.6	Reward	11.3							
sums								Sums	27.6125	27.6125	27.6125	27.6125
SS Reward								SS Reward	110.45			

Now we treat each no-reward score as the mean for the no-reward condition (6.6), then subtract the grand mean (8.95) to get –2.35. We treat each reward score as the mean for the reward condition (11.3), then subtract the grand mean to get +2.35. Then we square the differences and sum them up.

SS Distraction by Reward

We need to compute the SS for the interaction effect between distraction and reward. This is the new thing that we do in an ANOVA with more than one IV. How do we calculate the variation explained by the interaction?

What we will do is this. We will find the four condition means. Then we will see how much additional variation they explain beyond the group means for reward and distraction. To do this we treat each score as the condition mean for that score. Then we subtract the mean for the distraction group, and the mean for the reward group, and then we add the grand mean. This gives us the unique variation that is due to the interaction.

Here is a formula to describe the process for each score:

$$\bar{X}_{\text{condition}} - \bar{X}_{\text{IV1}} - \bar{X}_{\text{IV2}} + \bar{X}_{\text{Grand Mean}}$$

When you look at the following table, we apply this formula to the calculation of each of the differences scores. We then square the difference scores, and sum them up to get $SS_{\text{Interaction}}$, which is reported in the bottom yellow row.

	All Conditions				Interaction Differences				Squared Differences			
	No Reward		Reward		No Reward		Reward		No Reward		Reward	
	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction
	A	B	C	D	A-ND-NR+GM	B-D-NR+GM	C-ND-R+GM	D-D-R+GM	(A-ND-NR+GM)^2 A	(B-D-NR+GM)^2 B	(C-ND-R+GM)^2 C	(D-D-R+GM)^2 D
	10	5	12	9	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
	8	4	13	8	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
	11	3	14	10	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
	9	4	11	11	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
	10	2	13	12	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
Means	9.6	3.6	12.6	10								
Grand Mean	8.95											
sums								Sums	3.6125	3.6125	3.6125	3.6125
SS Interaction								SS Interaction	14.45			

SS Error

The last thing we need to find is the SS Error. We can solve for that because we found everything else in this formula:

$$SS_{\text{Total}} = SS_{\text{Effect IV1}} + SS_{\text{Effect IV2}} + SS_{\text{Effect IV1xIV2}} + SS_{\text{Error}}$$

$$\begin{aligned} SS_{\text{Error}} &= SS_{\text{Total}} - SS_{\text{Effect IV1}} - SS_{\text{Effect IV2}} - SS_{\text{Effect IV1xIV2}} \\ &= 242.95 - 92.45 - 110.45 - 14.45 \\ &= 25.6 \end{aligned}$$

This page titled 13.2: Worked Example - Calculating Sums of Squares for a Two-Factor ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

- 3.2I: Sanctions is licensed CC BY-SA 4.0.

13.3: Worked Example - Completing and Reporting a Two-Factor ANOVA

Worked Example: Completing and Reporting a Two-Factor ANOVA

On the previous page, we calculated the sums of squares for a study examining distraction and reward. There were two levels of distraction, two levels of reward, and five participants in each condition. Each participant completed only one condition, making this a 2 × 2 between-subjects design.

We found the following sums of squares:

- Distraction: 92.45
- Reward: 110.45
- Distraction × Reward: 14.45
- Error: 25.60
- Total: 242.95

Now we can use these values to calculate degrees of freedom, mean squares, and our three *F* ratios.

Calculating Degrees of Freedom

Each main effect has degrees of freedom equal to the number of levels of that factor minus one. Both factors have two levels:

$$df_{\text{Distraction}} = 2 - 1 = 1 \tag{13.3.1}$$

$$df_{\text{Reward}} = 2 - 1 = 1 \tag{13.3.2}$$

For the interaction, multiply the degrees of freedom for the two factors:

$$df_{\text{Interaction}} = (2 - 1)(2 - 1) = 1 \tag{13.3.3}$$

For error, subtract the number of conditions from the total number of participants. We have 20 participants across four conditions:

$$df_{\text{Error}} = 20 - 4 = 16 \tag{13.3.4}$$

Finally:

$$df_{\text{Total}} = 20 - 1 = 19 \tag{13.3.5}$$

The degrees of freedom for the two main effects, interaction, and error should add up to the total:

$$1 + 1 + 1 + 16 = 19 \tag{13.3.6}$$

Calculating Mean Squares

A mean square is a sum of squares divided by its degrees of freedom:

$$MS = \frac{SS}{df} \tag{13.3.7}$$

For our example:

$$MS_{\text{Distraction}} = \frac{92.45}{1} = 92.45 \tag{13.3.8}$$

$$MS_{\text{Reward}} = \frac{110.45}{1} = 110.45 \tag{13.3.9}$$

$$MS_{\text{Interaction}} = \frac{14.45}{1} = 14.45 \tag{13.3.10}$$

$$MS_{\text{Error}} = \frac{25.60}{16} = 1.60 \tag{13.3.11}$$

The first three mean squares happen to equal their sums of squares because their degrees of freedom are all 1. This will not be true for every factorial ANOVA.

Calculating the F Ratios

We calculate a separate *F* ratio for each main effect and the interaction. Each uses the same error mean square as its denominator:

$$F_{\text{Distraction}} = \frac{92.45}{1.60} = 57.78 \tag{13.3.12}$$

$$F_{\text{Reward}} = \frac{110.45}{1.60} = 69.03 \tag{13.3.13}$$

$$F_{\text{Interaction}} = \frac{14.45}{1.60} = 9.03 \tag{13.3.14}$$

Our completed ANOVA table is:

Completed two-factor ANOVA

Source	SS	df	MS	F	p
Distraction	92.45	1	92.45	57.78	< .001
Reward	110.45	1	110.45	69.03	< .001
Distraction × Reward	14.45	1	14.45	9.03	.008
Error	25.60	16	1.60	—	—
Total	242.95	19	—	—	—

The *p* values come from comparing each *F* ratio with an *F* distribution with 1 numerator degree of freedom and 16 denominator degrees of freedom. All three results are statistically significant at α = .05.

Understanding the Interaction

A significant interaction tells us that the effect of one factor depends on the level of the other. To understand the pattern, we need to look at the four cell means:

Mean number of targets spotted in each condition

Condition	No reward	Reward
No distraction	9.60	12.60
Distraction	3.60	10.00

Without distraction, the reward group spotted 3.00 more targets than the no-reward group:

$$12.60 - 9.60 = 3.00 \tag{13.3.15}$$

With distraction, that difference was 6.40 targets:

$$10.00 - 3.60 = 6.40 \tag{13.3.16}$$

Reward was associated with higher scores in both conditions, but the difference was larger when participants were distracted. That difference in the size of the reward effect is the interaction.

Following Up the Interaction

The interaction does not tell us whether each individual reward comparison is statistically significant. We can examine this using simple-effects comparisons: compare reward with no reward separately within each distraction condition.

Using the ANOVA's pooled error estimate, the standard error for either comparison is:

$$SE = \sqrt{MS_{\text{Error}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{1.60\left(\frac{1}{5} + \frac{1}{5}\right)} = \sqrt{0.64} = 0.80 \tag{13.3.17}$$

Without distraction:

$$t = \frac{12.60 - 9.60}{0.80} = 3.75 \tag{13.3.18}$$

With distraction:

$$t = \frac{10.00 - 3.60}{0.80} = 8.00 \tag{13.3.19}$$

Both comparisons use the ANOVA error degrees of freedom, *df* = 16. Because we are conducting two comparisons, we apply a Bonferroni adjustment by multiplying each two-tailed *p* value by 2. The adjusted results are $p = .0003$ without distraction and $p < .0001$ with distraction.

We can also calculate partial eta squared for the interaction:

$$\eta_p^2 = \frac{SS_{\text{Interaction}}}{SS_{\text{Interaction}} + SS_{\text{Error}}} = \frac{14.45}{14.45 + 25.60} = .36 \tag{13.3.20}$$

APA Formatted Summary Example

A two-way between-subjects ANOVA revealed a significant interaction between distraction and reward on the number of targets spotted, $F(1, 16) = 9.03$, $p = .0008$, $\eta_p^2 = .36$. Bonferroni-adjusted simple-effects comparisons indicated that participants receiving a reward spotted significantly more targets than those receiving no reward both without distraction ($M = 12.60$ versus $M = 9.60$), $t(16) = 3.75$, $p = .0003$, and with distraction ($M = 10.00$ versus $M = 3.60$), $t(16) = 8.00$, $p < .0001$. Reward was therefore associated with better performance in both conditions, but the improvement was larger with distraction (6.40 targets) than without distraction (3.00 targets).

Example data and sums of squares are from Michelle Oja's Calculating Sum of Squares for the Factorial ANOVA Summary Table, available through Statistics LibreTexts under CC BY-SA 4.0. The calculations and reporting explanation on this page extend that example.

This page titled 13.3: Worked Example - Completing and Reporting a Two-Factor ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

INDEX

No index terms found.

This page titled Index is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

GLOSSARY

S

Sample Word 1 | Sample Definition 1

This page titled Glossary is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.

DETAILED LICENSING

Overview

Title: Statistical Methods in the Behavioral Sciences with JASP (Psych 10 at IVC)

Webpages: 55

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 92.7% (51 pages)
- Public Domain: 5.5% (3 pages)
- CC BY-SA 4.0: 1.8% (1 page)

By Page

- Statistical Methods in the Behavioral Sciences with JASP (Psych 10 at IVC) - *CC BY-NC-SA 4.0*
  - Front Matter - *CC BY-NC-SA 4.0*
    - TitlePage - *CC BY-NC-SA 4.0*
    - InfoPage - *Public Domain*
    - Table of Contents - *Public Domain*
    - Licensing - *Public Domain*
  - 1: Introduction to Statistics - *CC BY-NC-SA 4.0*
    - 1.1: Introduction to this Textbook - *CC BY-NC-SA 4.0*
    - 1.2: Why do we learn statistics? - *CC BY-SA 4.0*
    - 1.3: Why JASP? - *CC BY-NC-SA 4.0*
    - 1.4: Communicating Statistical Results - *CC BY-NC-SA 4.0*
  - 2: Research Design and Measurement - *CC BY-NC-SA 4.0*
    - 2.1: A brief introduction to research design - *CC BY-NC-SA 4.0*
  - 3: Getting Started with JASP - *CC BY-NC-SA 4.0*
    - 3.1: Getting started with JASP - *CC BY-NC-SA 4.0*
    - 3.2: Installing and Opening JASP - *CC BY-NC-SA 4.0*
    - 3.3: Accessibility and Other Settings - *CC BY-NC-SA 4.0*
    - 3.4: Creating Data Files in JASP - *CC BY-NC-SA 4.0*
    - 3.5: Importing Data from Excel into JASP - *CC BY-NC-SA 4.0*
  - 4: Descriptive Statistics - *CC BY-NC-SA 4.0*
    - 4.1: Descriptive statistics - *CC BY-NC-SA 4.0*
    - 4.2: Descriptive Statistics in JASP - *CC BY-NC-SA 4.0*
  - 5: Visualizing Data - *CC BY-NC-SA 4.0*
    - 5.1: Drawing graphs - *CC BY-NC-SA 4.0*
    - 5.2: Basic Plots - *CC BY-NC-SA 4.0*
    - 5.3: Tables - *CC BY-NC-SA 4.0*
    - 5.4: Exporting and Editing Plots from JASP - *CC BY-NC-SA 4.0*
  - 6: Probability and Distributions - *CC BY-NC-SA 4.0*
    - 6.1: Introduction to probability - *CC BY-NC-SA 4.0*
  - 7: Sampling and Estimation - *CC BY-NC-SA 4.0*
    - 7.1: Estimating unknown quantities from a sample - *CC BY-NC-SA 4.0*
  - 8: Hypothesis Testing - *CC BY-NC-SA 4.0*
    - 8.1: Hypothesis testing - *CC BY-NC-SA 4.0*
  - 9: Chi-Square Tests - *CC BY-NC-SA 4.0*
    - 9.1: Categorical data analysis - *CC BY-NC-SA 4.0*
    - 9.2: Chi-squared Test - *CC BY-NC-SA 4.0*
  - 10: Comparing Means - z-Tests and t-Tests - *CC BY-NC-SA 4.0*
    - 10.1: Comparing two means - *CC BY-NC-SA 4.0*
    - 10.2: Independent t-test and Mann-Whitney U - *CC BY-NC-SA 4.0*
    - 10.3: Paired t-test and Wilcoxon Signed Ranks - *CC BY-NC-SA 4.0*
    - 10.4: Worked Example- Calculating an Independent Samples t-Test - *CC BY-NC-SA 4.0*
  - 11: Correlation and Linear Regression - *CC BY-NC-SA 4.0*
    - 11.1: Correlation and linear regression - *CC BY-NC-SA 4.0*
    - 11.2: Pearson and Spearman Correlations - *CC BY-NC-SA 4.0*
    - 11.3: Calculating and Reporting Pearson's r - *CC BY-NC-SA 4.0*
  - 12: One-Way ANOVA - *CC BY-NC-SA 4.0*
    - 12.1: Comparing several means (one-way ANOVA) - *CC BY-NC-SA 4.0*
  - 13: Factorial ANOVA - Main Effects and Interactions - *CC BY-NC-SA 4.0*

- 13.1: Factorial ANOVA - *CC BY-NC-SA 4.0*
- 13.2: Worked Example - Calculating Sums of Squares for a Two-Factor ANOVA - *CC BY-NC-SA 4.0*
- 13.3: Worked Example - Completing and Reporting a Two-Factor ANOVA - *CC BY-NC-SA 4.0*

Back Matter - *CC BY-NC-SA 4.0*

- Index - *CC BY-NC-SA 4.0*
- Glossary - *CC BY-NC-SA 4.0*
- Detailed Licensing - *CC BY-NC-SA 4.0*

This page titled Statistical Methods in the Behavioral Sciences with JASP (Psych 10 at IVC) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Hartrisha Kaur Uchida.