A Contemporary Approach to Research and Statistics in Psychology

Klaire Somoray

James Cook University

The LibreTexts libraries are Powered by NICE CXOne and are supported by the Department of Education Open Textbook Pilot Project, 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.

TABLE OF CONTENTS

Licensing

Acknowledgements

Acknowledgement of Country

About the Book

About the Author

Why do we Need Another Book About Research and Statistics?

Chapter 1: Research and Statistical Thinking in Everyday Life

- 1.1: What can Statistics do for us?
- 1.2: If I Apply Statistical Thinking all the Time then why do I find it Difficult?
- 1.3: Big Ideas in Statistics
- 1.4: Data is/are

Chapter 2: Working with jamovi

- 2.1: Why jamovi?
- 2.2: Getting Started with jamovi
- 2.3: Analyses
- 2.4: The Spreadsheet
- 2.5: Loading Data in jamovi
- 2.6: Installing add-on Modules into jamovi

Chapter 3: Brief Review of Research Methods

- 3.1: How do we Measure Variables in Psychology?
- 3.2: Introduction to Psychological Measurement
- 3.3: What Makes a Good Measure?
- 3.4: Some Complexities
- 3.5: The Role of Variables - Predictors and Outcomes
- 3.6: Research Design I- Experimental Designs
- 3.7: Research Design II- Non-Experimental Designs

Chapter 4: The Replication Crisis

- 4.1: How we Think Science Should Work
- 4.2: Reasons for Non-Replication
- 4.3: What can we do About it?

Chapter 5: Aggregation

- 5.1: Why Summarise Data?
- 5.2: Summarising Data Using Tables
- 5.3: Summarising Data Using Graphs
- 5.4: The Middle of the Data
- 5.5: Variability - How Spread Out are the Values?

- 5.6: Z Scores

Chapter 6: Modelling Variations

- 6.1: A Simple Model
- 6.2: Statistical Modelling Using a Single Number
- 6.3: Sampling and Sampling Error
- 6.4: The Central Limit Theorem
- 6.5: Null Hypothesis Testing
- 6.6: Quantifying Effects

Chapter 7: The General Linear Model

- 7.1: General Linear Model
- 7.2: Modelling Continuous Relationships
- 7.3: Comparing Means
- 7.4: Working with Categorical Outcomes
- 7.5: Introduction to Multivariate Statistical Modelling

Chapter 8: Putting it all Together

- 8.1: Practical steps to Statistical Modelling

Chapter 9: Beyond Research and Statistics

- 9.1: Beyond Research and Statistics

Index

Glossary

Detailed Licensing

20: Accessibility Statement

Chapter 11: Versioning History

Chapter 10: Review Statement

Detailed Licensing

Licensing

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

Acknowledgements

4

This textbook represents a labour of love and a deep commitment to my students.

I would like to first thank Deborah King, Alice Luetchford and Sharon Bryan from the James Cook University Library. I came to them with an idea for a book that would incorporate the way that I see and understand statistics. They have been very supportive from the beginning. Thank you Deb for assisting with the formatting and proofreading and for being ever so patient with the whole process. However, any remaining errors are my responsibility alone. Thank you Alice for overseeing the project. I'm sure there are many behind-the-scenes tasks that I am not aware of that are involved in developing this book.

I would also like to express my gratitude to Russell Poldrack, Daniel Navarro, and Dustin Fife. My book is based largely on their existing open education resources.

I am also grateful to the Council of Australian University Librarians (CAUL) for supporting the creation of this Open Education Resource.

This book is dedicated to Michelle, who took four years of research methods and statistics in her psychology degree, and received distinctions and high distinctions in these subjects but admitted that she did not learn anything. She only knew how to click the right buttons in SPSS. I also dedicate this book to my niece and nephews, and I hope that they will develop the same interest and curiosity towards numbers and research.

Acknowledgement of Country

3

James Cook University is committed to building strong and mutually beneficial partnerships that work towards closing the employment, health and education gap for Australian Aboriginal and Torres Strait Islander peoples. Our students come from many backgrounds, promoting a rich cultural and experiential diversity on campus. We acknowledge the Aboriginal and Torres Strait Islander peoples as the Traditional Custodians of the Australian lands and waters where our staff and students live, learn and work. We honour the unique cultural and spiritual relationship to the land, waters and seas of First Australian peoples and their continuing and rich contribution to James Cook University and Australian society and we recognise that these lands have always been places of teaching and learning. We also pay respect to ancestors and Elders past and present.

Kassandra Savage (JCU Alumni), 'Coming Together and Respecting Difference', acrylic on canvas, 2014, 90cm x 90cm. © Kassandra Savage, reproduced with permission of the artist

About the Book

This book was published via the Council of Australian University Librarians Open Educational Resources Collective. The online version is available at https://oercollective.caul.edu.au/psychstats

First published in 2023 by James Cook University

eISBN: 978-0-6455878-7-6

Recommended citation:

Somoray, K. (2023). *A contemporary approach to research and statistics in psychology.* James Cook University. https://doi.org/10.25120/6xg7-djxk

About the Author

5

Dr Klaire Somoray (they/she) is a Lecturer in Psychology at the College of Healthcare Sciences, James Cook University. They completed their PhD at Queensland University of Technology, within the Centre for Accident Research and Road Safety Queensland (CARRS-Q).

Dr Somoray has a diverse research portfolio. Due to their primary expertise in statistical modelling and quantitative research methodologies, they have been involved in different research projects, including research on traffic psychology, workplace health, safety and wellbeing, and criminology.

Currently, they have a strong interest in examining the positive and negative impacts of data and technology at the individual- and societal-level.

If you would like to provide feedback on this OER, please do so by contacting Klaire by email: klaire.somoray@jcu.edu.au

Why do we Need Another Book About Research and Statistics?

6

Psychology is currently facing a crisis. In 2015, a study published by the Open Science Collaboration found that only 36% of 100 experimental and correlational studies from three top-ranking journals could be replicated.[1] This news was widely covered by media outlets around the world, including the *New York Times* and the *Atlantic*. As a wide-eyed (and very naive) undergraduate psychology student at that time, who was eager to cure humanity of its mental health issues, this was not what I had signed up for.

While some researchers argue that psychology is not in crisis,[2] it is still important to understand the reasons behind the replicability crisis and how the findings of the Open Science Collaboration study can be used to promote better scientific practices. This Open Education Resource is a humble attempt to assist future psychology students in conducting research that is robust and reliable.

It is also my hope that this resource will help ease the burden of mandatory research and statistics coursework for students who aspire to make a positive impact on people's lives through the field of psychology.

My Experience with Statistics

My relationship with statistics is unique compared to that of my peers. Many people are surprised or confused when I tell them that I love statistics. Sometimes, I am even surprised by this myself. If someone had told my 16-year-old self that I would one day be teaching statistics at the undergraduate level, I would have thought that my 30-something self had lost their mind.

I do admit that I have always loved puzzles, and I was fortunate enough to have passionate educators who taught me statistics in a fun and non-threatening manner, using resources such as Andy Field's book.[3] I viewed statistics class as another puzzle to be solved, albeit a very challenging one.

Because of this, I developed a deep curiosity about statistics and began to see it in a different light. I realised that maybe statistics isn't as difficult as it seems, or maybe, a lot of educators (even myself) make it more difficult than it needs to be. Maybe, just *maybe*, there is a way to make it more intuitive for students.

Dustin Fife, who inspired me to write this Open Education Resource, even suggested renaming statistics as "simplistics" to make it more approachable.[4]

The Philosophy Behind this Little Book

When I first started teaching undergraduate statistics, it was challenging to find a textbook that aligns with my perspective on statistics. I was also teaching intensive classes. While I loved Andy Field's textbook, I couldn't use it as a prescribed text because it required a lot of reading through stories to get to the statistical knowledge.[5] Additionally, it didn't cover certain topics, such as a brief introduction to research methods, open science principles and critiquing research. I ultimately felt that my students would be better served by a book that closely follows my philosophy of teaching research methods and statistics in psychology.

The statistics section of this book will be similar to Field's book, but the content is substantially less comprehensive and less engaging.

The goal of this book is to help future scholars of psychology conduct better science by teaching statistics in an intuitive manner. I also hope that my students will not only learn a thing or two about research and statistics, but also find it useful in their everyday life and future practice. Perhaps some students may even be inspired to become statisticians or data analysts in the future!

1. Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi.org/10.1126/science.aac4716 ↵
2. Barrett, L. F. (2015). Psychology is not in crisis. *The New York Times*. www.nytimes.com/2015/09/01/opinion/psychology-is-not-in-crisis.html ↵
3. Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage. ↵
4. quantpsych.net/stats_modeling/ ↵
5. If you think this approach will work for you, please purchase a copy of Andy Field's textbook - it is very entertaining! ↵

CHAPTER OVERVIEW

Chapter 1: Research and Statistical Thinking in Everyday Life

Learning Objectives

After reading this chapter, you should be able to:

- discuss research and statistical thinking in everyday life
- describe the central goals and fundamental concepts of statistics.

We are drowning in information, but we are starved for knowledge – John Naisbitt[1]

When was the last time you made a decision? Perhaps you decided to sleep in this morning instead of going for a run, or maybe you chose to make coffee when you woke up. Even reading this book is a decision you've made.

As you made these decisions, what thoughts were going through your mind? Maybe when you decided to sleep in, you rationalised that getting more sleep is better for you than exercising. Perhaps you decided to make coffee because it gives you a caffeine boost and helps you focus on the task at hand. And maybe you're reading this book because you want to see if my approach will help you get a better grade on your next statistics assignment.

What if I told you that statistical thinking can help you make better decisions? That may sound like a pyramid scheme, but I believe this sentiment is true.

Many people don't realise that they already use statistical thinking on a daily basis. For example, students are often adept at this when they calculate how many points they need on the next assessment to pass the overall subject.

A Note on Dealing with Anxiety Related to Statistics

Before we continue, I want to acknowledge the anxiety that many people feel when they first take a statistics class. The use of unfamiliar statistical software can also feel overwhelming. However, feeling anxious does not mean that you are bad at statistics or that you don't understand it. Statistics are not necessarily intuitive.

Anxiety can be uncomfortable, but psychology tells us that this kind of emotional arousal can actually help us perform better on tasks by focusing our attention. So, if you start to feel anxious about the materials in your statistics course, remind yourself that other students in your class are likely feeling the same way and that the arousal could actually help you perform better, even if it doesn't seem like it.

Before we delve into the details, let's first discuss what statistics can do for us.

1.1: What can Statistics do for us?
1.2: If I Apply Statistical Thinking all the Time then why do I find it Difficult?
1.3: Big Ideas in Statistics
1.4: Data is/are

1. Naisbitt, J. (1982). *Megatrends: Ten new directions transforming our lives*. Warner Books. ↵

1.1: What can Statistics do for us?

There are four main things we can do with statistics, which we already do on a regular basis:

- **Describe**: The world is complex, and we often need to simplify it to understand it.
- **Explain**: We can use data to explain a phenomenon.
- **Decide**: We also often need to make decisions based on data, usually in the face of uncertainty.
- **Predict**: We often wish to make predictions about new situations based on our knowledge of previous situations.

Statistical thinking is a tool that can help us achieve these goals. As you read through the points above, you may have noticed that we already do these behaviours in our everyday lives. Every day, we describe things (e.g., the weather is not great today), explain things (e.g., the food was salty because I put too much salt while cooking it), make decisions (e.g., I will take this route because it will be shorter), and predict things (e.g., the food in this restaurant is going to be great because it's been recommended to me by many people). The only difference with statistical thinking is that we use data as evidence for these everyday behaviours.

Let's Look at an Example

To illustrate this point, let's take a common question that many people are interested in: How do we decide what's healthy to eat? There are many different sources of guidance, from government dietary guidelines to diet books and different bloggers.

Now, let's focus on a specific question: *Is saturated fat in our diet a bad thing*?

One way to answer this question is through common sense. If we eat fat, it's going to turn into fat in our bodies, right? And we've all seen photos of arteries clogged with fat, so eating fat is going to clog our arteries, right?

Another way to answer this question is by listening to authority figures. For example, the third point under the Australian Dietary Guidelines states: "Limit intake of foods high in saturated fat such as many biscuits, cakes, pastries, pies, processed meats, commercial burgers, pizza, fried foods, potato chips, crisps and other savoury snacks."[1][2]

Finally, we might look at actual scientific research. Let's start by looking at a large study called the long-running **Prospective Urban and Rural Epidemiological** study aka the **PURE** study, which has examined diets and health outcomes (including death) in more than 135,000 people from 18 different countries. In one of the analyses of this dataset (published in The Lancet in 2017 by Dehghan et al., 2017),[3] the PURE investigators reported an analysis of how intake of various classes of macronutrients (including saturated fats and carbohydrates) was related to the likelihood of dying during the time that people were followed. People were followed for a median of 7.4 years.

Figure 1.1.1 below plots some of the data from the study (extracted from the paper), showing the relationship between the intake of both saturated fats and carbohydrates and the risk of dying from any cause.

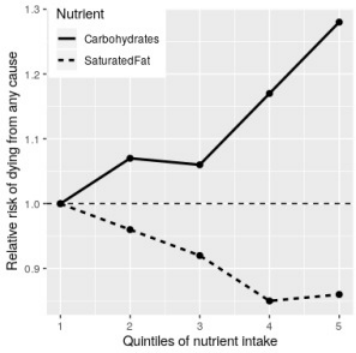


Figure 1.1.1 "A plot of data from the PURE study, showing the relationship between death from any cause and the relative intake of saturated fats and carbohydrates" by Russell A. Poldrake is licensed under CC BY-NC 4.0

This plot is based on 10 numbers. To obtain these numbers, the researchers split the group of 135,335 study participants (which we call the "sample") into five groups ("quintiles") after ordering them in terms of their intake of either of the nutrients; the first quintile contains the 20% of people with the lowest intake, and the 5th quintile contains the 20% with the highest intake. The researchers then computed how often people in each of those groups died during the time they were being followed. The figure expresses this in terms of the *relative risk* of dying in comparison to the lowest quintile: If this number is greater than 1 it means that people in that group are *more* likely to die than the people in the lowest quintile, whereas if it's less than one it means that people in that group are *less* likely to die than the people in the lowest quintile.

The figure is pretty clear: People who ate more saturated fat were *less* likely to die during the study, with the lowest death rate seen for people who were in the fourth quintile (that is, who ate more fat than the lowest 60% but less than the top 20%). The opposite is seen for carbohydrates; the more carbs a person ate, the more likely they were to die during the study. This example shows how we can use statistics to *describe* a complex dataset in terms of a much simpler set of numbers; if we had to look at the data from each of the study participants at the same time, we would be overloaded with data and it would be hard to see the pattern that emerges when they are described more simply.

The numbers in Figure 1.1.1 seem to show that *deaths decrease with saturated fat and increase with carbohydrate intake*, but we also know that there is a lot of uncertainty in the data; there are some people who died early even though they ate a low-carb diet, and, similarly, some people who ate a ton of carbs but lived to a ripe old age. Given this variability, we want to *decide* whether the relationships that we see in the data are **large enough** that we wouldn't expect them to occur randomly if there was not truly a relationship between diet and longevity.

Statistics provide us with the tools to make these kinds of decisions, and often people from the outside view this as *the* main purpose of statistics. But as we will see throughout the book, this need for black-and-white decisions based on fuzzy evidence has often led researchers astray.

We can also make predictions about future outcomes based on the available data that we have. For example, a life insurance company might want to use data about a particular person's intake of fat and carbohydrate to predict how long they are likely to live. An important aspect of prediction is that it requires us to generalise from the data we already have to some other situation, often in the future. However, if our conclusions were limited to the specific people in the study at a particular time, then the study would not be very useful. In general, researchers must assume that their particular sample is representative of a larger *population*, which requires that they obtain the sample in a way that provides an unbiased picture of the population. For example, if the PURE study had recruited all of its participants from religious sects that practice vegetarianism, then we probably wouldn't want to generalise the results to people who follow different dietary standards.

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

---

1. You can read the guidelines here: www.eatforhealth.gov.au/sites/default/files/2022-09/n55a_australian_dietary_guidelines_summary_131014_1.pdf ↵
2. You might hope that these guidelines would be based on good science, and in some cases they are, but as Nina Teicholz outlined in her book *Big Fat Surprise* (Teicholz 2014), this particular recommendation seems to be based more on the dogma of nutrition researchers than on actual evidence. ↵
3. Dehghan, M., Mente, A., Zhang, X., Swaminathan, S., Li, W., Mohan, V., ... Garcia, R. (2017). Associations of fats and carbohydrate intake with cardiovascular disease and mortality in 18 countries from five continents (PURE): A prospective cohort study. *The Lancet*, *390*(10107), 2050-2062. https://doi.org/10.1016/S0140-6736(17)32252-3 ↵

1.2: If I Apply Statistical Thinking all the Time then why do I find it Difficult?

> *If you tell students that math is logical and students don't understand something that's supposed to be logical, then students will start thinking that there must be something wrong with them – Bruce Hoskins*[1]

I hear you ask: "If I apply statistical thinking all the time – then why do I find it difficult?" Bruce Hoskins, a great educator who I look up to, argued that high school mathematics education is to blame for anxiety about statistics and numbers.[2] We have been told as young children that math is supposed to be logical. So, if students do not understand something that's supposed to be logical, then there must be something wrong with them.

Therefore, in this book, we will treat statistics as learning a new language. Think of this learning process as developing a new way of speaking and a new way of thinking. Even if you have never learned another language before, it is like visiting a new country, with different cultures and different norms. Similar to learning a new language, there are many ways to say the same things (e.g., an independent variable is also called a predictor variable).

More importantly, to be able to understand complex ideas, you need to have a solid grounding in the fundamentals. Therefore, statistics is not something you can cram at the last minute.

> *Statistics is not intuitive – Harvey Motulsky*

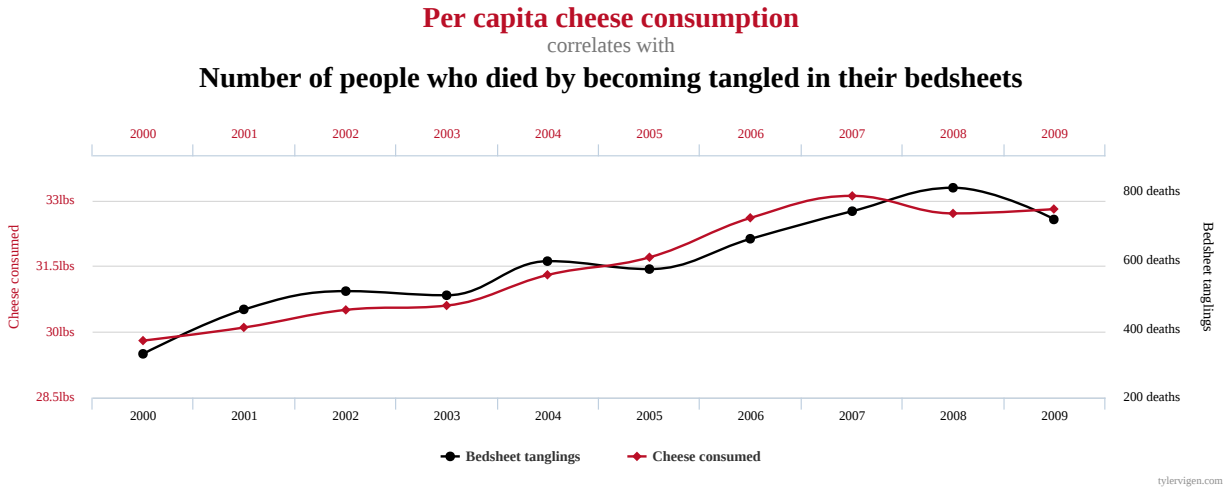
Harvey Motulsky, another great statistics educator, tells us how statistics is not intuitive. There are many biases within the human mind that often lead us astray when dealing with statistical concepts like probabilities. For example, human beings tend to jump to conclusions. My niece (who was five at that time) learned that I do boxing as a hobby and told me that, "boxing is for boys only"! She must have only seen boys to be doing hard-contact sports (like boxing) and made conclusions based on her small sample.

There is evidence to suggest that we may be hardwired to generalise from a sample to a population – even 8-month-old babies do it! (Xu & Garcia, 2008).[3] Hopefully, by the end of this book, you will understand that there are many problems with these types of generalisations. As Motulsky stated, *to avoid our natural inclination to make overly strong conclusions from limited data, scientists need to use statistics*.

> *Correlation does not imply causation – age-old statistical wisdom*

If I ever do a class in person, I would get everyone in the class to shout the above quote "Correlation does not imply causation!"

Just because two variables are correlated and the difference is statistically significant (we will learn more about this later), it does not mean that changes in the X variable caused the changes in the Y variable. Let's look at the following chart:



The above chart shows that cheese consumption has a positive relationship with the number of people who died from being tangled by their bedsheets. In other words, as cheese consumption increases, more people die from bedsheet entanglement.

Ummm… Is cheese to blame for people dying on their beds?

The figure above is an example of spurious correlation. A spurious correlation, also known as a false correlation or a coincidental correlation, is a relationship between two variables that appear to be related, but in reality, the relationship is not causal. In other words, the correlation between the two variables is a random occurrence that is not caused by any underlying mechanism or factor.

While the above example is a bit of a ridiculous one, we sometimes make the mistake that, just because two variables are highly correlated, we assume that one of the variables can cause changes to another. Let's go back to the PURE study.

Causality and Statistics

The PURE study seemed to provide pretty strong evidence for a positive relationship between eating saturated fat and living longer, but this doesn't tell us what we really want to know: If we eat more saturated fat, will that cause us to live longer? This is because we don't know whether there is a direct causal relationship between eating saturated fat and living longer. The data is consistent with such a relationship, but it is equally consistent with some other factor causing both higher saturated fat and longer life. For example, it is likely that people who are richer eat more saturated fat and richer people tend to live longer, but their longer life is not necessarily due to fat intake — it could instead be due to better health care, reduced psychological stress, better food quality, or many other factors. The PURE study investigators tried to account for these factors, but we can't be certain that their efforts completely removed the effects of other variables. The fact that other factors may explain the relationship between saturated fat intake and death is an example of why introductory statistics classes often teach that "correlation does not imply causation", though the renowned data visualisation expert Edward Tufte has added, "but it sure is a hint."

Although observational research (like the PURE study) cannot conclusively demonstrate causal relations, we generally think that causation can be demonstrated using studies that experimentally control and manipulate a specific factor. In medicine, such a study is referred to as a *randomised controlled trial* (RCT). Let's say that we wanted to do an RCT to examine whether increasing saturated fat intake increases life span. To do this, we would sample a group of people, and then assign them to either a treatment group (which would be told to increase their saturated fat intake) or a control group (who would be told to keep eating the same as before). It is essential that we assign the individuals to these groups randomly. Otherwise, people who choose the treatment might be different in some way than people who choose the control group – for example, they might be more likely to engage in other healthy behaviours as well. We would then follow the participants over time and see how many people in each group died. Because we randomised the participants to treatment or control groups, we can be reasonably confident that there are no other differences between the groups that would *confound* the treatment effect; however, we still can't be certain because sometimes randomisation yields treatment versus control groups that *do* vary in some important way. Researchers often try to address these confounds using statistical analyses, but removing the influence of a confound from the data can be very difficult.

A number of RCTs have examined the question of whether changing saturated fat intake results in better health and longer life. These trials have focused on *reducing* saturated fat because of the strong dogma among nutrition researchers that saturated fat is deadly; most of these researchers would have probably argued that it was not ethical to cause people to eat *more* saturated fat! However, the RCTs have shown a very consistent pattern: Overall there is no appreciable effect on death rates of reducing saturated fat intake.

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

---

1. Hoskins, B. (Host). (2020, October 29). *(Re)Teach: Teaching statistics pt 1* [Audio podcast]. https://reteach.buzzsprout.com/428977/6117973-teaching-statistics-pt1 ↵
2. I would like to clarify that this comment is more about the systemic issues we face rather than blaming individual math teachers. Teachers, I know you're already underfunded, stressed and constantly burned out, I don't want to place this burden on you! ↵
3. Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012-5015. ↵

1.3: Big Ideas in Statistics

Now that you've understood some of the problems we usually come across when conducting statistics and interpreting the results, let's look at the main ideas that cut through nearly all aspects of statistical thinking. These ideas will be threaded throughout the whole book.

Before going into the "meat" of the book, introducing these main ideas can help you on your path to understanding statistics. My goal for you is to not just *remember* concepts but to *understand* them. By giving you the fundamental ideas first, hopefully, it will help you organise the knowledge you will learn in this book and make them more flexible and powerful.

Several of these ideas came from Ji Y. Son, a learning scientist who co-wrote a fantastic interactive textbook, *Introduction to Statistics: A Modelling Approach* and James Stigler's (2016) outstanding book *The Seven Pillars of Statistical Wisdom*, which are augmented here.

Big Idea 1: Aggregation

One way to think of statistics is "the science of throwing away data". In the example of the PURE study previously, we took more than 100,000 numbers and condensed them into 10. This kind of *aggregation* is one of the most important concepts in statistics. However, you maybe asking yourself: If we throw out all of the details about every one of the participants, then how can we be sure that we aren't missing something important?

Statistics offers us ways to describe the structure of aggregated data, backed by theoretical principles that explain its effectiveness. However, it's crucial to remember that aggregation can go too far and later we will encounter cases where a summary can provide a misleading picture of the data being summarised.

Big Idea 2: Variation

Statistics can also be thought of as "the study of variation". If variation didn't exist, then we wouldn't need statistics.

But variation is everywhere. If everyone experiencing chronic pain took a particular drug and their symptoms improve and those who didn't take the drug got worse, then we wouldn't need statistics. However, this scenario rarely occurs. Typically, some individuals who take the drug improves, while others do not. Conversely, some individuals who don't take the drug can recover. It can be challenging to determine if the drug truly cures the ailment or if the recovery is simply a coincidence. Statistics helps us understand such scenarios by providing a set of concepts and tools that have developed over centuries to detect patterns and make sense of variation.

Big Idea 3: Modelling

When my niece, Bella, turned 1 year old, my brother and I bought her a 6V battery-powered Frozen toy car that was advertised for 18 months and up. She had to wait another 8 months to actually ride her toy car. This toy car is a model of a real car. It has nearly all the components to make it a car – it has four wheels, a steering wheel, a pedal, side mirror. You can basically drive it like you would a real car (but you need to be 5 or under, otherwise you would not fit. Trust me, I tried). One could argue that my niece's toy car is a good 'model' of a real car because this toy car was constructed using existing information about a real car.

As quantitative social scientists, we can also build (statistical) models of *real-world processes in an attempt to predict these processes in certain conditions.*[1] Unlike ride-on toy makers, however, social scientists are mostly working with *non-physical constructs* and therefore, we can only make *inferences* about the psychological process that our models are based upon. Like ride-on toy makers, we do still want our models to be representative of reality (or at least close to it). The closer the model is to reality, the better the *fit* of the model. We want to build a model that has *a better fit* (i.e., closer to reality), because if we use this model to make predictions about the real world then, we can be more confident that these predictions will be accurate.

Modelling is an important concept that we will touch on over and over again in this book.

Big Idea 4: Uncertainty

The world is an uncertain place. We now know that cigarette smoking causes lung cancer, but this causation is probabilistic: A 68-year-old man who smoked two packs a day for the past 50 years and continues to smoke has a 15 per cent (1 out of 7) risk of

getting lung cancer, which is much higher than the chance of lung cancer in a nonsmoker. However, it also means that there will be many people who smoke their entire lives and never get lung cancer. Statistics provides us with the tools to characterise uncertainty, to make decisions under uncertainty, and to make predictions whose uncertainty we can quantify.

One often sees journalists write that scientific researchers have "proven" some hypothesis. But statistical analysis can never "prove" a hypothesis, in the sense of demonstrating that it must be true (as one would in a logical or mathematical proof). Statistics can provide us with evidence, but it's always tentative and subject to the uncertainty that is always present in the real world.

Big Idea 5: Sampling from a Population

The concept of aggregation implies that we can make useful insights by collapsing across data – but how much data do we need? The idea of sampling says that we can summarise an entire population based on just a small number of samples from the population, as long as those samples are obtained in the right way. As we already discussed above, the way that the study sample is obtained is critical, as it determines how broadly we can generalise the results. Another fundamental insight about sampling is that while larger samples are always better (in terms of their ability to accurately represent the entire population), there are diminishing returns as the sample gets larger. In fact, the rate at which the benefit of larger samples decreases follows a simple mathematical rule, growing as the square root of the sample size, such that to double the quality of our data we need to quadruple the size of our sample.

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

---

1. Field, A. (2017). *Discovering statistics using IBM SPSS statistics*. Sage. ↵

1.4: Data is/are

Before we go any further, a note on the word data. While working as a data analyst for the state government, my colleagues and I had an argument on whether we say, data *is* or data *are* when finalising a report. This is an old debate that comes up every now and again among people who deal with data on a regular basis.

The American Psychological Association's (APA) style manual treats data as "plural" in its strict form, as the word datum is singular and the word data is plural.[1] However, I take the modern approach of saying "the data is" and my reason is that *no one really uses the singular form anymore*.[2]

There are other things that we should be focusing on though. Like, what is the nature of the data? Is it qualitative? Quantitative? How is it measured?

Data are composed of variables, where a variable reflects a unique measurement or quantity. In the following chapters, we will be talking about different types of data that we will come across.

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, under a CC BY-NC 4.0 licence.

1. APA Style Blog. (2012). *Data is/are*. American Psychological Association. blog.apastyle.org/apastyle/2012/07/data-is-or-data-are.html ↵
2. Data is or data are? (2010). *The Guardian*. https://www.theguardian.com/news/datablog/2010/jul/16/data-plural-singular ↵

CHAPTER OVERVIEW

Chapter 2: Working with jamovi

Learning Objectives

After reading this chapter, you should be able to:

- install jamovi on your computer
- get familiar with jamovi.

In this section, we'll discuss how to get started in jamovi. First, we will download and install jamovi, but most chapters will focus on how you can interact with the jamovi interface. You will not be learning any statistical concepts here, just learning the basics of how to use the statistical software. Therefore, we will spend some time looking at datasets and variables. This way, you can get a feel for what it's like to work with jamovi.

2.1: Why jamovi?
2.2: Getting Started with jamovi
2.3: Analyses
2.4: The Spreadsheet
2.5: Loading Data in jamovi
2.6: Installing add-on Modules into jamovi

2.1: Why jamovi?

The open-source (meaning free!) software, jamovi, offers statistical analysis. It is user-friendly and intuitive but sophisticated enough to allow for advanced analyses and sophisticated data transformation and recording.

It is now being used in a lot of undergraduate statistics programs and is being adopted by many universities.

How is it Different from SPSS and Other Statistical Software out There?

SPSS, developed by IBM, is a very common statistical software that many universities use for their programs. I went through my undergraduate program using SPSS, and while I am well-versed with this software, its pricing makes me wince. As an educator who promotes open knowledge, getting students to pay for statistical software does not align with my values.

The business model for SPSS and most statistical software companies is to sell "student versions" of their software at a low price and then charge a high price for "educational versions" and even higher prices for commercial licenses (Navarro and Foxcroft, 2022).[1] This can lead to students becoming reliant on these tools and feeling obligated to continue paying high fees after they graduate. One way to avoid this is by using open-source software like jamovi, which is free and does not require payment of licensing fees.

How is it Different from R (a Statistical Programming Language)?

I know some students would be horrified if I asked them to code AND learn statistics. While I believe that coding and programming are the best ways to learn statistics, I don't want to add to the mental load that students already experience when learning statistics.

As a graphical statistical spreadsheet, jamovi is different from R, which is a programming language. However, jamovi and R are compatible because the analyses in jamovi are written in R. In fact, users can switch to "syntax mode" in jamovi to view the equivalent R code for the analyses, or they can use the Rj editor to type and run R code directly within the spreadsheet. This feature is useful for those who want to transition from using a spreadsheet to learning R. Overall, jamovi is a good starting point for those who prefer a graphical spreadsheet but are interested in learning R.

To change to syntax mode, select the Application menu at the top right of jamovi (a button with three vertical dots) and click the "Syntax mode" checkbox there. You can turn off syntax mode by clicking this a second time.

In syntax mode, analyses continue to operate as before but now they produce R syntax and 'ASCII output' like an R session. Like all results objects in jamovi, you can right-click on these items (including the R syntax) and copy and paste them, for example, into an R session. At present, the provided R syntax does not include the data import step, so this must be performed manually in R. There are many resources explaining how to import data into R and if you are interested we recommend you take a look at these – just search on the interweb.

Chapter attribution

This chapter contains material taken and adapted from *The jamovi project* by Sebastian Jentschke, used under a CC BY-NC 4.0 licence.

1. Navarro, D. J., & Foxcroft, D. R. (2022). *Learning statistics with jamovi: A tutorial for psychology students and other beginners* (Version 0.75). https://doi.org/10.24384/hgc3-7p15 ↵

2.2: Getting Started with jamovi

Before you start, there are a few things you need to do to get ready. First, make sure you have jamovi installed on your computer and that it is ready to go.

You can still use this book if you have other statistical software (like SPSS or R) but the steps and outputs presented here are from jamovi.

Windows

The jamovi software is available for Windows Vista (64-bit) and above. Installation on Windows is quite straightforward and should be familiar to anyone who has installed software on Windows before. Download the latest version from the jamovi website.

At some institutions (particularly universities, or if you are using your work computer), the "normal" approach to installing software is blocked by IT security policies. If you cannot install the software raise the issue with your IT staff.

macOS

The 1.6 series of jamovi (and newer) are available for macOS 10.13 (High Sierra) and newer. To install jamovi on macOS, download the .dmg file from the jamovi website. This will present you with a "file view", and you can install jamovi by "dragging-and-dropping" the jamovi application to the Applications folder (in the usual way). After this, jamovi will appear in your applications and can be started like any other installed application.

Other Operating Systems

For other operating systems, please visit the user manual page of the jamovi website.

Getting Started

When first starting jamovi, you will be presented with a user interface that looks something like the figure below:

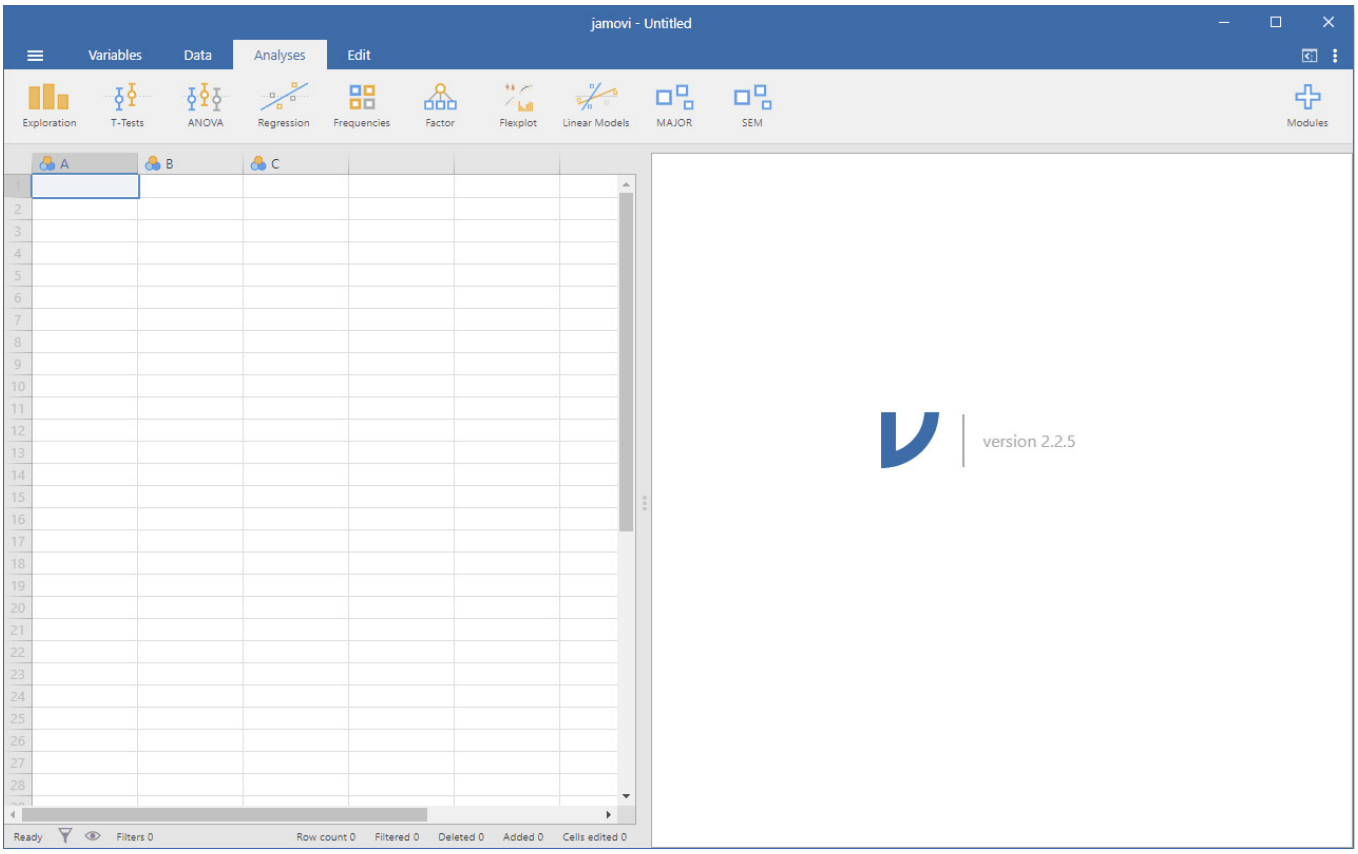


Figure 2.2.1. The jamovi start up screen. To the left is the spreadsheet view, and to the right is where the results of statistical tests appear. Down the middle is a bar separating these two regions and this can be dragged to the left or the right to change their sizes.

It is possible to simply begin typing values into the jamovi spreadsheet as you would in any other spreadsheet software. There are also several sample data sets available in jamovi (you can find them under the data library tab).

Alternatively, existing data sets in the CSV (.csv) file format can be opened in jamovi. Additionally, you can easily import SPSS, SAS, Stata and JASP files directly into jamovi. We will talk more about this later on, but in short, you can open a file by selecting the File tab (three horizontal lines signify this tab) at the top left-hand corner, selecting “Open” and then choosing from the files listed on “Browse” depending on whether you want to open an example or a file stored on your computer.

Chapter attribution

This chapter contains material taken and adapted from *The jamovi project* by Sebastian Jentschke, used under a CC BY-NC 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

2.3: Analyses

To begin your analysis, select **Analyses** from the top menu. From there, choose the type of statistical analysis you wish to perform, such as t-tests, ANOVA and regression (as shown in the figure below). Upon selecting **Analysis**, an "options panel" specific to that analysis will become available. This panel allows you to allocate different variables to different parts of the analysis and choose different settings. As you make changes, the results of your analysis will appear in real time on the right-hand "Results panel".

When you have the analysis set up correctly, you can dismiss the analysis options by clicking the arrow to the top right of the options 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. The analysis can be removed by right-clicking on the analysis results to bring up a menu, and then selecting **Analysis**, and then **Remove.**

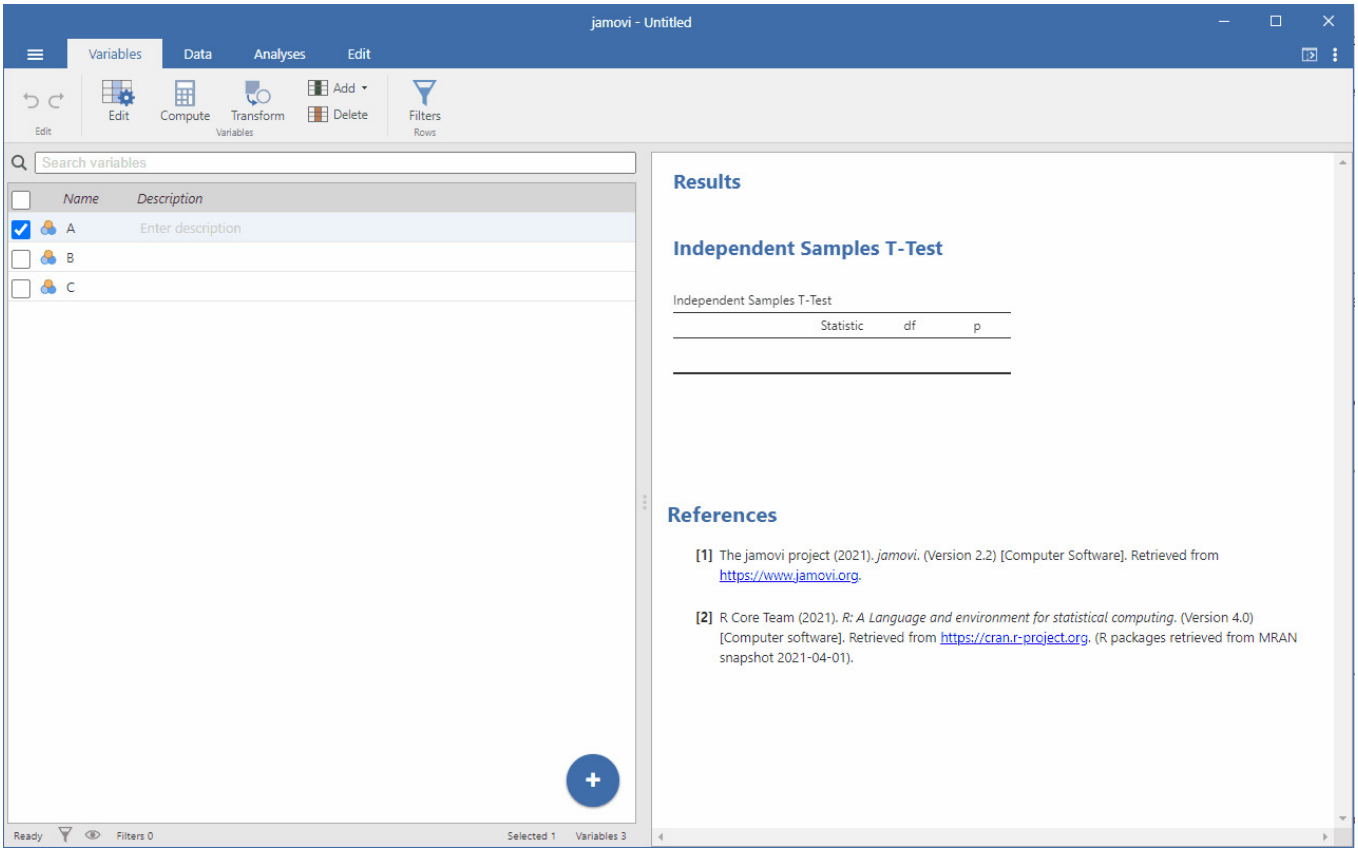


Figure 2.3.1 *Navigating the analyses menu*

Copy and Paste Feature

The jamovi program 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 right-click on the object of interest and from the menu select exactly what you want to copy. The menu allows you to choose to copy only the image or the entire analysis. 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.

Chapter attribution

This chapter contains material taken and adapted from *The jamovi project* by Sebastian Jentschke, used under a CC BY-NC 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

2.4: The Spreadsheet

In jamovi, data is represented in a spreadsheet with each column representing a “variable” and each row representing a “case” or “participant”.

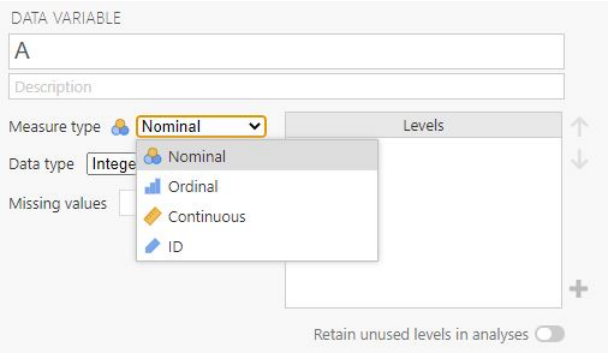


Figure 2.4.1. Data variables and different measurement levels in jamovi

These levels are designated by the symbol in the header of the variable’s column. Let’s discuss each measurement level below:

- The *ID* variable type is unique to jamovi. It’s intended for variables that contain identifiers that you would rarely want to analyse, for example, a person’s name, or a participant ID. Specifying an ID variable type can improve performance when interacting with very large data sets.
- *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, as would a person’s name. Nominal variable values can also have a numeric value. These variables are used most often when importing data that 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 of this is ranking (e.g., first place, second place, last place).
- *Continuous* variables are variables that exist on a continuous scale. Examples might be height or weight. This is also referred to as the “Interval” or “Ratio scale”.

In addition, you can also specify different data types: variables have a data type of either “Text”, “Integer” or “Decimal”.

Changing Data from One Level 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.

Earlier 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 jamovi data view for that variable. Just click the check box for the measurement level you want – continuous, ordinal, or nominal.

When starting with a blank spreadsheet and typing values in, the variable type will change automatically depending on the data you enter. This is a good way to get a feel for which variable types go with which sorts of data. Similarly, when opening a data file, jamovi 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.

The variable editor can be opened by selecting “Setup” from the data tab or by double-clicking on the variable column header. The variable editor allows you to change the name of the variable and, for data variables, the variable type, the order of the levels, and the label displayed for each level. Changes can be applied by clicking the “tick” at the top right. The variable editor can be dismissed by clicking the “Hide” arrow.

New variables can be inserted or appended to the data set using the “add” button from the data ribbon. The “add” button also allows the addition of computed variables.

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 “add” button available on the data tab. This will produce a formula box where you can specify the formula. The usual arithmetic operators are available. Some examples of the common formulas we use in psychological research include:

Sum of the variables to get a total score = Variable1 + Variable2 + Variable3

Average of the variables to get an average score = Mean(Variable1, Variable2, Variable3)

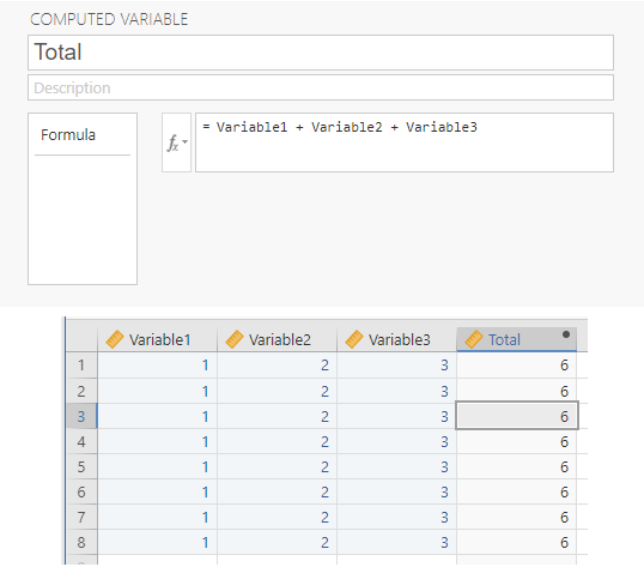


Figure 2.4.2. *A newly computed variable*

Filter

The filtering function in jamovi enables you to eliminate unwanted rows from your analysis. This includes filtering by specific criteria, such as only including survey responses from individuals who provided consent for their data to be used, excluding a specific age group, for example. Alternatively, you may choose to filter out extreme scores, such as those more than three standard deviations from the mean. These filters are constructed using jamovi’s computed variable formula system, which permits the creation of highly intricate formulas.

Chapter attribution

This chapter contains material taken and adapted from *Learning statistics with jamovi* by Danielle J. Navarro and David R. Foxcroft, used under a CC BY-SA 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

2.5: Loading Data in jamovi

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:

- jamovi files are those with a .omv file extension. This is the standard kind of file that jamovi 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.

There are also several other kinds of data files that you might want to import into jamovi. For instance, you might want to open Microsoft Excel spreadsheets (.xls files), or data files that have been saved in the native file formats for other statistics software, such as SPSS or SAS. Whichever file formats you are using, it's a good idea to create a folder or folders especially for your jamovi data sets and analyses and to make sure you keep these backed up regularly.

Loading 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. The .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 2.5.1, which shows a file called booksales.csv. 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.

	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

Figure 2.5.1. The booksales.csv data file

Figure 2.5.2. Opening the booksales.csv in jamovi

Loading SPSS Files

Since SPSS is probably the most widely used statistics package in psychology, it's worth mentioning that jamovi can also import SPSS data files (file extension .sav). Just follow the instructions above for how to open a .csv file, but this time navigate to the .sav file you want to import. For SPSS files, jamovi will regard all values as missing if they are regarded as "system missing" files in SPSS. The "Default missings" value does not seem to work as expected when importing SPSS files, so be aware of this.

Loading Excel Files

The way to handle Excel files is to open them up first in Excel or another spreadsheet program that can handle Excel files, and then export the data as a .csv file before opening/importing the .csv file into jamovi.

Chapter attribution

This chapter contains material taken and adapted from *Learning statistics with jamovi* by Danielle J. Navarro and David R. Foxcroft, used under a CC BY-SA 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

2.6: Installing add-on Modules into jamovi

A really great feature of jamovi is the ability to install add-on modules from the jamovi library. These add-on modules have been developed by the jamovi community, that is, jamovi users and developers who have created special software add-ons that do other, usually more advanced, analyses that go beyond the capabilities of the base jamovi program.

To install add-on modules, just click on the large plus sign (+) in the top right of the jamovi window, select "jamovi-library" and then browse through the various available add-on modules. Choose the one(s) you want, and then install them, as in the figure below. It's that easy. The newly installed modules can then be accessed from the "Analyses" button bar. Try it!

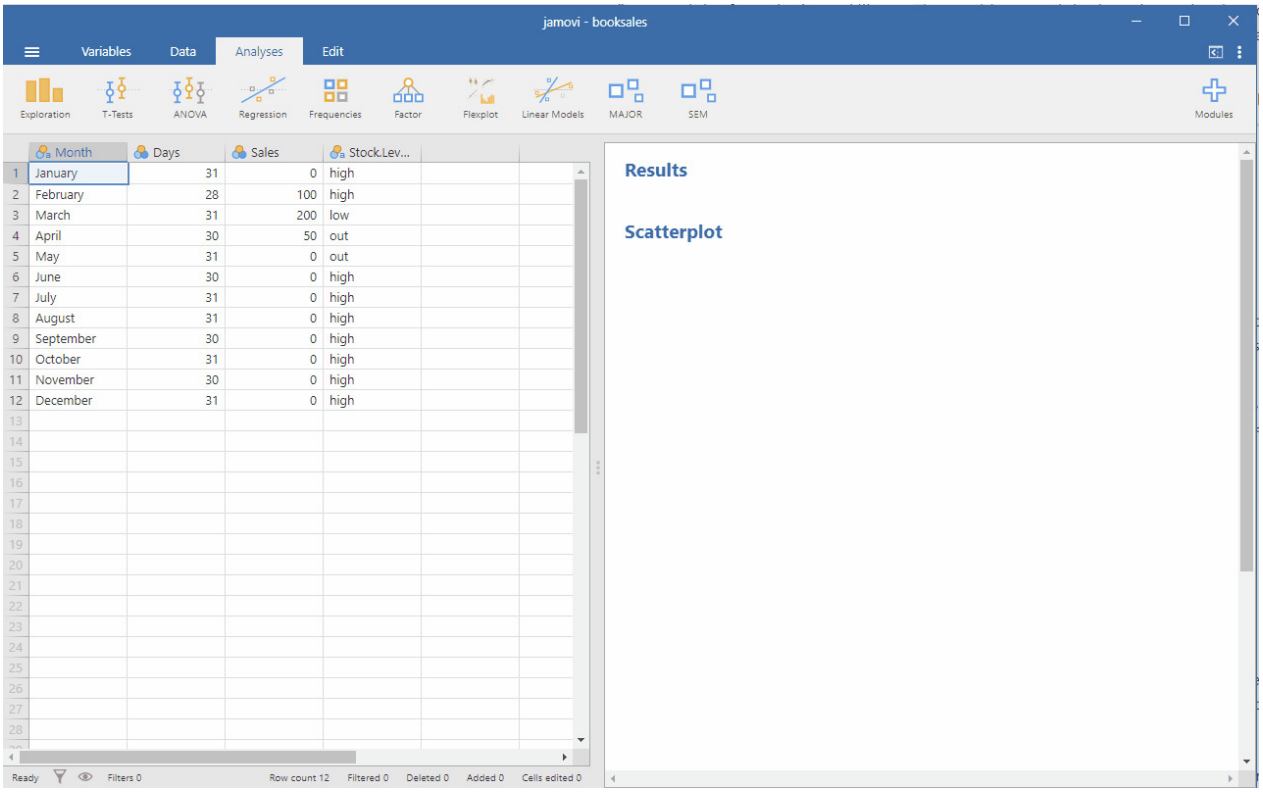


Figure 2.6.1. Installing add-on modules in jamovi

Chapter attribution

This chapter contains material taken and adapted from *Learning statistics with jamovi* by Danielle J. Navarro and David R. Foxcroft, used under a CC BY-SA 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

CHAPTER OVERVIEW

Chapter 3: Brief Review of Research Methods

Learning Objectives

After reading this chapter, you should be able to:

- distinguish between different types of variables (quantitative/qualitative, binary/integer/real, discrete/continuous) and give examples of each of these kinds of variables
- distinguish between the concepts of reliability and validity and apply each concept to a particular dataset.

Until now, our discussions have predominantly centred on statistics — however, it's also important to have a good understanding of research methodology to conduct effective statistical analysis. As one of my favourite lecturers in statistics used to say: "garbage in, garbage out".

"Garbage in, garbage out" means that if your data is bad, your results will be bad. It's like cooking a fancy dish with spoiled ingredients — no matter how skilled the chef, the dish will still taste bad. In statistics, it's crucial to have accurate and reliable data to get valuable insights and conclusions and research design can play a big role in getting this accurate and reliable data for your analysis.

Research design is just as critical as data analysis, and this book will briefly cover research methods that you will mostly encounter in psychological research. Statistics provides a universal set of core tools useful for most types of research, but research methods are not as universal. While there are general principles to consider, much of research design is specific to the area of discipline. Therefore, we will only consider the general principles that we often see in psychological research.

3.1: How do we Measure Variables in Psychology?
3.2: Introduction to Psychological Measurement
3.3: What Makes a Good Measure?
3.4: Some Complexities
3.5: The Role of Variables - Predictors and Outcomes
3.6: Research Design I- Experimental Designs
3.7: Research Design II- Non-Experimental Designs

3.1: How do we Measure Variables in Psychology?

To students who are just getting started in psychological research, the challenge of measuring such variables might seem insurmountable. Is it really possible to measure things as intangible as self-esteem, mood, or an intention to do something? The answer is a resounding yes, in the next few chapters, we will talk about the nature of the variables that psychologists study and how they can be measured.

How dark is your personality?

According to personality theorists, there are several characteristics that reflect the dark side of our personality. A combination of these characteristics is called the dark triad. The dark triad is a relatively new concept in personality psychology. Conceptualised by Paulhus and Williams (2002), the dark triad is comprised of anti-social characteristics of Machiavellianism, Narcissism, and Psychopathy. The following is an example of how we can measure the dark triad:

The dark triad is assessed using a scale called the Dirty Dozen (Jonason & Webster, 2010)[1] which has been designed for use with a non-clinical population. Jonason and Webster's measure asks people to rate themselves against these questions (on a scale of 1 to 7):

1. I tend to manipulate others to get my way.
2. I have used deceit or lied to get my way.
3. I have used flattery to get my way.
4. I tend to exploit others towards my own end.
5. I tend to lack remorse.
6. I tend to not be too concerned with morality or the morality of my actions.
7. I tend to be callous or insensitive.
8. I tend to be cynical.
9. I tend to want others to admire me.
10. I tend to want others to pay attention to me.
11. I tend to seek prestige or status.
12. I tend to expect special favours from others.

The total scores range from 12 to 84. What did you get?

1. Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: A concise measure of the dark triad. *Psychological Assessment*, *22*(2), 420–432. doi.org/10.1037/a0019265 ↵

3.2: Introduction to Psychological Measurement

Research starts with identifying **what** you want to learn, and then determining **how** you plan to study it. Therefore, we will start our brief introduction to research methods with variables and psychological measurement.

What do we Mean by Psychological Measurement?

Measurement is the assignment of scores to individuals so that the scores represent some characteristic of the individuals. This very general definition is consistent with the kinds of measurement that everyone is familiar with — for example, weighing oneself by stepping onto a bathroom scale or checking the internal temperature of a roasting turkey using a meat thermometer.

This general definition of measurement is consistent with measurement in psychology too. You may imagine a clinical psychologist who is interested in how depressed a person is. He administers the Beck Depression Inventory, which is a 21-item self-report questionnaire in which the person rates the extent to which they have felt sad, lost energy, and experienced other symptoms of depression over the past two weeks. The sum of these 21 ratings is the score and represents the person's current level of depression.

Variables that we Study in Psychological Research

Many variables studied by psychologists are straightforward and simple to measure. These include age, height, weight and birth order. You can ask people how old they are and be reasonably sure that they know and will tell you. Other variables studied by psychologists — perhaps the majority — are not so straightforward or simple to measure. We cannot accurately assess people's level of intelligence by looking at them, and we certainly cannot put their self-esteem on a bathroom scale to measure it. We sometimes call these variables "constructs".

Psychological constructs are difficult to observe directly. One reason is that they often represent tendencies to think, feel or act in certain ways. Often, these constructs often involve internal processes. For example, to say that a particular university student is highly extroverted does not necessarily mean that she is behaving in an extroverted way right now. In fact, she might be sitting quietly by herself, reading a book. Instead, it means that she has a general tendency to behave in extraverted ways (e.g., being outgoing, enjoying social interactions etc.) across a variety of situations.

How do we Measure These Psychological Constructs?

Even though psychological constructs are difficult to measure, we still try anyway. We call this process **operationalisation**. We define and explain variables in terms of how they will be measured. In other words, "operationalisation is the process by which we take a meaningful but somewhat vague concept and turn it into a precise measurement" (Navarro and Foxtrot, 2022).[1] Navarro and Foxtrot (2022) also provide the process of operationalisation, which involves:

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

> *the set of possible values to include all verbal response, how will you interpret their answers?*

Let's focus on the second point, determining what method you will use to measure your variables. Methods of measurement generally fall into one of three broad categories. Self-report measures are those in which participants report their own thoughts, feelings, and actions, such as the Beck Depression Inventory. Behavioural measures are those in which some other aspect of participants' behaviour is observed and recorded. This is an extremely broad category that includes the observation of people's behaviour both in highly structured laboratory tasks and in more natural settings. Finally, physiological measures are those that involve recording any of a wide variety of physiological processes, including heart rate and blood pressure, galvanic skin response, hormone levels, and electrical activity and blood flow in the brain.

Levels of Measurement

Now, let's further discuss the third point. The psychologist S. S. Stevens suggested that scores can be assigned to individuals in a way that communicates more or less quantitative information about the variable of interest (Stevens, 1946).[2] For example, the officials at a 100 metre race could simply rank order the runners as they crossed the finish line (first, second, etc.), or they could time each runner to the nearest tenth of a second using a stopwatch (11.5 sec, 12.1 sec, etc.). In either case, they would be measuring the runners' times by systematically assigning scores to represent those times. Stevens provided a framework for categorising variables based on their level of measurement or amount of information that they provide, which is called scales of measurement (also known as levels of measurement). The four levels are nominal, ordinal, interval and ratio, which we will discuss further below.

Nominal Variable

Nominal level of measurement is one in which variables are classified based on their names or categories. They do not have any order, ranking or mathematical significance. For example, gender (male, female), hair colour (black, brown, blonde), and country of origin are all examples of nominal level variables.

Ordinal Variable

Ordinal level of measurement involves variables that have an inherent order or ranking, but the difference between values is not necessarily equal. For example, social class (lower, middle, upper), education level (primary, high school, university) and levels of satisfaction (unsatisfied, neutral, satisfied) are all examples of ordinal variables. They can be arranged in a sequence, but the difference between the values is not meaningful in mathematical terms.

Unlike nominal scales, ordinal scales allow comparisons of the degree to which two individuals rate a particular variable. For example, we know that unsatisfied would be a lower rating of satisfaction compared to neutral.

Interval Variable

An interval scale has all of the features of an ordinal scale, but in addition, the intervals between units are meaningful. A standard example is physical temperature measured in Celsius or Fahrenheit; the physical difference between 10 and 20 degrees is the same as the physical difference between 90 and 100 degrees, but each scale can also take on negative values.

Interval variables **do not have a true zero point**. The Celsius scale illustrates this issue. Zero degrees Celsius does not represent the complete absence of temperature (the absence of any molecular kinetic energy). In psychology, the intelligence quotient (IQ) is often considered to be measured at the interval level. While it is technically possible to receive a score of 0 on an IQ test, such a score would not indicate the complete absence of IQ.

Ratio Variable

Finally, the ratio level of measurement assigns scores with a **true zero point** that represents the complete absence of the quantity being measured. Examples include height measured in metres and weight measured in kilograms. This level also applies to counts of discrete objects or events, such as the number of siblings one has or the number of questions answered correctly on an exam.

Why is this important?

Why does it matter whether a variable is nominal, ordinal, interval or ratio?

The most important takeaway from this is that some calculations would not make sense on some types of data. For example, imagine that we were to collect postal code data from 200 people in living in Brisbane, Australia. Even though postal codes may look like an interval variable, they don't actually refer to a numeric scale. Each postcode basically serves as a label for a different region. For this reason, it wouldn't make sense to talk about the average postal code, for example.

Furthermore, statistical software like jamovi assumes that the variables we are trying to analyse will have a specific measure. For instance, it would not make sense to compute an average of people's hair colour (given that this is a nominal variable). For the most part however, researchers just mostly care about distinguishing between nominal variables and all the others.

As shown in the previous chapter, there are only three measures that you can choose for your variables in jamovi: continuous (which can be used for ratio and interval variables), ordinal and nominal. See the image below for the options available in jamovi.

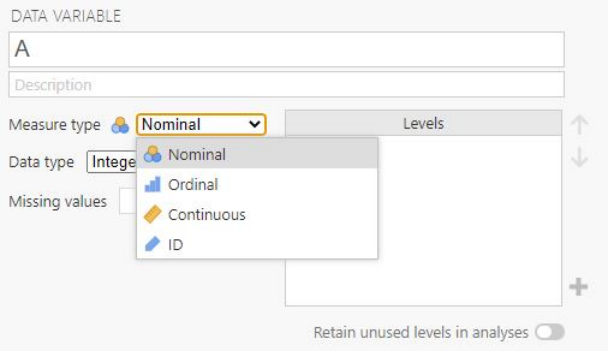


Figure 3.2.1. Data variables and different measurement levels in jamovi

Discrete Versus Continuous Measurements

There's a second kind of distinction that you need to be aware of regarding what types of variables you can run into. The difference between these is as follows:

- A ***discrete*** measurement is one that takes one of a set of particular values. These could be qualitative values (for example, different breeds of dogs) or numerical values (for example, how many friends one has on Facebook). Importantly, there is no middle ground between the measurements; it doesn't make sense to say that one has 33.7 friends.
- A ***continuous*** measurement is one that is defined in terms of a real number. It could fall anywhere in a particular range of values — for instance, response time is continuous. If Bella takes 3.1 seconds and Isaac takes 2.3 seconds to respond to a question, then Gabby's response time will lie in between if he took 3.0 seconds to respond.

Chapter attribution

This chapter contains taken and adapted material from several sources:

- *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.
- *Research methods in psychology* by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler and Dana C. Leighton, used under a CC BY-NC-SA 4.0 licence.
- Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

1. Navarro, D. J., & Foxcroft, D. R. (2022). *Learning statistics with jamovi: A tutorial for psychology students and other beginners* (Version 0.75). https://doi.org/10.24384/hgc3-7p15 ↵
2. Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680. ↵

3.3: What Makes a Good Measure?

It's usually impossible to measure a construct without some error. For example, you may know the answer but misread the question and get it wrong. In other cases, the error is intrinsic to the thing being measured, such as in a simple reaction time test where the time it takes a person to respond can vary from trial to trial for various reasons. We generally strive to minimise measurement error as much as possible.

There are times when there's a "gold standard" against which other measurements can be compared to. For example, sleep can be measured using various devices, such as those that measure movement in bed, but they are considered inferior to the gold standard of polysomnography, which uses brain waves to quantify the time a person spends in each stage of sleep. However, the gold standard is often more difficult or expensive to perform, so a cheaper method is used even though it may have greater error.

When evaluating the quality of a measurement, we generally distinguish between two different aspects: **reliability** and **validity**. Put simply, the reliability of a measure tells you how precisely you are measuring something, whereas the validity of a measure tells you how accurate the measure is.

Reliability

Reliability refers to the consistency of our measurements. One form of reliability is **test-retest reliability**, which measures how well the measurements agree if the same measurement is performed twice. For example, if a questionnaire is given to a person about their attitude towards statistics today, and the same questionnaire is repeated tomorrow, it would be expected that the answers would be similar, unless something significant changed the person's view of statistics (like reading this book!).

Assessing test-retest reliability requires using the measure on a group of people at one time, using it again on the *same* group of people at a later time, and then looking at the **test-retest correlation** between the two sets of scores. This is typically done by graphing the data in a scatterplot and computing the correlation coefficient.

Another way to assess reliability is in cases where the data includes subjective judgments. For example, let's say that a researcher wants to determine whether a treatment changes how well a child with Attention Deficit Hyperactivity Disorder (ADHD) interacts with other children, which is measured by having experts watch the child and rate their interactions with the other children. In this case, we would like to make sure that the answers don't depend on the individual rater — that is, we would like for there to be high **inter-rater reliability**. This can be assessed by having more than one rater perform the rating, and then comparing their ratings to make sure that they agree well with one another.

Another kind of reliability is **internal consistency,** which is the consistency of people's responses across the items on a multiple-item measure. In general, all the items on such measures are supposed to reflect the same underlying construct, so people's scores on those items should be correlated with each other. On the dark triad personality test, if people's responses to the different items are not correlated with each other, then it would no longer make sense to claim that they are all measuring the same underlying construct. This is as true for behavioural and physiological measures as for self-report measures.

Like test-retest reliability, internal consistency can only be assessed by collecting and analysing data. One approach is to look at a **split-half correlation**. This involves splitting the items into two sets, such as the first and second halves of the items or the even- and odd-numbered items. Then a score is computed for each set of items, and the relationship between the two sets of scores is examined. A split-half correlation of +.80 or greater is generally considered good internal consistency.

Perhaps the most common measure of internal consistency used by researchers in psychology is a statistic called Cronbach's α (the Greek letter alpha). Conceptually, α is the mean of all possible split-half correlations for a set of items. Again, a value of +.80 or greater is generally taken to indicate good internal consistency.

Validity

Reliability is important but, on its own, it's not enough. After all, I could create a perfectly reliable measurement on a personality test by re-coding every answer using the same number, regardless of how the person actually answers. We want our measurements to also be **valid** (see Figure 3.3.1) — that is, we want to make sure that we are actually measuring the construct that we think we are measuring. There are many different types of validity that are commonly discussed; we will focus on three of them.

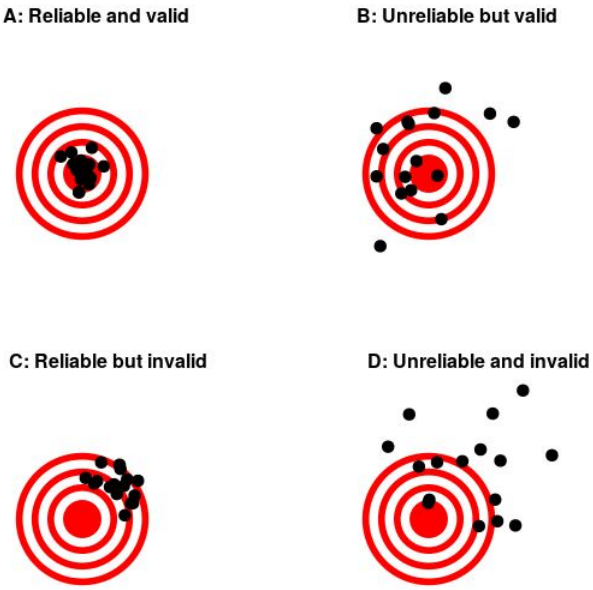


Figure 3.3.1. A figure demonstrating the distinction between reliability and validity, using shots at a bullseye. Reliability refers to the consistency of location of shots, and validity refers to the accuracy of the shots with respect to the centre of the bullseye. Poldrack (2019), Statistical thinking for the 21st Century. Licensed under CC BY-NC

Face Validity

Does the measurement make sense at face value? If I were to tell you that I was going to measure a person's blood pressure by looking at the colour of their tongue, you would probably think that, on the surface, this was not a valid measure. However, using a blood pressure cuff would have face validity. This is usually a first reality check before we dive into more complicated aspects of validity.

Content Validity

Content validity is the extent to which a measure "covers" the construct of interest. For example, if a researcher conceptually defines test anxiety as involving both sympathetic nervous system activation (leading to nervous feelings) and negative thoughts, then their measure of test anxiety should include items for both nervous feelings and negative thoughts.

Consider also that attitudes are usually defined as involving thoughts, feelings, and actions toward something. By this conceptual definition, a person has a positive attitude toward exercise to the extent that they think positive thoughts about exercising, feels good about exercising, and actually exercises. So to have good content validity, a measure of people's attitudes toward exercise would have to reflect all three of these aspects. Like face validity, content validity is not usually assessed quantitatively. Instead, it is assessed by carefully checking the measurement method against the conceptual definition of the construct.

Construct Validity

Is the measurement related to other measurements in an appropriate way? This is often subdivided into two aspects: convergent and divergent validity. **Convergent validity** means that the measurement should be closely related to other measures that are thought to reflect the same construct. Let's say that I am interested in measuring how extroverted a person is using a questionnaire or an interview. Convergent validity would be demonstrated if both of these different measurements are closely related to one another. However, measurements thought to reflect different constructs should be unrelated, known as **divergent validity**. If my theory of personality says that extraversion and conscientiousness are two distinct constructs, then I should also see that my measurements of extraversion are unrelated to measurements of conscientiousness.

Predictive Validity

If our measurements are truly valid, then they should also be **predictive** of other outcomes. For example, let's say that we think that the psychological trait of sensation seeking (the desire for new experiences) is related to risk-taking in the real world. To assess the

predictive validity of the sensation-seeking measurement, we would need to determine how effectively the scores on this test can predict scores on another survey that measures actual risk-taking behaviour (e.g., asking individuals if they would go sky-diving).

Assessing the Validity of a Study

When we read about psychology experiments with a critical view, one question to ask is, "is this study valid (accurate)?" Another one is to ask "can you trust the results of your study?" While the above types of validity are more applicable to measurements, we should also be assessing the validity of a study.

A study is said to be high in **internal validity** if the way it was conducted supports the conclusion that the predictor caused any observed differences in the outcome variable. Thus, experiments are high in internal validity because the way they are conducted — with the manipulation of the predictor and the control of extraneous variables (such as through the use of random assignment to minimise confounds) — provides strong support for causal conclusions. In contrast, non-experimental research designs (e.g., correlational designs), in which variables are measured but are not manipulated by an experimenter, are low in internal validity.

At the same time, the way that experiments are conducted sometimes leads to a different kind of criticism. Specifically, the need to manipulate the predictors and control extraneous variables means that experiments are often conducted under conditions that seem artificial (for instance, Bauman et al., 2014, had undergraduate students come to a laboratory on campus and complete a math test while wearing a swimsuit).[1] Furthermore, in many psychology experiments, the participants are all undergraduate students and come to a classroom or laboratory to fill out a series of paper-and-pencil questionnaires or to perform a carefully designed computerised task.

The issue we are confronting is that of external validity. **External validity** relates to the **generalisability** 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? An empirical study is high in external validity if the way it was conducted supports generalising the results to people and situations beyond those actually studied. 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.

In saying that, however, a study that uses only psychology students *does not necessarily* have a problem with external validity (Navarro & Foxtrot, 2022).[2] Psychology undergraduates differ from the general population in many ways, therefore using only psychology students in a study may compromise its external validity. However, if the differences between the groups are not relevant to the phenomenon under investigation, then external validity should not be a concern. Navarro and Foxtroxt (2022) provided examples to make this distinction more concrete:

> - *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*

Chapter attribution

This whole section contains taken and adapted material from several sources:

- *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.
- *Research methods in Psychology* by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler and Dana C. Leighton, used under a CC BY-NC-SA 4.0 licence.

1. Fredrickson, B. L., Roberts, T.-A., Noll, S. M., Quinn, D. M., & Twenge, J. M. (1998). The swimsuit becomes you: Sex differences in self-objectification, restrained eating, and math performance. *Journal of Personality and Social Psychology, 75*, 269–284 ↵
2. Navarro, D. J., & Foxcroft, D. R. (2022). *Learning statistics with jamovi: A tutorial for psychology students and other beginners* (Version 0.75). https://doi.org/10.24384/hgc3-7p15 ↵

3.4: Some Complexities

The previous classification scheme that we discussed is very useful in most variables that we would come across in psychological research. However, there are some variables that may not fit neatly into this classification scheme.

Let's take a classic example of a psychological measurement tool, the Likert scale. The Likert scale is the bread and butter of survey design. You've likely filled out hundreds (if not thousands) of these and may have even used one yourself. Consider a survey question like this:

> *Which of the following best describes your opinion of the statement 'pizzas are 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 handy but limited tools. The question is, what measurement are they? They're obviously discrete, since you can't give a response of 2.5. They're not nominal scale, since the items are ordered, and they're not ratio scale, since there's no natural zero.

But are they ordinal or interval scale? One argument is that we can't prove that the difference between "strongly agree" and "agree" is the same as the difference between "agree" and "neither agree nor disagree". In fact, it's pretty obvious in everyday life that they're not the same. This suggests that Likert scales should be treated as ordinal variables. On the other hand, most participants tend to take the "on a scale from 1 to 5" aspect seriously, acting as if the differences between the five response options are similar to one another. As a result, many researchers treat Likert scale data as interval scale. It's not interval scale but, in practice, it's close enough that it's often thought of as quasi-interval scale.

If you're interested in these kinds of debates, you can read the following commentary on this very important (ahem, nerdy) matter. [1]

Chapter attribution

This chapter contains taken and adapted material from *Learning statistics with jamovi* by Danielle J. Navarro and David R. Foxcroft, used under a CC BY-SA 4.0 licence.

1. Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education, 42*(12), 1150–1152. doi.org/10.1111/j.1365-2923.2008.03172.x ↵

3.5: The Role of Variables - Predictors and Outcomes

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 as Y and denote the variables "doing the explaining" as X_1, X_2 and so on.

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 X depends on Y, and if we have designed our study "properly" then Y isn't dependent on anything else.

I personally find those names unintuitive. 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.

A lot of statistical books still use these terms however, so it's still good to know them. 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). Navarro and Foxtrot (2022)[1] provided a summary of the differences which can be found in Table 3.5.1.

Table 3.5.1. *Variable distinctions*

Role of the variable	Classical name	Modern name
"to be explained"	dependent variable (DV)	outcome
"to do the explaining"	independent variable (IV)	predictor

Chapter attribution

This chapter contains taken and adapted material from *Learning statistics with jamovi* by Danielle J. Navarro and David R. Foxcroft, used under a CC BY-SA 4.0 licence.

1. Navarro, D. J., & Foxcroft, D. R. (2022). L*earning statistics with jamovi: A tutorial for psychology students and other beginners* (Version 0.75). https://doi.org/10.24384/hgc3-7p15 ↵

3.6: Research Design I- Experimental Designs

Psychologists agree that if their ideas and theories about human behaviour are to be taken seriously, they must be backed up by data. However, the research of different psychologists is designed with different goals in mind, and the different goals require different approaches. These varying approaches are known as research designs. In this section, we will talk about the different research designs that we use in psychology.

Experimental Research: Understanding the Causes of Behaviour

In psychology, the "gold standard" is an experimental design. Utilising an experimental design can assist in determining the impact of the predictor on the outcome by isolating the predictor as the probable cause. By comparing the outcomes of the experimental group and the control group, researchers can evaluate whether there are any differences. Since random assignment ensures that both groups are identical, with the only variation being the treatment or interventions, researchers can conclude that the difference in outcomes is likely due to the treatment. Given these factors, an experimental design is best suited to establish causation in research.

There are three critical components to experimental design: random assignment, manipulation of treatment, and the presence of a control group. Generally, there are two groups: an experimental group and a control group. The experimental group receives treatment, while the control group does not. The purpose of the control group is to represent what the experimental group would look like if it were not given the treatment.

We can diagram a research hypothesis in experimental research using an arrow pointing in one direction. This demonstrates the expected direction of causality (Figure 3.6.1):

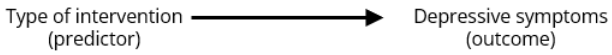


Figure 3.6.1. Expected direction of causality between the predictor and the outcome

Experimental designs can be conducted both in a lab or a field setting.

An Example of a Lab Experimental Design

Suppose a researcher aims to examine the effectiveness of cognitive-behavioural therapy (CBT) and pharmacological intervention in treating depression. To achieve this, the researcher employs an experimental design with random assignment, manipulation of treatment, and a control group. Thirty participants are recruited, and they are randomly assigned to one of three groups: a CBT group, a pharmacological intervention group, or a control group. The treatment outcome, such as reduced symptoms of depression, is the dependent variable (DV), while the type of intervention is the independent variable (IV), with three levels: CBT group, pharmacological intervention group, and control group.

To ensure the groups are comparable, the researcher may match participants based on demographic characteristics such as age, gender, and severity of depression. In the CBT group, participants may receive a standardised protocol of CBT sessions, while those in the pharmacological intervention group receive a standardised medication regimen. Participants in the control group received no treatment, simulating the natural course of depression without any intervention. By comparing the changes in the DV between the groups, the researcher can determine whether the intervention caused changes in the dependent variable. Overall, this experimental design enables the researcher to make causal inferences about the effectiveness of CBT and pharmacological intervention in treating depression.

As you can see from the above example, the researcher attempts to control all aspects of the study – especially what participants experience during the study. 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.

While lab-based experiments can help in establishing cause and effect between variables, they also have limitations that researchers should consider when selecting a research design. Some of the limitations of lab-based experiments include:

1. Artificiality: Lab experiments often take place in a highly controlled environment that may not reflect real-world situations. Participants may behave differently in a lab than they would in their natural environment. Therefore, the results may not be generalizable to real-world situations.

2. Demand characteristics: Participants in lab experiments may behave in a way that they think the researcher expects them to behave, rather than behaving naturally. This can happen due to the artificial setting or because participants may try to please the researcher. This can lead to biased results.
3. Limited external validity: Lab experiments may not be representative of real-world populations or situations. The participants in a lab experiment may not represent the larger population, and the experimental task may not accurately represent real-world situations. This can limit the external validity of the results.

An Example of a Field Experimental Design

Field experiments take place in a real-world setting, such as a workplace, school or community. Field experiments are conducted in a natural setting, and the researcher does not have as much control over the experimental conditions as in a lab experiment. For example, a researcher may conduct a field experiment to examine the effect of a job training program on job performance by randomly assigning participants in a company to either a training or non-training condition. The researcher would observe the participants in their natural work environment rather than in a controlled lab setting.

One advantage of field experiments is that they provide greater ecological validity, meaning that the findings are more generalisable to real-world situations. Some of the limitations of field-based experiments include:

1. Limited control: Field experiments are often conducted in real-world settings, which means the researcher has less control over the experimental conditions than in lab-based experiments. This can make it more difficult to isolate the effects of the independent variable and to control for extraneous variables that may affect the outcome.
2. Confounding variables: In field experiments, there may be more confounding variables that can influence the outcome of the study. These variables are not controlled by the researcher, and therefore, their effects cannot be isolated from the effects of the independent variable.
3. Difficulties in randomisation: Randomisation, which is the process of assigning participants to different groups, can be more difficult in field experiments than in lab experiments. For example, it may be difficult to ensure that participants in the control and experimental groups are similar in terms of their demographics, attitudes and behaviours.

3.7: Research Design II- Non-Experimental Designs

Researchers who are simply interested in describing characteristics of people, describing relationships between variables, and using those relationships to make predictions can use a non-experimental research design. Using the non-experimental approach, the researcher simply measures variables as they naturally occur, but they do not manipulate them.

For instance, if a researcher is interested in measuring the number of traffic fatalities in Queensland last year that involved mobile phones, a researcher may not be able to manipulate 'mobile phone use while driving', but can simply collect data about a phenomenon that has already occurred. Another example would be standing at a busy intersection and recording the driver's gender and whether or not they were using a mobile phone while they pass through the intersection, and then analysing the data to see whether men or women are more likely to use a mobile phone when driving. Again, this time, the researcher is just observing the variables (use of mobile phone and gender) and is not manipulating anything.

It is important to point out that 'non-experimental' does not mean nonscientific. Non-experimental research is still scientific in nature. It can be used to fulfil two of the three goals of science (to describe and to predict). However, unlike experimental research, we cannot make causal conclusions using this method as the researcher does not have full control of all aspects of the design. With the example we used above, it is possible that there is another variable that is not part of the research hypothesis but that causes both the predictor and the outcome variable and thus produces the observed correlation between them.

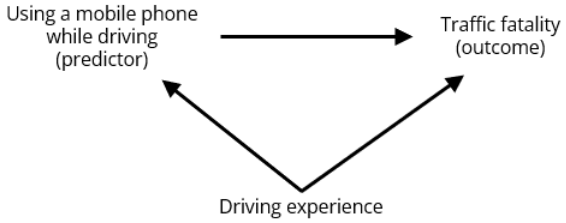


Figure 3.7.1 How a potential confounding variable can affect the relationship between two variables

There are different examples of non-experimental designs and we will cover some of these types below.

Case Studies

Case studies involve the idiographic, observational (and/or interview) study of an individual or individuals by the researcher, such as within a clinical context where the focus might be on the participant's lived experience. A famous example of a case study in psychology is the case of Phineas Gage. In 1848, Gage, a railroad worker, survived a severe brain injury when a metal rod was accidentally driven through his skull, damaging his frontal lobes. Remarkably, Gage survived the injury and was able to walk and talk normally, but his personality and behaviour underwent dramatic changes.

Case studies allow for a detailed and in-depth examination of the individual, group, or situation under investigation. Researchers can gather a wealth of information about the person or phenomenon being studied. Unique or rare cases (like Phineas Gage) are particularly useful for studying unique or rare cases that may not be easily observed or studied in other ways. Case studies can be used to generate hypotheses or ideas about potential cause-and-effect relationships that can be tested in future research.

However, as with any research designs, case studies are limited due to the following reasons:

1. Limited generalisability: Due to the focus on a single individual, group, or situation, case studies have limited generalisability to larger populations. It may be difficult to generalise findings from a case study to the wider population.
2. Subjectivity: Case studies are often subjective, as researchers may have personal biases or interpretations that influence their analysis and conclusions.
3. Lack of control: Case studies lack experimental control, which makes it difficult to establish cause-and-effect relationships. It is also difficult to replicate the same conditions across different cases, making it difficult to determine if the findings are consistent.

Quasi-Experimental Designs

A quasi-experimental design is essentially a hybrid of experimental and non-experimental designs. It aims to establish a cause-and-effect relationship between an independent and dependent variable. However, unlike a true experimental design, a quasi-experiment does not rely on random assignment. Instead, subjects are assigned to groups based on non-random criteria.

Quasi-experimental designs are common in psychology research and feature non-random assignment to condition and/or non-manipulation of independent variables, often through necessity. As an example, imagine that some school authorities in Queensland want to implement a new math curriculum and they are interested in determining whether the curriculum is effective in improving student performance. The school authorities decide to implement the new curriculum in one school but not in another. They then compare the test scores of the students in both schools before and after the implementation of the new curriculum.

Since the assignment of the schools to the different conditions (new versus old curriculum) was not random, this is a quasi-experimental design. The study attempts to establish a causal relationship between the new curriculum and student performance by comparing the pre-post scores of the two groups. However, there may be other factors that could account for the differences observed between the two groups, such as differences in student populations or teacher quality. Therefore, the study has limited internal validity.

CHAPTER OVERVIEW

Chapter 4: The Replication Crisis

Learning Objectives

After reading this chapter, you should be able to:

- describe what is meant by the “replicability crisis” in psychology
- describe some questionable research practices
- identify some ways in which scientific rigour may be increased
- understand the importance of openness in psychological science.

The replication of findings is a key characteristic of science. For a study’s results to be considered part of scientific knowledge, they must be replicable. This process helps to prevent false positive results (i.e., when you think something is true when it is actually false–a false alarm) and increases confidence in the validity of the findings.

As mentioned in the introduction of this book, the field of psychology is currently facing a replication crisis. In the era of instant news, the inability to replicate research raises serious concerns about the reliability of the scientific process. The public has a right to know if they can trust research evidence, and as psychologists, it’s in our best interest to ensure our methods and findings are trustworthy.

In this chapter, we will look at what we meant by the replicability crisis in psychology and what are the conditions that allowed for this to occur. Then, we will talk about potential solutions to this issue.

4.1: How we Think Science Should Work
4.2: Reasons for Non-Replication
4.3: What can we do About it?

4.1: How we Think Science Should Work

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

The scientific method movement believes that truth can be objectively discovered, supposedly separate from the subjective biases of the person seeking it. The scientific method — a systematic approach that includes forming a hypothesis, conducting experiments, and analysing and refining results — was introduced as a way to achieve this objective truth.

How we Think Science Should Work

Let's say that we are interested in a research project on how children choose what to eat. This is a question that was asked in a study by the well-known eating researcher Brian Wansink and his colleagues in 2012.[1] The standard (and, as we will see, somewhat naive) view goes something like this:

- You start with a hypothesis
  - Branding with popular characters should cause children to choose "healthy" food more often
- You collect some data
  - Offer children the choice between a cookie and an apple with either an Elmo-branded sticker or a control sticker, and record what they choose
- You do statistics to test the **null hypothesis.** The null hypothesis test states that you don't find an effect – we will learn more about this later.
  - "The preplanned comparison shows Elmo-branded apples were associated with an increase in a child's selection of an apple over a cookie, from 20.7% to 33.8%".
- You make a conclusion based on the data
  - "This study suggests that the use of branding or appealing branded characters may benefit healthier foods more than they benefit indulgent, more highly processed foods. Just as attractive names have been shown to increase the selection of healthier foods in school lunchrooms, brands and cartoon characters could do the same with young children."

However, it has since been recognised that the illusion of objectivity inherent in the scientific method can lead to a false sense of confidence in one's ability to uncover the truth and that the results of scientific research are not certain. This was exemplified by the replication crisis[2], which revealed the extent to which subjective biases and self-interest can influence scientific findings.

How Science (Sometimes) Actually Works

Let's look at what happened with Brian Wansink's studies below:

Examples

Brian Wansink is well known for his books on "Mindless Eating", and his fee for corporate speaking engagements is in the tens of thousands of dollars. In 2017, a set of researchers began to scrutinise some of his published research, starting with a set of papers about how much pizza people ate at a buffet. The researchers asked Wansink to share the data from the studies but he refused, so they dug into his published papers and found a large number of inconsistencies and statistical problems in the papers. The publicity around this analysis led a number of others to dig into Wansink's past, including obtaining emails between Wansink and his collaborators. As reported by Stephanie Lee at Buzzfeed, these emails showed just how far Wansink's actual research practices were from the naive model:

> *…back in September 2008, when Payne was looking over the data soon after it had been collected, he found no strong apples-and-Elmo link — at least not yet. … "I have attached some initial results of the kid study to this message for your report," Payne wrote to his collaborators. "Do not despair. It looks like stickers on fruit may work (with a bit more wizardry)." … Wansink also acknowledged the paper was weak as he*

> *was preparing to submit it to journals. The p-value was 0.06, just shy of the gold standard cutoff of 0.05. It was a "sticking point," as he put it in a Jan. 7, 2012, email. … "It seems to me it should be lower," he wrote, attaching a draft. "Do you want to take a look at it and see what you think. If you can get the data, and it needs some tweeking, it would be good to get that one value below .05." … Later in 2012, the study appeared in the prestigious JAMA Pediatrics, the 0.06 p-value intact. But in September 2017, it was retracted and replaced with a version that listed a p-value of 0.02. And a month later, it was retracted yet again for an entirely different reason: Wansink admitted that the experiment had not been done on 8- to 11-year-olds, as he'd originally claimed, but on preschoolers (Lee, 2017).*[3]

This kind of behaviour finally caught up with Wansink; fifteen of his research studies have been retracted and in 2018 he resigned from his faculty position at Cornell University.

Chapter attribution

This chapter contains taken and adapted material from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

1. Wansink, B., Just, D. R., & Payne, C. R. (2012). Can branding improve school lunches? *Archives of pediatrics & adolescent medicine*, *166*(10), 967-968. https://doi.org/10.1001/archpediatrics.2012.999 ↵
2. https://en.wikipedia.org/wiki/Replication_crisis ↵
3. Lee, S. (2017, September 25). How A Star Cornell Food Scientist Wowed Prestigious Journals With His "Artful Pizzazz". BuzzFeed News. https://www.buzzfeednews.com/article/stephaniemlee/brian-wansink-cornell-p-hacking ↵

4.2: Reasons for Non-Replication

Brian Wansink is just one of many examples of the replication crisis that is currently facing psychology. There are many other reasons for the replication crisis, which we will discuss below.

Questionable Data Practices

Some suggested that the low replicability of many studies is evidence of the widespread use of questionable research practices by psychological researchers. These may include:

1. The selective deletion of outliers to influence (usually by artificially inflating) statistical relationships among the measured variables.
2. The selective reporting of results, that is, cherry-picking only those findings that support one's hypotheses.
3. Mining the data without an a priori hypothesis, only to claim that a statistically significant result had been originally predicted, a practice referred to as "HARKing" or hypothesising after the results are known (Kerr, 1998).[1]
4. A practice colloquially known as "*p*-hacking", in which a researcher might perform inferential statistical calculations to see if a result was significant before deciding whether to recruit additional participants and collect more data (Head et al., 2015).[2] As you will learn later on, the probability of finding a statistically significant result is influenced by the number of participants in the study.
5. Outright fabrication of data (as the case for Brian Wansink's studies) — although this would be a case of fraud rather than a "research practice."

Small Sample Sizes

Another reason for non-replication is that, in studies with small sample sizes, statistically significant results may often be the result of chance. For example, if you ask five people if they believe that aliens from other planets visit Earth and regularly abduct humans, you may get three people who agree with this notion — simply by chance. Their answers may, in fact, not be at all representative of the larger population. On the other hand, if you survey one thousand people, there is a higher probability that their belief in alien abductions reflects the actual attitudes of society. Now consider this scenario in the context of replication: if you try to replicate the first study — the one in which you interviewed only five people — there is only a small chance that you will randomly draw five new people with exactly the same (or similar) attitudes. It's far more likely that you will be able to replicate the findings using another large sample because it is simply more likely that the findings are accurate.

Results may be True for Some People, in Some Circumstances

Another reason for non-replication is that, while the findings in an original study may be true, they may only be true for some people in some circumstances and not necessarily universal or enduring. Imagine that a survey in the 1950s found a strong majority of respondents trust government officials. Now imagine the same survey administered today, with vastly different results. This example of non-replication does not invalidate the original results. Rather, it suggests that attitudes have shifted over time.

Systemic Issues

Others have interpreted this situation as evidence of systemic problems with conventional scholarship in psychology, including a publication bias that favours the discovery and publication of counter-intuitive but statistically significant findings instead of the duller (but incredibly vital) process of replicating previous findings to test their robustness (Aschwanden, 2015; Pashler & Harris, 2012).[3][4]

Chapter attribution

This chapter contains taken and adapted material from *Research methods in psychology* by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler and Dana C. Leighton, used under a CC BY-NC-SA 4.0 licence.

1. Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196-217. doi.org/10.1207/s15327957pspr0203_4 ↵
2. Head M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3): e1002106. http://doi.org/10.1371/journal.pbio.1002106 ↵

3. Aschwanden, C. (2015, August 19). *Science isn't broken: It's just a hell of a lot harder than we give it credit for*. FiveThirtyEight. http://fivethirtyeight.com/features/science-isnt-broken/ ↵
4. Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments explained. *Perspectives on Psychological Science*, *7*(6), 531-536. doi.org10.1177/1745691612463401 ↵

4.3: What can we do About it?

It is important to shed light on these questionable research practices to ensure that current and future researchers (such as yourself) understand the problems these research practices create for our discipline. However, in addition to highlighting *what not to do*, we should also talk about potential solutions to this so-called "crisis". Easy changes we can make now include:

1. Designing and conducting studies that have sufficient statistical power, in order to increase the reliability of findings.
2. Publishing both null and significant findings (thereby counteracting the publication bias and reducing the file drawer problem).
3. Describing one's research designs in sufficient detail to enable other researchers to replicate a study using an identical or at least very similar procedure.
4. Conducting high-quality replications and publishing these results (Brandt et al., 2014).[1]

Furthermore, there has been a movement to develop tools to help protect the reproducibility of scientific research. We will discuss each of them below.

Pre-Registration

One of the ideas that has gained the greatest traction is *pre-registration*, in which one submits a detailed description of a study (including all data analyses) to a trusted repository (such as the Open Science Framework or AsPredicted.org). By specifying one's plans in detail prior to analysing the data, pre-registration provides greater faith that the analyses do not suffer from p-hacking or other questionable research practices. Pre-registration is a vital part of the Open Science Framework (see Figure 4.3.1 below).

The use of pre-registration in clinical trials has shown significant results. For example, the National Heart, Lung, and Blood Institute began requiring pre-registration of all clinical trials in 2000 through ClinicalTrials.gov. A study by Kaplan and Irvin (2015)[2] found that the number of positive results in clinical trials decreased after pre-registration was implemented, suggesting that pre-registration reduced the ability of researchers to manipulate their methods and hypotheses for a positive outcome.

Replication

As mentioned previously, the ability to replicate results is a critical aspect of science. To increase the likelihood of replicability, researchers should first attempt to replicate their own findings using a new, adequately powered sample. However, failure to replicate does not necessarily mean the original finding was incorrect. Multiple replications are needed to determine the validity of a finding. In the past, many fields, including psychology, have neglected this principle, resulting in "textbook" findings that may be false.

It's important to note that p-values do not provide an estimate of the replicability of a finding. The p-value only reflects the likelihood of the data under a specific null hypothesis, not the probability that the finding is true. In order to know the likelihood of replication, we need to know the probability that the finding is true, which we generally don't know.

Reproducible Practices

The paper by Simmons, Nelson, and Simonsohn (2011)[3] laid out a set of suggested practices for making research more reproducible, all of which should become standard for researchers:

> - *Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.*
> - *Authors must collect at least 20 observations per parameter or else provide a compelling cost-of-data-collection justification.*
> - *Authors must list all variables collected in a study.*
> - *Authors must report all experimental conditions, including failed manipulations.*
> - *If observations are eliminated, authors must also report what the statistical results are if those observations are included.*
> - *If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.*

Doing Reproducible Data Analysis

The focus on replication so far has been on repeating experiments to verify findings by other researchers. However, computational reproducibility – the ability to reproduce someone's data analysis – is also crucial. This requires researchers to share both their data and analysis code, allowing others to validate the results and test different methods. There is a growing trend in psychology towards open sharing and including the "open science badges" provided by the Centre for Open Science to encourage pre-registration and to share data, code and research materials.

Figure 4.3.1 "Open science badges" by Open Science Collaboration is licensed under CC BY 3.0

I also recommend using scripted analysis tools like R and free and open-source software (like jamovi!) rather than commercial ones, to promote reproducibility. Code can be shared on version control sites like Github, while datasets can be shared on portals like OSF.

Doing Better Science

It is every scientist's responsibility to improve their research practices in order to increase the reproducibility of their research. It is essential to remember that the goal of research is not to find a significant result, rather, it is to ask and answer questions about nature in the most truthful way possible. Most of our hypotheses will be wrong, and we should be comfortable with that, so that when we find one that's right, we will be even more confident in its truth.

Chapter attribution

This chapter contains taken and adapted material from *Research methods in psychology* by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler and Dana C. Leighton, used under a CC BY-NC-SA 4.0 licence.

1. Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217-224. http://doi.org/10.1016/j.jesp.2013.10.005 ↵
2. Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PloS one*, *10*(8), e0132382. https://doi.org/10.1371/journal.pone.0132382 ↵
3. Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science 22*(11): 1359–66. doi.org/10.1177/0956797611417632 ↵

CHAPTER OVERVIEW

Chapter 5: Aggregation

Learning Objectives

After reading this chapter, you should be able to:

- generate a tabular and graphical representation of a frequency distribution
- understand the importance of data visualisation
- describe different measures of central tendency and dispersion, how they are computed, and which are appropriate under what circumstances
- compute a z-score and describe why they are useful.

As mentioned previously, one of the main ideas behind statistics is the idea of **aggregation**. As a reminder, we discussed the idea that we can better understand the world by throwing away information. In other words, when we aggregate data, that's exactly what we are doing when we summarise a dataset.

We will also delve into another main idea behind statistics: **variation**. As mentioned in Chapter 1, if variation didn't exist, then we wouldn't need statistics. In this chapter, we will discuss why we summarise data and how we can summarise data and explain variation in data in meaningful ways.

5.1: Why Summarise Data?
5.2: Summarising Data Using Tables
5.3: Summarising Data Using Graphs
5.4: The Middle of the Data
5.5: Variability - How Spread Out are the Values?
5.6: Z Scores

5.1: Why Summarise Data?

When we summarise data, we are essentially throwing away information, and one may object to this. As an example, let's go back to the PURE study that we discussed in Chapter 1. You may think, what happens to participant information beyond what was summarised in the dataset? What about the specific details of how the data were collected such as the time of day or whether it was on a weekend or a weekday? What about the mood of the participant? All of these details are lost when we summarise the data.

We summarise data because it allows us to *describe* and *compare*. Two actions that we do all the time in everyday life.

Curiously, students often complain that statistics is confusing and irrelevant. But, the very same students are the ones comparing their GPAs with their classmates. They may even be calculating the grade they need for the next assessment in order to pass. Most people don't realise this, but we are cognizant of **descriptive statistics** because we use numbers to summarise information on a daily basis.

For example, GPA is often used by universities to assess high school students' academic potential. In Australia, high school students are given a ranking called the Australian Tertiary Admission Rank (ATAR) as the primary criterion to enter undergraduate courses. A student with an ATAR of 95 is clearly a stronger student than someone in the same class with an ATAR of 80.[1] Using Charles Wheelan's words from *Naked Statistics*,[2] GPA (or in our case, ATAR) makes a nice descriptive statistic: it's easy to calculate, it's easy to understand, and it's easy to compare across students. However, it is important to note that it is not *perfect* as it often does not reflect the difficulty of the subjects that different students may have taken.

We also summarise data because it enables us to *generalise*. That is, to make general statements that extend beyond specific observations. Psychologists have long recognised the importance of generalisation, including the process of categorisation. For example, we can easily recognise different types of birds, despite their differing surface features. We know that an ostrich, a robin and a chicken belong to the "bird category", but we understand that these birds individually differ from one another. Importantly, generalisation allows us to make predictions. In the case of birds, we can predict that they can fly and eat worms and that they probably can't drive a car or speak English. These predictions won't always be right, but they are often useful in the real world.

Chapter attribution

This chapter contains taken and adapted material from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

1. It's a percentile ranking between 0.00 and 99.95. For more information: https://en.wikipedia.org/wiki/Australian_Tertiary_Admission_Rank ↵
2. Wheelan, C. (2013). *Naked statistics: Stripping the dread from the data*. WW Norton & Company. ↵

5.2: Summarising Data Using Tables

A simple way to summarise data is to generate a table representing counts of various types of observations. This type of table has been used for thousands of years (see Figure 5.2.1).

Figure 5.2.1. A Sumerian tablet from the Louvre, showing a sales contract for a house and field, by Jastrow, is in the public domain

Let's look at some examples of the use of tables. We will use the crash_data_ardd.csv dataset. This dataset is collected by the Australian Road Deaths Database (ARDD) which provides basic information on fatalities resulting from road transport crash fatalities in Australia as reported monthly by the police to state and territorial road safety authorities. Data is available from 1989 to 2021. It's a large dataset! If you scroll through the data, you will see that there are 52,843 entries in this dataset representing 52,843 casualties between 1989 to 2021.

Within this dataset, there's a variable named *Dayweek*, which represents the day of the week the casualty occurred. We can make more sense of the data by creating a frequency table. Doing so will group the data by the different values of the variable, and then count how many values there are in each group:

Frequencies

Frequencies of Dayweek

Levels	Counts	% of Total	Cumulative %
Monday	6108	11.6 %	11.6 %
Tuesday	6145	11.6 %	23.2 %
Wednesday	6663	12.6 %	35.8 %
Thursday	7106	13.4 %	49.2 %
Friday	8665	16.4 %	65.6 %
Saturday	9696	18.3 %	84.0 %
Sunday	8460	16.0 %	100.0 %

Figure 5.2.2. Frequency table of the variable Dayweek from the crash_data_ardd.csv dataset

This table shows the frequencies of each of the different values. Between 1989 and 2021, there were 6,108 casualties recorded on a Monday, 6,145 casualties on a Tuesday and so on. By looking at the counts, we can tell that the highest number of casualties occurred on Saturday. However, it can be hard to tell from absolute numbers how big the difference is. For this reason, we would often present the data using percentages. In the table above, **% of Total** shows how many casualties occurred on each weekday by percentages. From here, we can see that 18.3% of the casualties occurred on a Saturday.

Note that jamovi also calculates the **cumulative percentage** (which adds the percentages of each value from the top of the table to the bottom). It's not a particularly useful statistic if you are more interested in the specific percentage per category. In our example above, the more interesting information you can get from the cumulative percentage is that 65.6% of the casualties occurred in during the weekday.

Cumulative percentage is more useful when the variable of analysis is ranked or ordinal, as it makes it easy to get a sense of what percentage of cases fall below (or above) each rank. I also think the cumulative percentage is more meaningful if it's presented as a graph.[1]

Jamovi Exercises – Frequency Table

Create a frequency table in jamovi from the crash_data_ardd.csv dataset. You can do this by opening the file into jamovi. Go to Analyses > Exploration. Check Frequency Tables. Put the variables you are interested in under the **Variables** window. See

below:

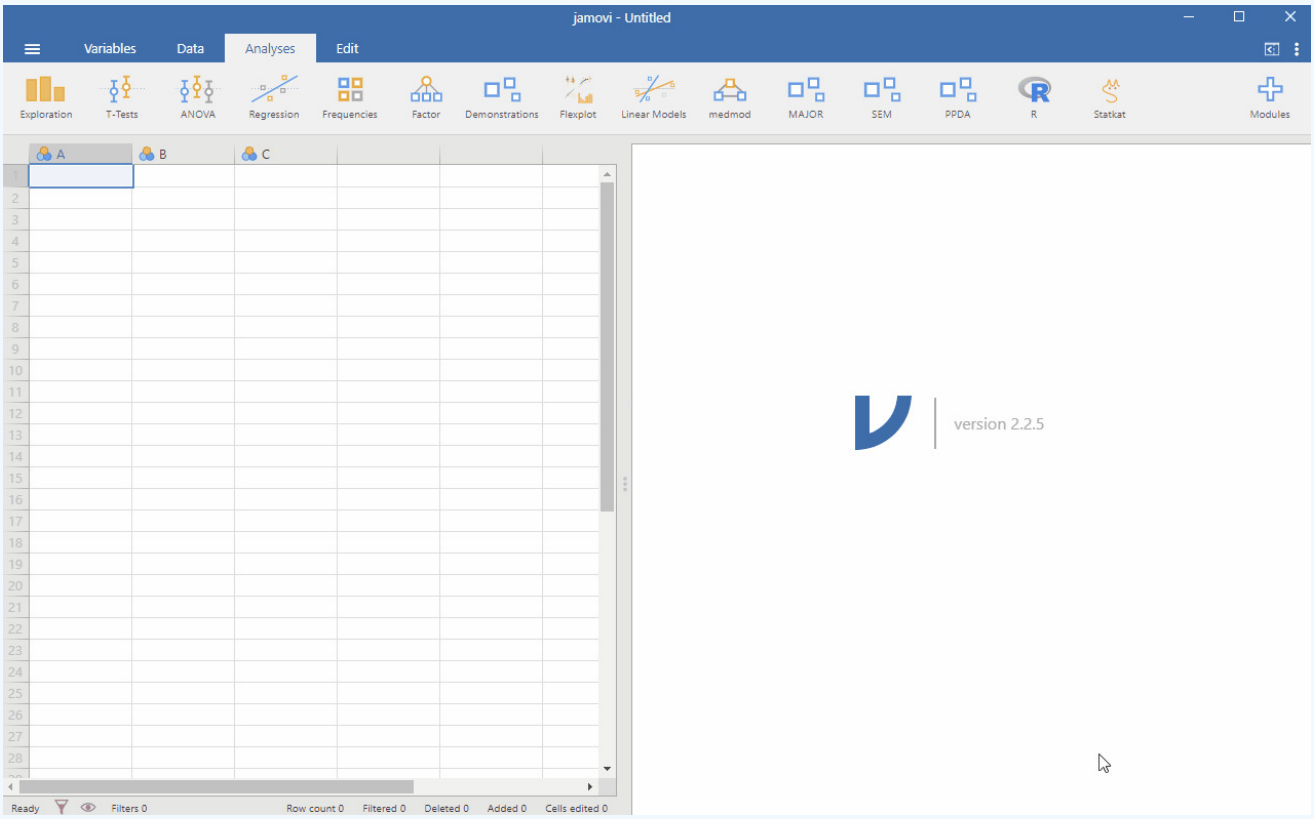


Figure 5.2.3 Creating a frequency table in jamovi

- Recreate the frequency table above for Dayweek variable.
- Create a new frequency table for Gender. From your frequency table, what percentage of men are involved in the casualties?

Chapter attribution

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

---

1. See this wonderful interactive graph showing the Cumulative confirmed COVID-19 cases by world region from *Our World in Data*: https://ourworldindata.org/grapher/cumulative-covid-cases-region ↵

5.3: Summarising Data Using Graphs

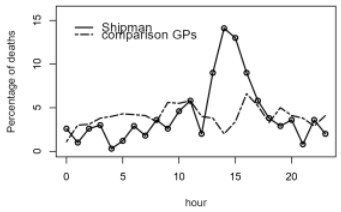


Figure 5.3.1 The time at which Harold Shipman's patients died, compared to the times at which patients of other local general practitioners died. The pattern does not require sophisticated statistical analysis.

Figure 5.3.1 shows a line graph that illustrates the time of day when Shipman's patients died (represented by a solid line) in contrast to the times of death among a sample of patients treated by other local GPs (represented by a broken line). As you can see in the figure, the majority of Shipman's patients tended to pass away during the early afternoon hours. According to Spiegelhalter, the data itself doesn't provide an explanation for this pattern, however, a deeper inquiry into his practices uncovered that he conducted his home visits post-lunch, a time when he was typically alone with his elderly patients.

Why do we Need to Visualise Data?

The above story is pretty grim. However, I hope that you can see the importance of data visualisation from reading the story. 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 and 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.

Plotting Frequencies

In addition to creating tables to look at frequency, we can also plot them in a graph. Now let's look at another type of variable: **Age**. Since age is a continuous variable – and in this dataset, we have ages from 0 to 101 – this time, it wouldn't make sense to put them in a frequency table. If we do, it will create a really long table with all available ages in the dataset (see Figure 5.3.2 for what happens when I graphed age in a frequency table).

Frequencies

Frequencies of Age

Levels	Counts	% of Total	Cumulative %
0	226	0.4 %	0.4 %
1	189	0.4 %	0.8 %
2	204	0.4 %	1.2 %
3	183	0.3 %	1.5 %
4	189	0.4 %	1.9 %
5	190	0.4 %	2.2 %
6	161	0.3 %	2.5 %
7	152	0.3 %	2.8 %
8	186	0.4 %	3.2 %
9	160	0.3 %	3.5 %
10	184	0.3 %	3.8 %
11	148	0.3 %	4.1 %

Figure 5.3.2. What happens when we put age in a frequency table

Instead, we will plot age using histograms, density plots and violin plots. First, let's plot the age variable for all of the individuals in the crash_data_ardd.csv dataset using histograms (see Figure 5.3.3-A). Histograms use bars to display the frequency of values in a dataset within specified ranges (or bins). It is not clear what range is used below, but we can assume that the ranges used were 0-5, 5-10, 10-15, and so on.[3] From the figure above, you may notice a large spike in deaths at around age 20-25.

We can ask jamovi to add a density plot to your histogram. A density plot depicts the data distribution with a smooth curve representing the proportion of values in each range. In essence, a histogram shows value counts in ranges with bars, while a density

plot presents a continuous distribution curve. The spike is clearer with the density plot (Figure 5.3.3-B), but another visualisation may be useful here.

Let's ask jamovi to instead create a violin plot (Figure 5.3.3-C). This spike is clearer in the violin plot. There is a higher density of crashes just below the age of 25. What do you think that spike is about?

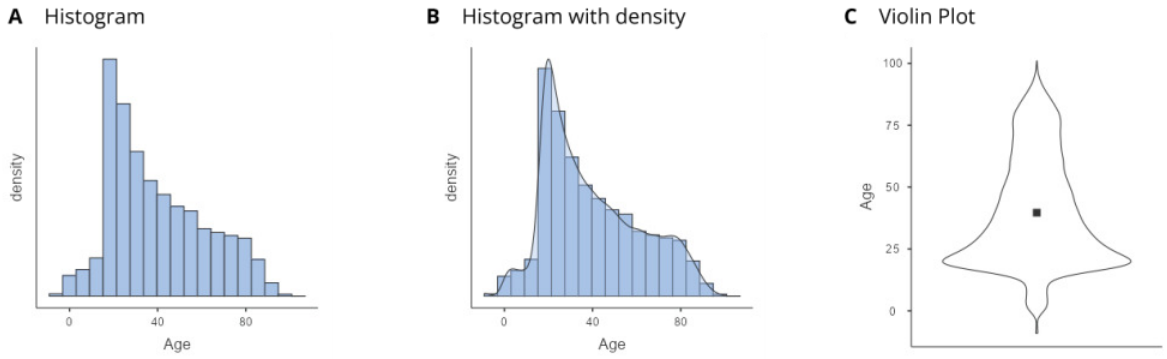


Figure 5.3.3 A histogram of the age variable in the crash_data_ardd.csv dataset without a density plot (A) and with the density plot (B) and a violin plot of the age variable (C).

According to the Bureau of Infrastructure, Transport and Regional Economics (BITRE),[4] around 20% of 1 in 5 individuals who are killed on the road were aged 17 to 25 years. Our violin plot clearly shows this.

We will later discuss more ways that we can visualise data, but violin plots (along with density plots) visualise the distribution of data over a continuous interval or time period. These plots are especially useful when we want to make a comparison of distributions between multiple groups. The peaks, valleys and tails of each group's density curve can be compared to see where groups are similar or different.

Figure 5.3.4 shows three different ways to plot these data and we will cover each one below.

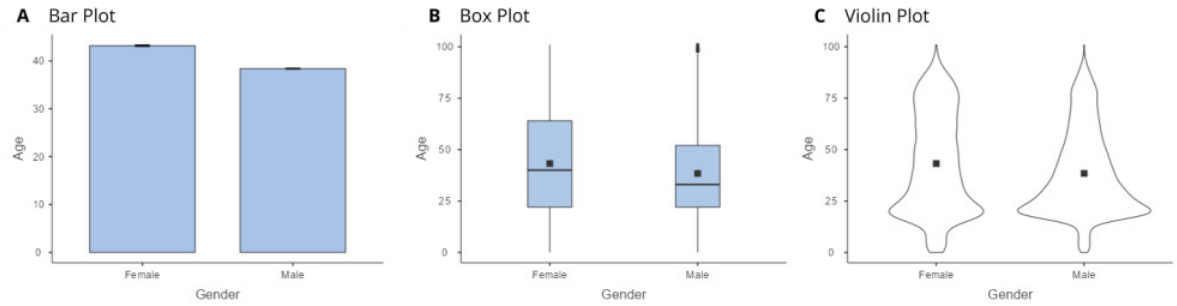


Figure 5.3.4 Three different ways of plotting the difference in ages between men and women in the ARDD dataset. Panel A plots the means of the two groups, which gives no way to assess the relative overlap of the two distributions. Panel B shows a box plot, which highlights the spread of the distribution along with any outliers (which are shown as individual points). Panel C shows a violin plot, which shows the distribution of the datasets for each group

Bar Graphs

The bar graph in panel A shows the difference in means, but doesn't show us how much spread there is in the data around these means. As we will see later, knowing this is essential to determine whether we think the difference between the groups is large enough to be important.

Box Plots and Violins

Another option is the **box plot** shown in panel B, which shows the median (central line), a measure of variability (the width of the box, which is based on a measure called the interquartile range), and any outliers (noted by the points at the ends of the lines). Since the box plot automatically separates out those observations that lie outside a certain range (depicting them with a dot in

jamovi) people often use them as an informal method for detecting **outliers**: observations that are "suspiciously" distant from the rest of the data.

In panel C, we see one example of a **violin plot**, which plots the distribution of data in each condition. 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 standard box plots. In jamovi you can achieve this sort of functionality by checking both the "Violin" and the "Box plot" checkboxes. You can also turn on "Data" to show the actual data points on the plot. This does tend to make the graph a bit too busy, in my opinion. Clarity is simplicity, so in practice, it might be better to just use a simple box plot.

In general, we prefer box plots and violins as plotting techniques as they provide a clearer view of the distribution of the data points.

Learning how to draw graphs in jamovi is reasonably simple as long as you're not too picky about what your graph looks like. Figure 5.3.5 below shows the different plots currently available in jamovi. In jamovi there are a lot of 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 base graphics functionality in jamovi is not yet capable of supporting advanced work or detail editing. Instead, you will need to install other modules in jamovi for other graphing functionalities.

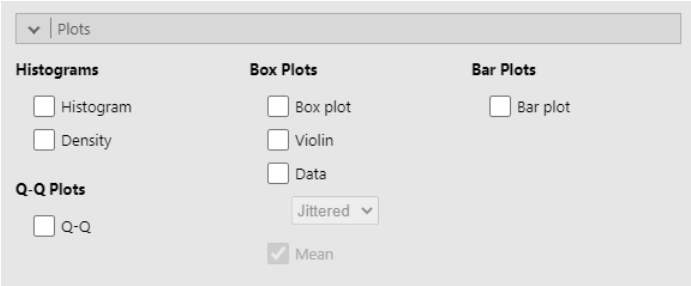


Figure 5.3.5 Available plots in jamovi

Given that jamovi is limited in its capacity for detail editing, I won't delve deep into the principles of good visualisation. However, if you'd like to know about how to make effective visualisations, I suggest the following readings:

- Poldrack, R. (2022). Principles of good visualization. In *Statistical thinking in the 21st century.* https://statsthinking21.github.io/statsthinking21-core-site/data-visualization.html#principles-of-good-visualization (this is an open textbook!)
- Knaflic, C. N. (2015). *Storytelling with data*. Wiley. onlinelibrary.wiley.com/doi/book/10.1002/9781119055259

Chapter attribution

This chapter contains taken and adapted material from *Learning statistics with jamovi* by Danielle J. Navarro and David R. Foxcroft, used under a CC BY-SA 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

---

1. Tufte, E. R. (1983). *The visual display of quantitative information*. Graphics Press. ↵
2. https://en.wikipedia.org/wiki/Harold_Shipman ↵
3. This is one of the limitations of jamovi. ↵
4. Bureau of Infrastructure and Transport Research Economics (2021). *Road trauma Australia 2021 statistical summary.* www.bitre.gov.au/sites/default/files/documents/road_trauma_2021.pdf ↵

5.4: The Middle of the Data

Creating tables or drawing pictures of the data as seen previously 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.

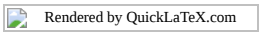
The Mean

The **mean** of a set of observations is just a normal, old-fashioned average. To calculate the mean, we add all of the values up and then divide by the total number of values. Let's look at the ages of those under Crash ID 19891001, 19891002, 19891003 and 19891004.

Table 5.4.1 Ages of those with Crash ID 19891001, 19891002, 19891003 and 19891004

Crash ID	Age
19891001	66
19891002	42
19891003	18
19891004	76

The mean of these observations is:



Calculating averages in jamovi

Averages (that is, means) are used so often in everyday life that this should be pretty familiar to you. We can find the average age of all people in the ARDD dataset by going to Analyses > Exploration. Drag *Age* into the Variables window. Under Descriptives, choose Variables across rows (this is optional – I prefer this setting because I think it looks cleaner this way). See below for the steps:

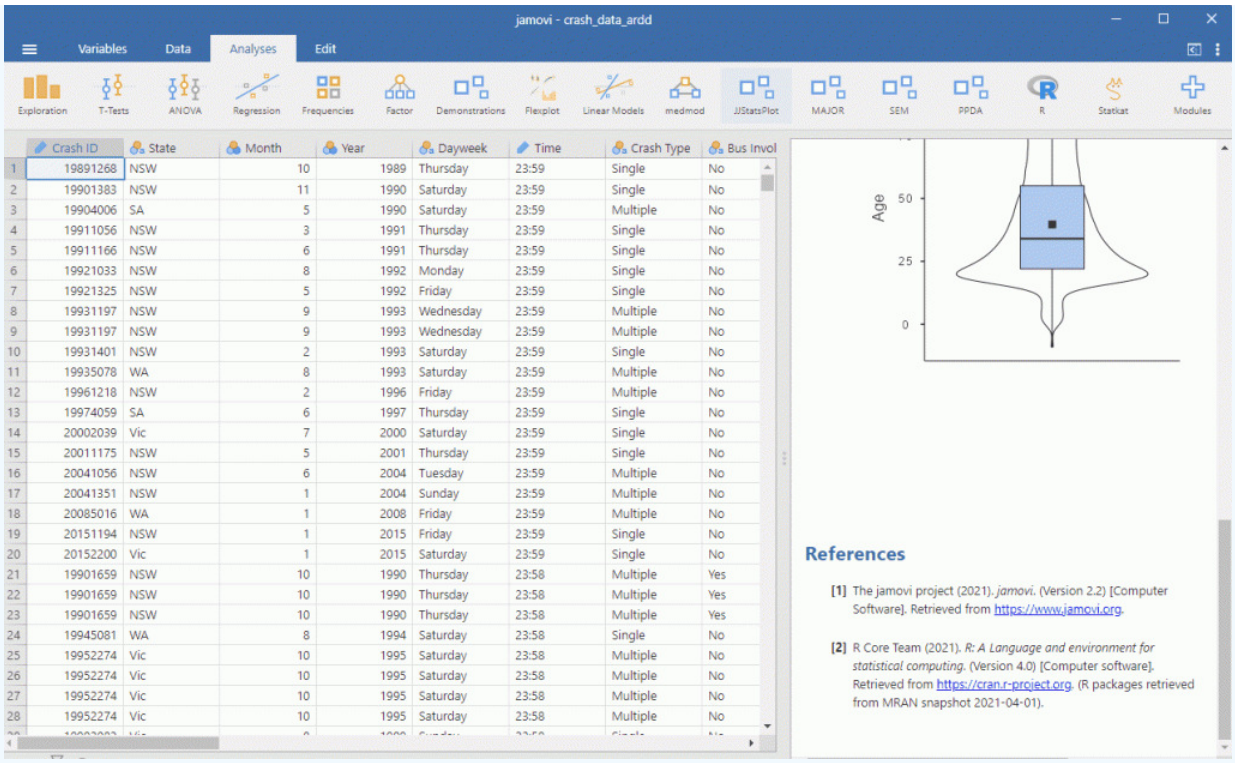


Figure 5.4.1. How to run Descriptives statistics in jamovi

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 is the "middle" value. As before, let's imagine we are only interested in the four sets of Crash ID's we have identified earlier. To figure out the median age, we sort age into ascending order:

Table 5.4.2 *Ages of those with Crash ID 19891001, 19891002, 19891003 and 19891004, highlighting the middle of the data.*

Crash ID	Age
19891003	18
19891002	**42**
19891001	**66**
19891004	76

In this case, we will have two data in the middle, 42 and 66. If we find two data in the middle, we take the average of these two data to get the median If we only wanted to find the median of Crash ID 19891001, 19891002 and 19891003 then the middle of the dataset would be 42.

Again, we do not need to do any of this by hand and we can let jamovi do the heavy lifting for us. We can find the median age in our dataset by following the same example we did above.

Mean or Median?

The mean and the median are two helpful measures. However, we need to know when we use which one. Let's say that five people are in a bar, and we examine each person's income (Table 5.4.3).

Table 5.4.3 *Income for our five bar patrons*

income	person
48000	Bella
64000	Isaac
58000	William
72000	Gabby
66000	Alex

The mean (61600.00) seems to be a pretty good summary of the income of those five people. Now let's look at what happens if Beyoncé Knowles walks into the bar (Table 5.4.4).

Table 5.4.4 *Income for our five bar patrons plus Beyoncé Knowles.*

income	person
48000	Bella
64000	Isaac
58000	William
72000	Gabby
66000	Alex
54000000	Beyoncé

The mean is now almost 10 million dollars, which is not really representative of any of the people in the bar – in particular, it is heavily driven by the outlying value of Beyoncé. In general, the mean is highly sensitive to extreme values, which is why it's always important to ensure that there are no extreme values when using the mean to summarise data.

We will go back to these concepts in the later chapters.

Mode

Calculating the mode is very simple. It is the value that occurs most frequently. Sometimes we wish to describe the central tendency of a dataset that is not numeric. For example, let's say that we want to know which models of iPhone are most commonly used. To test this, we could ask a large group of iPhone users which model each person owns. If we were to take the average of these values, we might see that the mean iPhone model is 9.51, which is clearly nonsensical, since the iPhone model numbers are not meant to be quantitative measurements. In this case, a more appropriate measure of central tendency is the mode, the most common value in the dataset.

Similar to mean and the median, jamovi can calculate the mode by following the steps detailed above.

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

5.5: Variability - How Spread Out are the Values?

Once we have described the central tendency of the data, we often also want to describe how variable the data is – this is sometimes also referred to as **dispersion**. In other words, we want to how widely dispersed the data is as it helps us make sense of the data we have.

Let's say that we've gathered data on the heights of 250 university students in a statistics subject and also the heights of a sample of 250 professional basketball players from Australia's National Basketball League. Let's assume that the average height for both groups is about 170 centimetres.

Using the descriptive statistics we've discussed so far, you might initially think that the heights of the university students and the basketball players are quite similar. However, this is not the case. Anyone who's ever watched a basketball game would notice that those basketball players look like they all have the same height – otherwise, shorter players may be disadvantaged given that basketball is known as the sport for the "giants".[1]

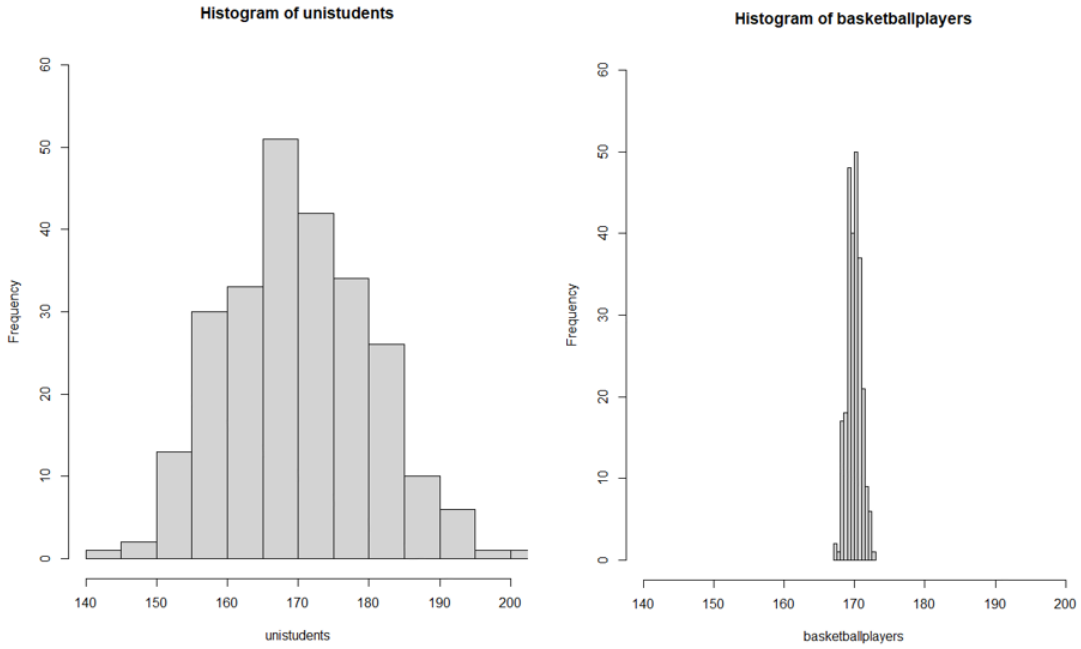


Figure 5.5.1. Heights of university students versus basketball players

While both groups have approximately the same "average" height, the university students have a much broader range of heights around that midpoint, indicating that their heights vary more and extend further from the middle. In contrast, the basketball players might all appear to have similar heights, making the heights of the uni students seem more diverse. To put it as simply as possible, if all our collected scores were found to be the same, there would be no variability. The more different the values collected are, the more variability there is.

This dispersion in height is an important characteristic when describing these two groups because it allows us to quantify precision and uncertainty.

Variance and Standard Deviation

So, let's dive into our dataset of university students' heights to explore how we can measure this variation. For simplicity's sake, let's just look at 5 of the student heights. The table below displays these heights in centimetres and how much their heights deviate from the mean height:

Table 5.5.1. Student heights in centimetres and how much their heights deviate from the mean height

Student Height (cm)	Height Deviation from Mean (cm)
165	– 5

170	0
175	5
180	10
185	15

We call the second column deviations. These deviations show how far each student's height is from the mean height.

Now, let's compute the average deviation or the average distance by which a student's height deviates from the mean. We start by summing the deviation scores:

$$-5 + 0 + 5 + 10 + 15 = 25$$

At first glance, it might seem like, on average, students' heights deviate by 25 centimetres from the mean, which doesn't make much sense.

However, here's where the magic of statistics comes into play. The positive and negative deviations cancel each other out, leading to an average of zero. This outcome highlights a critical point: deviations (or differences from the mean) always sum to zero. So, we're stuck at this point.

But, there's a clever solution. What if we compute the average of the absolute values of the deviations? Let's try it:

$$|5| + |0| + |5| + |10| + |15| = 35$$

35 divided by the number of students (5) gives us an average deviation of 7 centimetres. This makes sense and tells us that, on average, a student's height deviates by about 7 centimetres from the mean height.

However, this is not the standard deviation; when we compute the standard deviation, we don't use absolute values. Instead, we square the deviations. So, let's calculate the average of the squared deviations:

$$(-5)^2 + (0)^2 + (5)^2 + (10)^2 + (15)^2 = 350$$

The square root of 350 is approximately 18.71 centimeters. This is the standard deviation, and it tells us that, on average, a student's height deviates by about 18.71 centimetres from the mean height.

In summary, understanding dispersion in the context of university students' heights helps us measure how much individual heights vary from the average. It's not just about the average height; it's about the range of individual heights, and the standard deviation is a valuable tool for quantifying this variation.

Range

Range is another measure of variable in the data. The range is the difference between the lowest score and the highest score – the largest value minus the smallest value. For example, we may have a scale with a possible scoring range of 1 to 50.

However, it is also possible that no one in the sample scored higher than 40. Therefore, we have a "theoretical" range (i.e., the *possible* range) and the "actual range" (i.e., the range we actually got when we collected the data).

The interquartile range (IQR) is like a "fence" that encloses the middle portion of your data. It is calculated as the difference between the third quartile (the 75th percentile) and the first quartile (the 25th percentile). It helps us understand how tightly or loosely packed the data is, while ignoring outliers.

Available measures of variability in jamovi

Under Analyses > Exploration, you will find the available measures of variability in jamovi

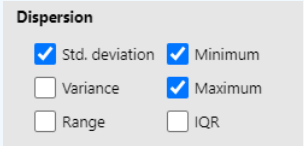


Figure 5.5.2. Available measures of dispersion in jamovi

Homogeneity of Variance

Homogeneity of variance (also known as homoscedasticity) is an important concept in statistics, particularly when we are comparing two different groups or examining relationships between two different variables. It can be a difficult name to say, but it's not a difficult concept to comprehend. Homogeneity comes from the word homogenous – meaning "same". Therefore, it refers to the idea that the variability (or spread) of data points should be roughly the *same* across different groups or levels of the variables.

Let's go back to the example above regarding the heights of university students and basketball players, as you can see in Figure 5.5.1, the basketball players' heights are tightly clustered together (low variance), and university students' heights are scattered widely (high variance). In this example, we *do not* have homogeneity of variance. This makes it challenging to draw meaningful conclusions because the variability within groups is not consistent.

If, instead, we compared the heights of university students and the heights of teachers, then it is more likely that the heights are spread out fairly evenly within each group. In other words, the spread of scores would be *fairly homogenous*. It is easier to compare the groups because you don't have one group with extremely tightly clustered scores and another with widely dispersed scores.

Homogeneity of variance is important because it ensures that the assumptions of various statistical tests are met (which we will learn more later on). When the assumption of homoscedasticity is violated, it can affect the validity and reliability of the results. To address this, researchers might need to use data transformation techniques or different statistical methods designed to handle unequal variances.

Chapter attribution

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

1. Although it is important to note that there are great basketball players who are less than 6 feet (or 183 cm). For an entertaining read: https://www.theguardian.com/sport/2022/nov/15/nba-basketball-height-tall-players-advantages ↵

5.6: Z Scores

Having characterised a distribution in terms of its central tendency and variability, it is often useful to express the individual scores in terms of where they sit with respect to the overall distribution. Let's say that we are interested in characterising the relative level of crashes across different states, in order to determine whether NSW is a particularly dangerous place for drivers.

Using our crash_data_ardd dataset, we can see that NSW had the highest number of fatalities from crashes in 2019.

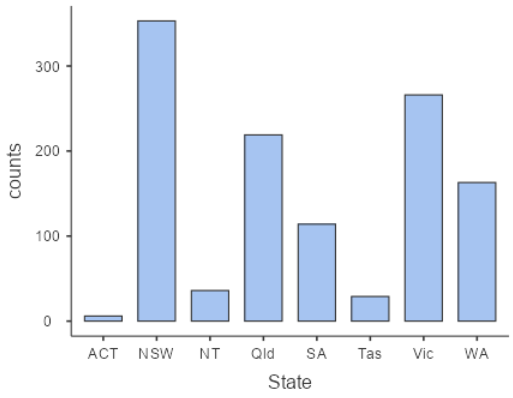
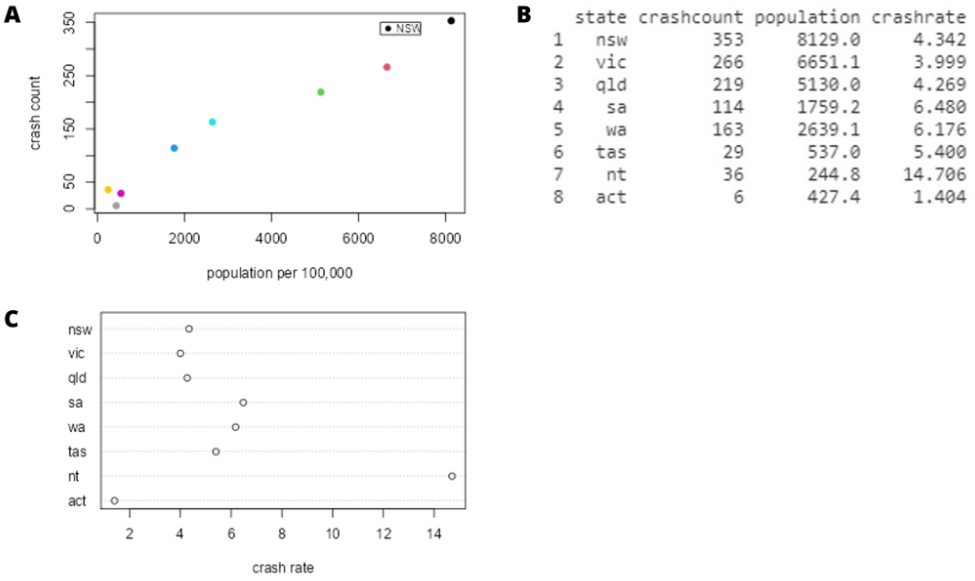


Figure 5.6.1. Crashes per state

It may have occurred to you, however, that NSW also has the largest population of any state in Australia, so it's reasonable that it will also have a larger number of crashes. If we plot the number of crashes against the population of each state (see Figure 5.6.2-A), we see that there is a direct relationship between the two variables. Instead of using the raw numbers of crashes, we should instead use crashes per capita, which we obtain by dividing the number of crashes per state by the population of each state. The original ARDD dataset did not have this information so I had to source this data from the Australian Bureau of Statistics.[1] Looking at the right panel of Figure 5.6.2 (B), we can see that NSW is not too bad after all. NSW has a crash rate of 4.342, which is slightly lower than the average crash rate of 5.847 across the states and territories. But what if we want to get a clearer view of how far it is from the rest of the distribution?



	state	crashcount	population	crashrate
1	nsw	353	8129.0	4.342
2	vic	266	6651.1	3.999
3	qld	219	5130.0	4.269
4	sa	114	1759.2	6.480
5	wa	163	2639.1	6.176
6	tas	29	537.0	5.400
7	nt	36	244.8	14.706
8	act	6	427.4	1.404

Figure 5.6.2. Crash count on y-axis and population per 100,000 per state (A) and crash rate per 100,000 population per state in table form (B) and dot chart form (C)

The ***z-score*** allows us to express data in a way that provides more insight into each data point's relationship to the overall distribution. It is calculated by subtracting the mean μ from the individual data point and then dividing by the standard deviation σ.

$$z = \frac{x-\mu}{\sigma}$$

Intuitively, you can think of a z-score as telling you how far away any data point is from the mean, in units of standard deviation. The z-score is positive if the value is above the mean and negative if it is below the mean. Calculating z-scores, allows researchers to calculate the probability of a score occurring within a standard normal distribution, and enables us to compare two scores that are from different samples or scales (which may have different means and standard deviations). Because of this, it has many practical applications. For example, z-scores can be used to compare the performance of students on different tests, even if the tests have different difficulty levels.

Z-scores are also used in many statistical tests. For example, the t-test is used to test for a difference in means between two groups (which we will learn more about later). The z-test is used to test for a difference in proportion between two groups. Both of these tests use z-scores to calculate the p-value, which is a measure of the statistical significance of the result.

We can compute the z-scores of the crash rates and let's plot the z-scores against the original crash rate, as shown in Figure 5.6.3. As you can see, the scatterplot shows us that the process of z-scoring doesn't change the relative distribution of the data points (visible in the fact that the original data and z-scored data still fall on a straight line when plotted against each other) – it just shifts them to have a mean of zero and a standard deviation of one. This provides us with a slightly more interpretable view of the data. If we look at Figure 5.6.3 again, we can see that NSW's crash rate is quite similar to the other states. Instead, we should be more worried about the Northern Territory (depicted with the yellow dot), as the state has a crash rate that is roughly two standard deviations above the mean.[2]

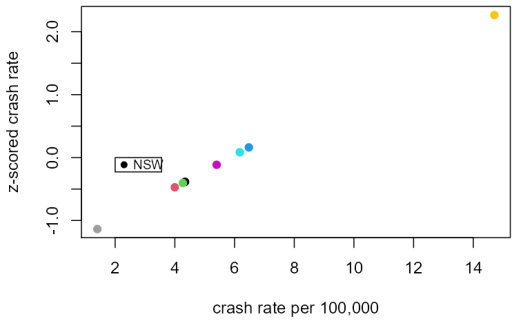


Figure 5.6.3. Z-scored crash rate on y-axis and crash rate per 100,000 on x-axis per state

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

1. https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/dec-2019 ↵
2. ABC reports that the Northern Territory leads in road deaths per capita. To read more, visit https://www.abc.net.au/news/2022-09-08/nt-road-toll-death-rate-four-times-national-average/101383908 ↵

CHAPTER OVERVIEW

Chapter 6: Modelling Variations

Learning Objectives

After reading this chapter, you should be able to:

- describe the basic equation for statistical models (data=model + error)
- distinguish between a population and a sample, and between population parameters and sample statistics
- describe the concepts of sampling error and sampling distribution
- describe how the Central Limit Theorem determines the nature of the sampling distribution of the mean.

In this chapter, we will delve into big ideas in statistics–**Modelling, Uncertainty and Sampling from Population**. As mentioned in Chapter 1, one of the fundamental activities in statistics is creating models that can summarise data using a small set of numbers, thus providing a compact description of the data.

Another foundational idea in statistics is that we can make **inferences** about an entire population based on a relatively small sample of individuals from that population. In this chapter, we will introduce the concept of statistical sampling and discuss why it works. As Charles Wheelan, the author of *Naked Statistics* aptly describes the process of inference as using *data from the "known world" to make informed inferences about the "unknown world."* There will always be uncertainties in our data and this could be due to the fact that we usually sample from a population. Therefore, to understand how we can use statistics to make these inferences, we will also talk about sampling, the central limit theory and null hypothesis testing.

6.1: A Simple Model
6.2: Statistical Modelling Using a Single Number
6.3: Sampling and Sampling Error
6.4: The Central Limit Theorem
6.5: Null Hypothesis Testing
6.6: Quantifying Effects

6.1: A Simple Model

What is a Model?

In the physical world, “models” are generally simplifications of things in the real world that nonetheless convey the essence of the thing being modelled. For instance, a model of a building conveys the structure of the building while being small and light enough to pick up with one’s hands (Figure 6.1.1-A). A model of a neuron that you see in textbooks is usually much larger than the actual thing, but it conveys the major parts of the cell and their relationships (Figure 6.1.1-B).

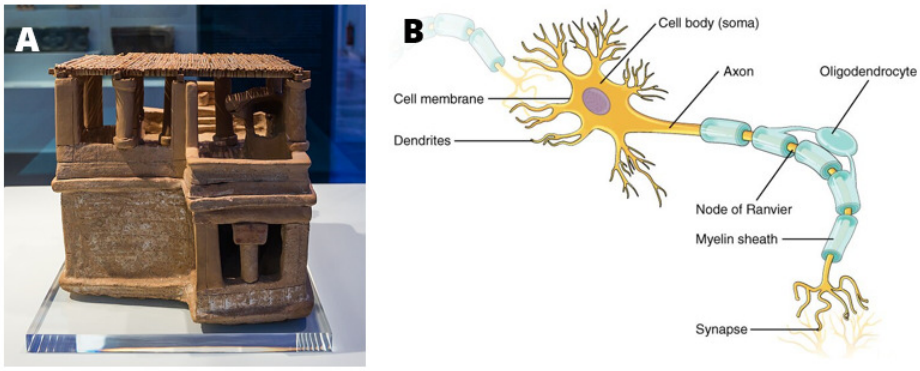


Figure 6.1.1. *A model of a house (A) and a model of a neuron you would typically see in a textbook (B). House by Jebulon is licensed under CC0. Model of neuron by OpenStax is licensed under CC BY 4.0*

In statistics, a model is meant to provide a similarly condensed description, but for data rather than a physical structure. Like physical models, a statistical model is generally much simpler than the data being described; it is meant to capture the “essence” of the data as simply as possible. In both cases, we realise that the model is a convenient fiction that necessarily glosses over some of the details of the actual thing being modelled. As the statistician George Box famously said: “All models are wrong but some are useful.”[1]

It can also be useful to think of a statistical model as a theory of how the observed data was generated; our goal then becomes to find the model that most efficiently and accurately summarises this data generation process. But as we will see later on, the desires for efficiency and accuracy will often be diametrically opposed to one another.

Let’s start with a simple model explained using an analogy.

Last weekend, I went to Bunnings[2] as I wanted to get a portable key holder for when I go diving with my partner.

Figure 6.1.2. A photo of a master lock key. Note: I am not affiliated with this product at all – I am just providing a visual image.

While in the store, I was presented with different products to choose from. I had two main criteria–1) our car keys needed to fit inside the key holder, and 2) it must be less than $50. The second criterion was easy to achieve, I just need to look at all the products that are less than $50. The first criterion (and arguably, the most important criterion) is a bit trickier to achieve because unfortunately, I left my car keys with my partner.

So I have a problem, I needed to estimate the dimensions of my car keys. Granted, I could have gone back to the car, got the keys and measured it against the locks. But alas, laziness took over.

One way of solving my problem is to model the dimensions of my car keys using the length of another object that is similar to it. I had my wallet, and sometimes, I put my keys into my wallet–so I got a measuring tape (I am in a hardware store after all) to measure the length of my wallet.

Figure 6.1.3. My wallet and my car keys

The first thing you would have noticed was that, from the outset, my model is not great. In fact, it's a terrible model. Although my model has a length, there are a lot of details that were not captured. For instance, I don't know the width and depth of the remote attached to the keys. Plus as you can see in Figure 6.1.3 above, my wallet is much longer than my keys.

If we were to write an equation in words to represent this model, it might look something like this:

$$Dimension\ of\ Keys = Length\ of\ Wallet + Error$$

The "error" bit in the equation represents the deviations from the model. We want to minimise errors as much as possible. But as you can see from above, there were a lot of errors in my model. Using the length of my wallet as a model for my keys is a gross oversimplification of the different necessary attributes of the real thing. According to CourseKata (2020), *models are always like this: they oversimplify some aspects of the world, and focus only on the dimension you are most interested in*.[3]

In the end, the portable key lock I bought was too big (and too expensive) and had to return to the store the next day to exchange it for a more suitable one. This time, I measured the dimensions of my keys before going back to the store.

1. https://en.wikipedia.org/wiki/All_models_are_wrong ↵
2. For non-Australian readers, Bunnings is an Australian household hardware and garden centre chain. It is akin to Home Depot. ↵
3. CourseKata (2020), Chapter 5: A simple model. In *Statistics and data science: A modelling approach*. https://coursekata.org/preview/book/e7ab06dd-53fd-4397-930b-72f21bcb1efb/lesson/8/0 ↵

6.2: Statistical Modelling Using a Single Number

Below is another way that we can represent the model that you saw in the previous chapter:

$$Dimension\ Of\ Keys = Length\ Of\ Wallet + Error$$

$$data = model + error$$

$data = model + error$ expresses the idea that the data can be broken into two portions: one portion that is described by a statistical model (the values that we expect the data to take, given our knowledge), and the other portion is *error* – the difference between the model's predictions and the observed data. In this section, we will start learning about statistical modelling using a single number.

Modelling Height Using a Single Number

Before reading on, prepare the datafile in jamovi by doing the exercise below:

jamovi exercise

We will use another dataset for the following example. The American *National Health and Nutrition Examination Survey* data set contains data on scores of variables. The data available is equivalent to a "a simple random sample from the American population" (Pruim 2015).[1]

In total, 10,000 observations on scores of variables are available (from the 2009/2010 and the 2011/2012 surveys). For our example in this section, we will try to build a model of the height of children in the NHANES dataset. First, let's load the data and plot them. The data can be downloaded from this link. [2]

Since the data includes all participants, not just children, we will use the filter function to conduct our analysis on those aged under 18. We do this by going to **Data** > **Filter** and typing next to the function button **Age < 18** and clicking somewhere in the window for the filter to take effect. To make it clearer for future analysis, you can write in the description: **Filter for those under 18.** You can see that the filter works when you see that some rows will be greyed out. Specifically, rows, where the participants were aged over 18, are not selected. See the image below to follow the procedure.

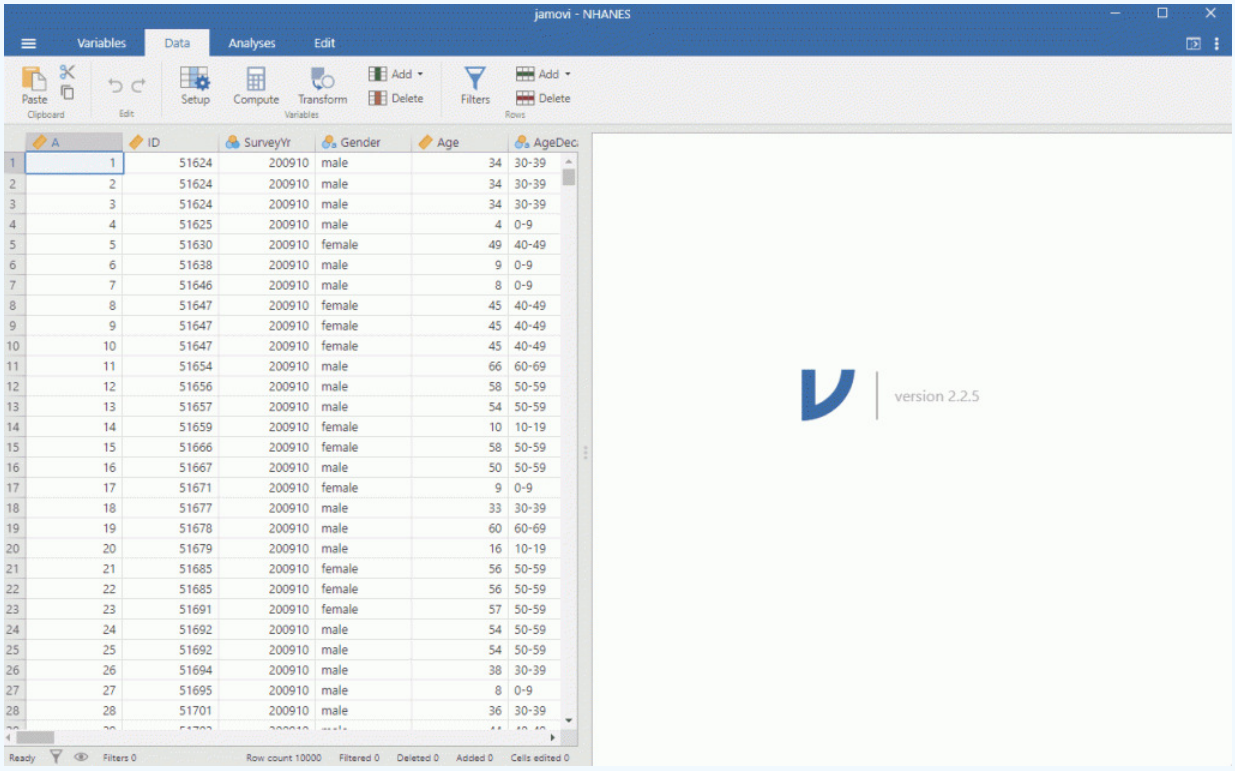


Figure 6.2.1. Using the filter button in jamovi.

As mentioned in the previous chapter, we always want to visualise the data to see what's going on. So, let's create a distribution by going to **Analyses > Exploration > put Height under Variable**. Under **Plots**, check **Histogram and Density**. You should get the following graph:

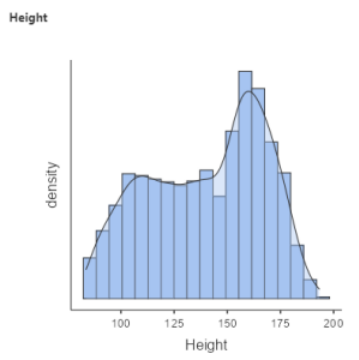


Figure 6.2.2 *Histogram of height of children in NHANES.*

As discussed previously, we use models to describe and communicate our data in a simple way. If we want to communicate the results of the height of children gathered from the NHANES dataset, we wouldn't just report this with an excel spreadsheet of 2,223 datapoints and say: "here you go!".

Remember that we want to describe the data as simply as possible while still capturing their important features. The simplest model that we can imagine would involve only a single number.

Statistical Lingo – Parameters

Whenever there are tricky concepts that I think readers may just gloss their eyes over, I'll put the text in a separate box. The information on these white boxes are not essential to have an overall grasp of the concepts discussed, but they are useful in understanding the nuances.

In statistics, we generally describe a model in terms of its ***parameters***. Parameters, in this instance, is just a fancy name to denote a numerical value that we can change in order to modify the predictions of the model. In this book and in other statistical texts, we refer to these using the Greek letter beta β. When the model has more than one parameter, we will use subscripted numbers to denote the different betas for example: (β_1, β_2).

It is also customary to refer to the values of the data using the letter y, and to use a subscripted version y_1, y_2, y_3, y_4 (and so on) to refer to the individual observations.

We generally don't know the true values of the parameters, so we have to estimate them from the data. For this reason, we will generally put a "hat" over the $\hat{\beta}$ symbol to denote that we are using an estimate of the parameter value rather than its true value (which we generally don't know).

So, how do we create a simple model to describe the data as simply as possible? One very simple estimator that we might imagine is the *mode*, which you have learnt, is the most common value in the dataset. This redescribes the entire set of 2,223 children in our data in terms of a single number. If we wanted to predict the height of any new children, then our predicted value would be the same number:

Using the word equation: $childheight_i = mode + error$

More specific equation: $\hat{y}_i = 166.5$

So in other words, if we were to get a new participant under 18 and ask for their height, based on our model, we predict that this new participant will have a height of 166.5 cm.

But anyone who has been around children would know that this is not the best estimate. I, for one, have a height of 153 cm. While Filipinos are generally shorter than the average Americans[3] (the data is based on American heights), this may not be the best model.

So, How Good of a Model is This?

In general, we define the goodness of a model in terms of the magnitude of the error, which represents the degree to which the data diverges from the model's predictions. All things being equal, the model that produces lower error is the better model (though as we will see later, all things are usually not equal…).

We can actually calculate this error using the formula below. The error for each individual is the difference between the predicted value $\hat{y}_i$ and their actual height y_i:

Using the word equation: $childheight_{error} = actualheight - predictedheight$

More specific equation: $error_i = y_i - \hat{y}_i$

Using the calculation above, the average child has a fairly large error of -28.4 centimeters when we use the mode as our estimator for our model parameter. This does not seem very good on its face value. You can calculate this yourself in jamovi by creating a computed variable using the formula, 166.5 – Height. Then use **Explore** to look at the average error rate.

Using the Mean as the Model

How might we find a better estimator for our model parameter?

We might start by trying to find an estimator that gives us an average error of zero. One good candidate is the mean. As previously discussed, we use the mean to measure the "central tendency" of a dataset – that is, what value the data is centered around? Most people don't think of computing a mean as fitting a model to data. However, that's exactly what we are doing when we compute the mean.

It turns out that if we use the mean as our estimator then the total average error will indeed be zero. You can calculate this yourself in jamovi by creating a computed variable using the formula: Average height – Individual height. In other words 137.5 – Individual height. Then use Explore to look at the average error rate. If you create a new variable in jamovi with the error from the mean, you will that there is some degree of error for each individual score – some are positive, some are negative. These will cancel each other out to give an average error of zero.

We want to take into account the magnitude of the error, regardless of its direction. In other words, we want to ignore the positive and negative aspects of the error. A common way to summarise errors to take account of their magnitude is to square the errors. If we take the average of the squared errors from the mean, instead of getting zero, we will get the value of 724.24.

There are several ways to summarise the squared error, but they all relate to each other. First, we could simply add them up – this is referred to as the *sum of squared errors (SSE)*. The magnitude of errors from this method relies heavily on the number of data points. So it can be difficult to interpret unless we are looking at the same number of observations. Second, we could take the average rate of the squared error values. This is referred to as the ***mean squared error (MSE)***. However, because we squared the values before averaging, they are not on the same scale as the original data. For the height example we have above, the new scale would be $centimeters^2$. For this reason, it's also common to take the square root of the MSE, which we refer to as the ***root mean squared error (RMSE)***, so that the error is measured in the same units as the original values (in this example, centimetres).

So, what can we do with this information? One way we can do this is to calculate the RMSE if we used the mode as an estimator and compare it to the RMSE if we used the mean as an estimator. The RMSE from the mode is 39.42 compared to the RMSE from the mean which is 26.91. The mean still has a substantial amount of error, but it's much better than the mode.

When is the Mode Most Useful as a Model?

As mentioned in the last chapter, the mode represents the value that occurs most frequently. The mode is most useful when we wish to describe the central tendency of a dataset that is not numeric. In other words, the mode is most meaningful when applied to values that are discrete in nature (for instance, the count of children or the frequency of past arrests) and when dealing with categorical variables (such as gender or political affiliation).

The Beauty of Sums of Squares

As mentioned above, the sum of squared errors (most commonly known as the sum of squares) is one way to summarise the squared error. The sum of squares is a measure of how much variability or spread there is in a set of data. This concept is something you'll see over and over again in statistics. Therefore, this is an important concept to understand.

How we do calculate the sum of squares? We take each person's height, subtract the average height from it, and square the result. Then you add up all those squared differences. The resulting number is the sum of squares. See the equation below – notice that we just added a few bits and pieces to the error equation presented above.

Using the word equation: $SS_{childheight} = total(actualheight - predictedheight)^2$

More specific equation: $SS = \Sigma(y_i - \hat{y}_i)^2$

So what does this tell us? The sum of squares is a measure of how much variation there is in the heights of the group. If everyone is exactly the same height, the sum of squares will be zero. But if there is a lot of variation in heights, the sum of squares will be larger.

Why is this useful? Well, in statistical modelling, we often want to know how much of the variation in a set of data can be explained by another variable. For example, if we're trying to predict someone's weight based on their height, we might want to know how much of the variation in weight can be explained by height. We can use the sum of squares to help us answer this question.

By comparing the sum of squares for different variables, we can see which variable explains the most variation in the data. This can help us build better models and make better predictions.

The Dark Side of the Mean

The minimisation of SSE is a good feature, and it's why the mean is the most commonly used statistic to summarise data. However, the mean also has a dark side. As explained in the previous chapter, the mean is highly sensitive to extreme values, which is why it's always important to ensure that there are no extreme values when using the mean to summarise data.

Summarising Data Robustly Using the Median

If we want to summarise the data in a way that is less sensitive to outliers, we can use the *median* to do so. While the mean minimises the sum of squared errors, the median minimises a slightly different quantity: ***The sum of the absolute value of errors***. This explains why it is less sensitive to outliers – squaring is going to exacerbate the effect of large errors compared to taking the absolute value.

In saying this however, mean is still regarded as an overall "best" estimator in the sense that it will vary less from sample to sample compared to other estimators. It's up to us to decide whether that is worth the sensitivity to potential outliers – statistics is all about tradeoffs.

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

---

1. Pruim, Randall. 2015. NHANES: Data from the US National Health and Nutrition Examination Study. https://CRAN.R-project.org/package=NHANES. ↵
2. The original data set is sourced from https://bookdown.org/pkaldunn/DataFiles/NHANES.html#ref-data:NHANES:Rpackage ↵
3. see https://en.wikipedia.org/wiki/Average_human_height_by_country ↵

6.3: Sampling and Sampling Error

Anyone living in Australia will be familiar with the concept of sampling from the political polls that have become a central part of our electoral process. In some cases, these polls can be incredibly accurate at predicting the outcomes of elections.

Nate Silver, an American statistician, was named one of the globe's top 100 influential individuals by Time magazine in 2009 after correctly predicting electoral outcomes for 49/50 states in 2008 (he had a better outcome in 2012 when he correctly predicted all 50 states). Silver did this by combining data from 21 different polls, which vary in the degree to which they tend to lean towards either the Republican or Democratic side. Each of these polls included data from about 1000 likely voters – meaning that Silver was able to almost perfectly predict the pattern of votes of more than 125 million voters using data from only about 21,000 people, along with other knowledge (such as how those states have voted in the past).

We don't have an equivalent of Nate Silver in Australia, but a recent article by the Australian Financial Review stated that Australian polls are getting more reliable.[1] According to the AFR, *pollsters attribute the uptick in accuracy to better sampling methods.*

How do we Sample?

Our goal in sampling is to determine the value of a statistic for an entire population of interest, using just a small subset of the population. We do this primarily to save time and effort – why go to the trouble of measuring every individual in the population when just a small sample is sufficient to accurately estimate the statistic of interest?

In the election example, the population is all registered voters in the region being polled, and the sample is the set of 1000 individuals selected by the polling organisation. The way in which we select the sample is critical to ensuring that the sample is representative of the entire population, which is the main goal of statistical sampling. It's easy to imagine a non-representative sample; if a pollster only called individuals whose names they had received from the local Greens Party, then it would be unlikely that the results of the poll would be representative of the population as a whole. In general, we would define a *representative poll as being one in which every member of the population has an equal chance of being selected*. When this fails, then we have to worry about whether the statistic that we computed using the sample is ***biased*** – that is, whether its value is systematically different from the population value (which we will refer to as a parameter). Keep in mind that we generally don't know this population parameter, because if we did then we wouldn't need to sample!

However, every now and again, we will use examples that have access to entire populations in order to explain some key ideas about sampling. Like the example used below.

Sampling Error and Standard Error of the Mean

Regardless of how representative our sample is, *it's likely that the statistic that we compute from the sample is going to differ at least slightly from the population parameter*. We refer to this as sampling error.

Suppose I wanted to know the average age of students in my stats class. The population average is 32.9 – this is the number that I am interested in knowing. However, it is usually rare to know the population parameter. In our example, maybe some of the students in the class are not comfortable sharing their age. So, instead, we take a sample.

Let's say we ask three people in the class and the responses we receive are 28, 29, and 36 years. By calculating the average of these ages, we obtain a sample mean of 31.0. This number is not too different from the population mean of 32.9. However, what if we get sample 2's average instead (see Figure 6.3.1 below), consisting of individuals aged 36, 40, and 45. In this sample, we have a mean of 40.3 – this sample has a larger sampling error. Yet again, we conducted another sampling, where the ages reported were 24, 28, and 29, and calculated the average age for this new group, we would obtain a sample of 27.0.

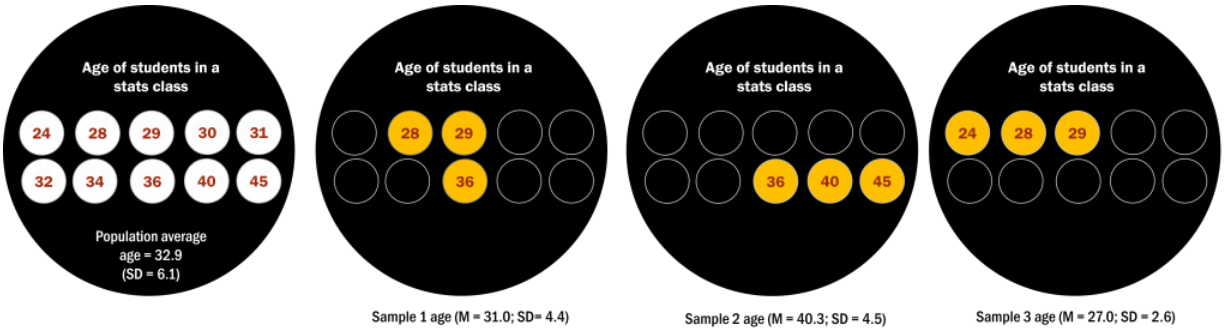


Figure 6.3.1. Age of students in a stats class (population, sample 1, sample 2 and sample 3)

The amount of variation between the average age of the different samples is a measure of **sampling error**. Sampling error is directly related to the quality of our measurement of the population**.** Clearly we want the estimates obtained from our sample to be as close as possible to the true value of the population parameter. However, even if our statistic is unbiased (that is, we expect it to have the same value as the population parameter), the value for any particular estimate will differ from the population value, and those differences will be greater when the sampling error is greater. Thus, reducing sampling error is an important step towards better measurement.

As you can see, it's important to characterise how variable our sample is, in order to make inferences about the sample statistic. If sampling error is the difference between a sample statistic and its corresponding population parameter, then the **standard error of the mean (SEM)** is a measure of how much sampling error is expected in the sample mean. In other words, the SEM is telling you how much discrepancy you can expect between the sample mean and the true population mean.

To compute SEM, we divide the estimated standard deviation by the square root of the sample size:

$$SEM = \frac{s}{\sqrt{(n)}}$$

with s as the standard deviation of the sample, and n is the sample size. Looking at the formula, you can intuitively see that the quality of our estimate will be dependent on two things, the variability of the population and the size of the sample. In other words, the larger the sample we have, the more accurate it will be in estimating the population mean due to a lower SEM. As you can see in Figure 6.3.2, the SEM for sample 4 with an n of 8 is much closer to the population SEM, compared to sample 1 with an n of 3.

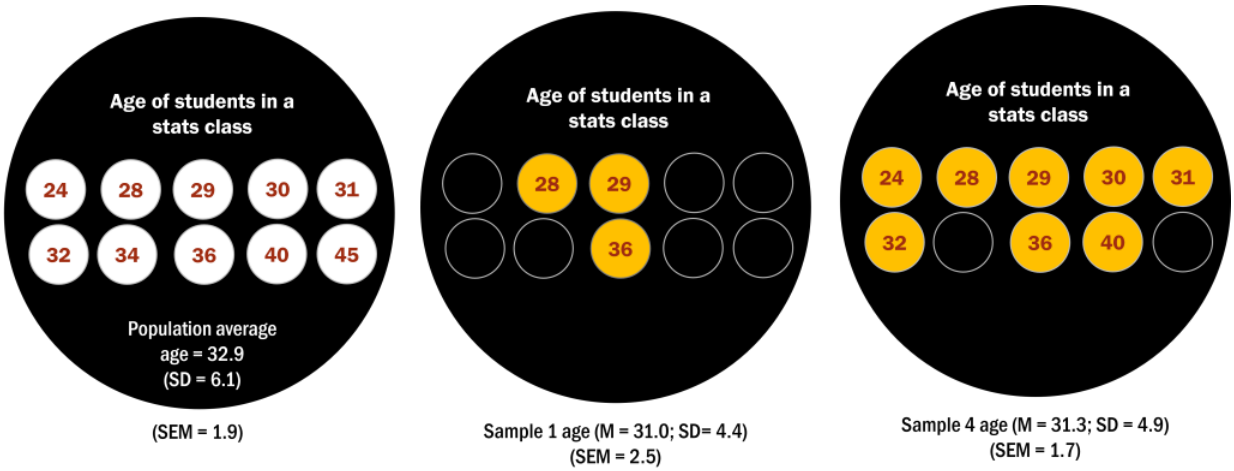


Figure 6.3.2. Population SEM, Sample 1 and Sample 4

If the variation in the population is large, then we should expect a more noisy estimate. We have no control over the population variability, but we *do* have control over the sample size. Thus, if we wish to improve our sample statistics (by reducing their sampling variability) then we should use larger samples. However, the formula also tells us something very fundamental about statistical sampling – namely, that the utility of larger samples diminishes with the square root of the sample size. This means that doubling the sample size will *not* double the quality of the statistics; rather, it will improve it by a factor of . Later on, we will discuss statistical power, which is intimately tied to this idea.

Chapter attribution

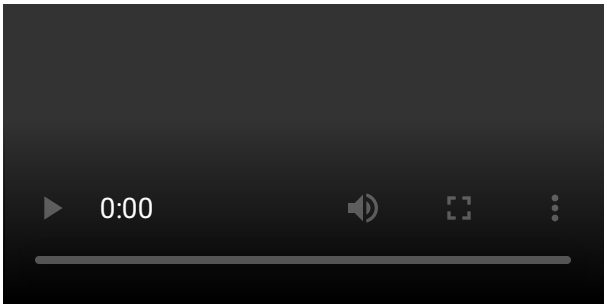
This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

1. *The power of data: Why polling is more accurate since Brexit and Trump*. https://www.afr.com/politics/federal/the-power-of-data-why-polling-is-more-accurate-since-brexit-and-trump-20230719-p5dph6 ↵

6.4: The Central Limit Theorem

The following video shows the Galton box in action. It is a device used to illustrate the principles of probability and the Gaussian distribution, which is often referred to as the normal distribution or bell curve. As you can see, seemingly random events of balls being released will accumulate into the bottom forming a bell-shaped curve. This visual device shows how randomness can lead to predictable patterns when many random events are combined. This also demonstrates the idea of regression to the mean.

Most importantly, it also demonstrates the central limit theorem (CLT), which states that, with enough sample size, the data will approximate to a normal distribution, *regardless of the original distribution of those variables.*



Media 6.4.1. Galton Box by Matemateca (IME/USP)/Rodrigo Tetsuo Argenton, licensed under CC BY-SA 4.0

The normal distribution is described in terms of two parameters: the mean (which you can think of as the location of the peak), and the standard deviation (which specifies the width of the distribution). The bell-like shape of the distribution never changes, only the location of the peak and width. The normal distribution is commonly observed in data collected in the real world – and the central limit theorem gives us some insight into why that occurs.

Let's use a module available within jamovi to understand how central limit theorem works. Under Modules, install CLT – Demonstrations. This is a simple jamovi module that contains simulations to help students visualise important lessons in probability, such as how the law of big numbers and how central limit theorem work. Students can also use this module to visualise correlations of different sizes and grasp important concepts when testing hypotheses.

Under this module, we can see how central limit theorem works by looking at different sources of distribution (e.g., normal, uniform, lognormal, etc.) and how the distribution of the sample means can lead to a normal distribution if the sample is large enough. For example, the right panel in Figure 6.4.1 shows the source distribution which has a lognormal shape. Lognormal distributions are commonly seen in variables such as the population's wealth when the data is skewed to the right. In other words, the distribution has a long tail towards the right.

Now, let's look at the sampling distribution of the mean for this source distribution (left panel in Figure 6.4.1). This graph was simulated by repeatedly drawing 500 trials from the source distribution and taking the mean. Despite the clear non-normality of the original data, the sampling distribution is remarkably close to the normal.

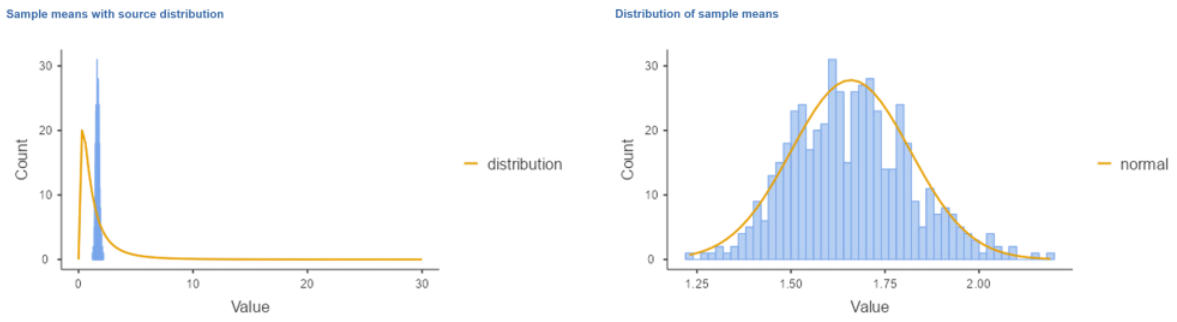


Figure 6.4.1. Demonstration of the central limit theorem using the CLT – Demonstrations module in jamovi. The panel on the right shows the source distribution with a lognormal shape. The panel on the left shows the distribution of the sample means with a normal distribution.

The central limit theorem is important to understand because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution. It's also important because it tells us why normal distributions are so common in the real world: any time we combine many different factors into a single number, the result is likely to be a normal distribution. For example, the height of any adult depends on a complex mixture of their genetics and experience — when we get enough data on height, the resulting distribution of the data will be in a bell-shaped curve.

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

6.5: Null Hypothesis Testing

In the first chapter, we discussed the three major goals of statistics: to describe, decide and predict. So far, we have talked about how we can use statistics to describe. In the next few sections, we will introduce the ideas behind the use of statistics to make decisions – in particular, decisions about whether a particular hypothesis is supported by the data.

The specific type of hypothesis testing that we will discuss is known (for reasons that will become clear) as ***null hypothesis statistical testing (NHST)***. You would be very familiar with this concept if you ever read any scientific literature. In their introductory psychology textbook, Gerrig et al. (2002)[1] referred to NHST as the "backbone of psychological research". Thus, learning how to use and interpret the results from hypothesis testing is essential to understanding the results from many fields of research.

It is also important for you to know, however, that NHST is deeply flawed, and that many statisticians and researchers (including myself) think that it has been the cause of serious problems in science. For more than 50 years, there have been calls to abandon NHST in favour of other approaches (like those that we will discuss in the following chapters).

$$H_0 : BMI_{active} = BMI_{inactive}$$

In words: *There is no difference in BMI between people who do not engage in physical activity compared to those who do.*

$$H_A : BMI_{active} \neq BMI_{inactive}$$

In words: *There will be a difference in BMI between people who do not engage in physical activity compared to those who do.*

A directional hypothesis, on the other hand, predicts which direction the difference would go. For example, we have strong prior knowledge to predict that people who engage in physical activity should weigh less than those who do not, so we would propose the following directional alternative hypothesis:

Figure 6.5.1 shows an example of such a sample, with BMI shown separately for active and inactive individuals, and Table 6.5.1. shows summary statistics for each group.

Table 6.5.1. *Summary of BMI data for active versus inactive individuals*

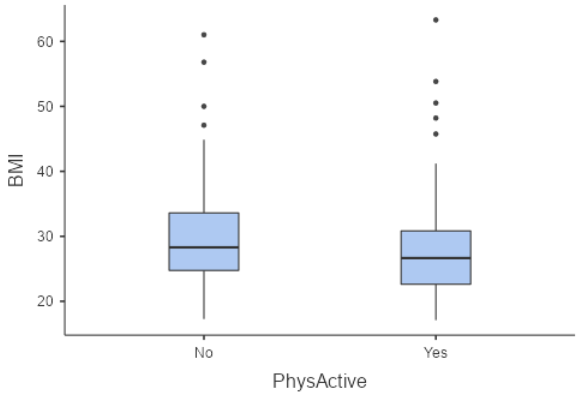


Figure 6.5.1. Box plot of BMI data from a sample of adults from the NHANES dataset, split by whether they reported engaging in regular physical activity

Step 4: Fit a Model to the Data and Compute a Test Statistic

Figure 6.5.2).

In your coin toss, you got 52 heads. Does that give you some doubts that maybe the coin is not fair after all? Well, maybe not because the probability of getting 52 heads out of 100 flips is still quite high (see below). In fact, if we use a probability calculator, the probability of getting at least 52 heads is 0.38 or 38% chance of success. 38% is quite close to 50% – so it could be just by chance that we got 52 heads out of 100 tosses.

What if, in our 100 tosses, we got 66 heads instead? Maybe doubts are starting to form. Around 2 thirds of the tosses are coming up as heads. If you can see that the 66 heads out of 100 tosses got you questioning the fairness of the coin (compared to the other sample) – then congratulations, you have the right mindset to understand the intuition behind hypothesis testing!

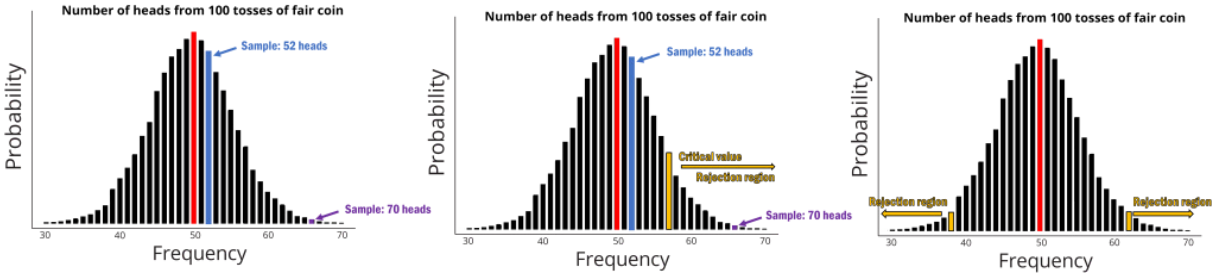


Figure 6.5.2. Are the $1 coins head-biased? The right panel shows two different samples. The middle panel shows the critical regions where we reject the null hypothesis when we have a specific direction for a hypothesis and the left panel shows the critical regions if we have a non-directional hypothesis

Basically, when we conduct NHST, we are essentially testing if the result that we get is extreme enough. We then make a decision that this result is not due to chance. With the example we provided above, the probability of getting at least 66 heads is 0.00089 – in other words, there is only 0.089% chance that we will get 66 heads out of 100 tosses!

As researchers, we try to make it easy on ourselves by appointing a value that we use to decide whether our result is rare enough to reject the null hypothesis. This is where critical values come in. Critical values are designated points in which we are happy to say that the result is rare enough that it may not be due to chance. Critical values can either be on only one side of the probability distribution (Figure 6.5.2 – middle panel) or on both sides of the probability distribution (Figure 6.5.2 – left panel).

Our hypothesis determines whether we use one side or two sides of the probability distribution (we call these one-tailed or two-tailed tests). If we are asking if the coin is biased in general (i.e., we don't know whether it is tail-biased or head-biased) then you will use these two critical points. If our results fall *beyond* either side of the critical value region, then we can reject the null hypothesis. If we are asking if the coin is only head-biased (or tail-biased) then we will use the one-tailed test. If our results fall *beyond* on our chosen side of the critical value region, then we can reject the null hypothesis.

Computing p-Values Using the t Distribution

Now, let's go back to our BMI example and compute a p-value for our BMI example using the *t* distribution. First, we compute the *t* statistic using the values from our sample that we calculated above, where we find that $t = 2.38$. The question that we then want to ask is: What is the likelihood that we would find a *t* statistic of this size, if the true difference between groups is zero (i.e. the directional null hypothesis)?

We can use the *t* distribution to determine this probability. According to jamovi, the probability value (or more commonly known as p-value) of getting t = 2.38 with a df of 275.83 is 0.018. In other words, our p-value is 0.018. This tells us that our observed *t* statistic value of 2.38 is relatively unlikely if the null hypothesis really is true.

The p-value provided in jamovi is usually set to non-directional hypothesis as a default (see under Hypothesis, the Group 1 =\= Group 2 will be selected). If we want to use a directional hypothesis (in which we only look at one end of the null distribution), then we click the option Group 1 > Group 2 or Group 1 < Group 2. What you choose will depend on what you have written for your hypothesis.

In our data, group 1 was the physically non-active group and group 2 is the physically active group. If we used the one-tailed test with Group 1 > Group 2, then we get a p-value of 0.009. Here, we see that the p value for the two-tailed test is twice as large as that for the one-tailed test, which reflects the fact that an extreme value is less surprising since it could have occurred in either direction.

How do you choose whether to use a one-tailed versus a two-tailed test?

The two-tailed test is always going to be more conservative, so it's always a good bet to use that one unless you had a very strong prior reason for using a one-tailed test. In that case, you should have written down the hypothesis before you ever looked at the data. In Chapter 4, we discussed the idea of pre-registration of hypotheses, which formalises the idea of writing down your hypotheses before you ever see the actual data. You should *never* make a decision about how to perform a hypothesis test once you have looked at the data, as this can introduce serious bias into the results.

Step 6: Assess the "Statistical Significance" of the Result

The next step is to determine whether the p-value that results from the previous step is small enough that we are willing to reject the null hypothesis and conclude instead that the alternative is true. How much evidence do we require? This is one of the most controversial questions in statistics, in part because it requires a subjective judgment – there is no "correct" answer.

Historically, the most common answer to this question has been that we should reject the null hypothesis if the p-value is less than 0.05. This comes from the writings of Ronald Fisher, who has been referred to as "the single most important figure in 20th-century statistics" (Efron 1998):

> *If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 … it is convenient to draw the line at about the level at which we can say: Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials. (Fisher, 1925)*[3]

However, Fisher never intended $p < 0.05$ to be a fixed rule:

> *no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. (Fisher, 1956)*[4]

Instead, it is likely that $p < 0.05$ became a ritual due to its simplicity – before computing, people had to rely on reading tables of p-values. Since all tables had an entry for 0.05, it was easy to determine whether one's statistic exceeded the value needed to reach that level of significance.

The choice of statistical thresholds remains deeply controversial, and researchers have been proposing to change the default threshold to be more conservative (e.g., 0.05 to 0.005), making it substantially more stringent and thus more difficult to reject the null hypothesis. In large part, this move is due to growing concerns that the evidence obtained from a significant result at $p < 0.05$ is relatively weak.

Hypothesis Testing as Decision-Making: The Neyman-Pearson Approach

Whereas Fisher thought that the p-value could provide evidence regarding a specific hypothesis, the statisticians Jerzy Neyman and Egon Pearson disagreed vehemently. Instead, they proposed that we think of hypothesis testing in terms of its error rate in the long run:

> *no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis. But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong. (Neyman & Pearson 1933)*[5]

That is: We can't know which specific decisions are right or wrong, but if we follow the rules, we can at least know how often our decisions will be wrong in the long run.

To understand the decision making framework that Neyman and Pearson developed, we first need to discuss statistical decision making in terms of the kinds of outcomes that can occur. There are two possible states of reality (H_0 is true or H_0 is false) and two

possible decisions (reject H_0 or retain H^0).

There are two ways in which we can make a correct decision:

- We can reject H_0 when it is false (in the language of signal detection theory, we call this a *hit*).
- We can retain H_0 when it is true (somewhat confusingly in this context, this is called a *correct rejection*).

There are also two kinds of errors we can make:

- We can reject H_0 when it is actually true (we call this a *false alarm*, or *Type I error*).
- We can retain H_0 when it is actually false (we call this a *miss*, or *Type II error*).

Neyman and Pearson coined two terms to describe the probability of these two types of errors in the long run:

- P(Type I error) = α">α
- P(Type II error) = β">β

That is, if we set α">α to 0.05, then in the long run we should make a Type I error 5% of the time. Whereas it's common to set α">α as 0.05, the standard value for an acceptable level of β">β is 0.2 – that is, we are willing to accept that 20% of the time we will fail to detect a true effect when it truly exists. We will return to this later when we discuss statistical power, which is the complement of Type II error.

What does a Significant Result Mean?

There is a great deal of confusion about what p-values actually mean (Gigerenzer, 2004).[6] Let's say that we do an experiment comparing the means between conditions, and we find a difference with a p-value of .05. There are a number of possible interpretations that one might entertain.

Does it mean that the probability of the null hypothesis being true is 0.01?

No. Remember that in null hypothesis testing, the p-value is the probability of the data given the null hypothesis. For those who think in formula, p-value is calculated as P(data|H_0). It does not warrant conclusions about the probability of the null hypothesis given the data – this suggestsP(H0|data).

Does it mean that the probability that you are making the wrong decision is 0.01?

No – this suggests P(H_0|data). But remember as above that p-values are probabilities of data under H_0 – not the other way around.

Does it mean that if you ran the study again, you would obtain the same result 99% of the time?

No. The p-value is a statement about the likelihood of a particular dataset under the null; it does not allow us to make inferences about the likelihood of future events such as replication.

Does it mean that you have found a practically important effect?

No. There is an essential distinction between *statistical significance* and *practical significance*. As an example, let's say that we performed a randomised controlled trial to examine the effect of a particular diet on body weight, and we find a statistically significant effect at p<.05. What this doesn't tell us is how much weight was actually lost, which we refer to as the *effect size* (to be discussed in more detail later on. If we think about a study of weight loss, then we probably don't think that the loss of one ounce (i.e. the weight of a few potato chips) is practically significant. Let's look at our ability to detect a significant difference of 1 ounce as the sample size increases.

Figure 6.5.3 shows how the proportion of significant results increases as the sample size increases, such that with a very large sample size (about 262,000 total subjects), we will find a significant result in more than 90% of studies when there is a 1 ounce difference in weight loss between the diets. While these are statistically significant, most physicians would not consider a weight loss of one ounce to be practically or clinically significant. We will explore this relationship in more detail when we return to the concept of *statistical power* in the next few section, but it should already be clear from this example that statistical significance is not necessarily indicative of practical significance.

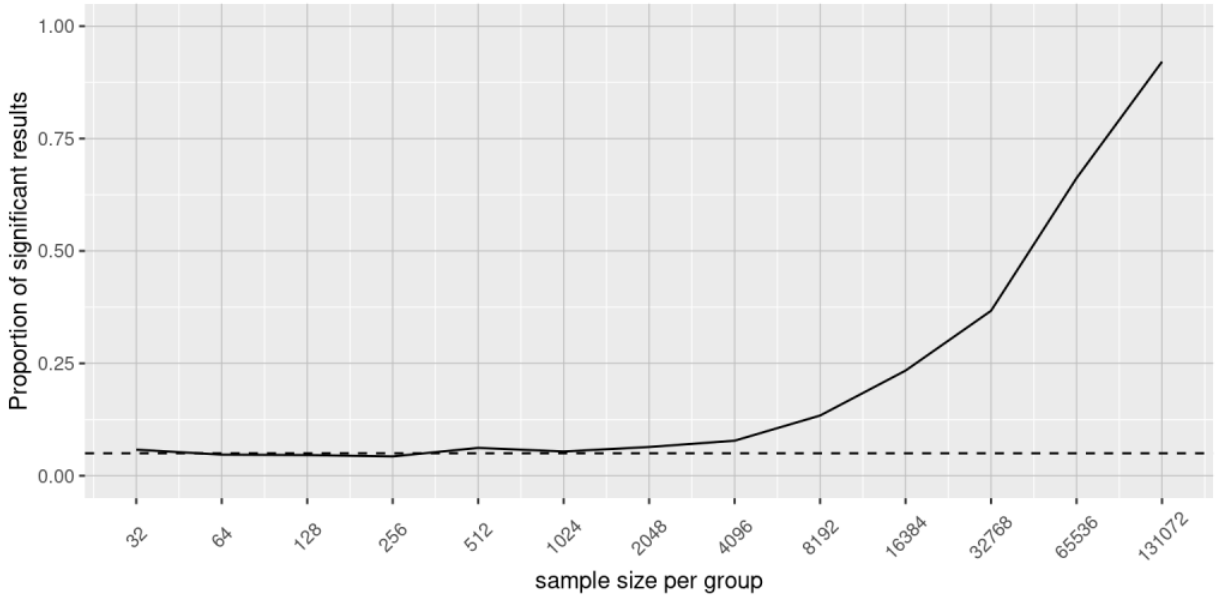


Figure 6.5.3. The proportion of significant results for a very small change (1 ounce, which is about .001 standard deviations) as a function of sample size. Image by Poldrack, licensed under CC BY-NC 4.0

In summary, being able to reject the null hypothesis is indirect evidence of the experimental or alternative hypothesis. However, rejecting the null hypothesis

- DOES NOT say whether the scientific conclusion is correct.
- DOES NOT tell us anything about the mechanism behind any differences or the relationship (e.g., does not tell us WHY there's a difference in BMI compared with physically active and physically inactive adults).
- DOES NOT tell us whether the study is well designed or well controlled.
- DOES NOT PROVE ANYTHING.

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

1. Gerrig, R. J., Zimbardo, P. G., Campbell, A. J., Cumming, S. R., & Wilkes, F. J. (2015). *Psychology and life*. Pearson Higher Education. ↵
2. Clare, J., Henstock, D., McComb, C., Newland, R., & Barnes, G. C. (2021). The results of a randomized controlled trial of police body-worn video in Australia. *Journal of Experimental Criminology*, *17*(1), 43–54. https://link.springer.com/article/10.1007/s11292-019-09387-w ↵
3. Fisher, R. A. 1925. *Statistical methods for research workers*. Oliver & Boyd. ↵
4. Efron, B. (1998). R. A. Fisher in the 21st Century (invited paper presented at the 1996 R. A. Fisher Lecture). *Statistical Science*, *13*(2), 95–122. https://doi.org/10.1214/ss/1028905930. ↵
5. Neyman, J., and K. Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 231*(694-706): 289–337. doi.org/10.1098/rsta.1933.0009. ↵
6. Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 62–88). Blackwell Publishing. doi.org/10.1002/9780470752937.ch4 ↵

6.6: Quantifying Effects

In the previous sections, we discussed how we can use data to test hypotheses. Those methods provided a binary answer: we either reject or fail to reject the null hypothesis. However, this kind of decision overlooks a couple of important questions. First, we would like to know how much uncertainty we have about the answer (regardless of which way it goes). In addition, sometimes we don't have a clear null hypothesis, so we would like to see what range of estimates are consistent with the data. Second, we would like to know how large the effect actually is, since as we saw in the weight loss example in the previous section, a statistically significant effect is not necessarily a practically important effect.

In this section, we will discuss methods to address these two questions: confidence intervals to provide a measure of our uncertainty about our estimates, and effect sizes to provide a standardised way to understand how large the effects are. We will also discuss the concept of *statistical power* which tells us how likely we are to find any true effects that actually exist.

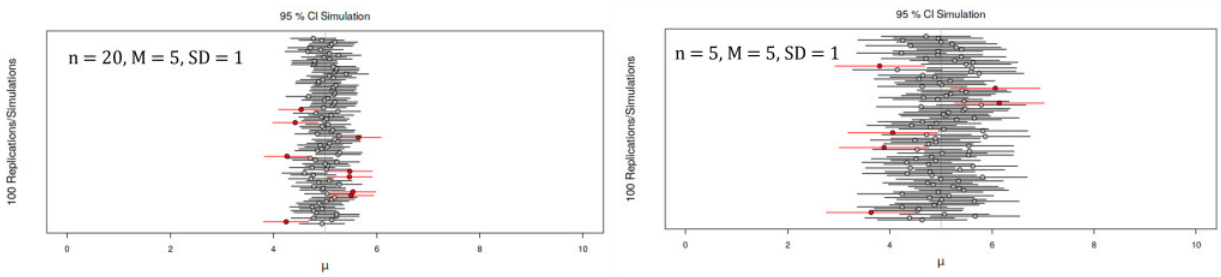


Figure 6.6.1. 95 % Confidence Intervals with different sample sizes (n) but with the same population parameters

Figure 6.6.2. Confidence intervals for the mean in jamovi

Relation of Confidence Intervals to Hypothesis Tests

There is a close relationship between confidence intervals and hypothesis tests. In particular, if the confidence interval does not include the null hypothesis, then the associated statistical test would be statistically significant. For example, if you are testing whether the mean of a sample is greater than zero with $_{\alpha}\alpha=0.05"\gt=0.05$, you could simply check to see whether zero is contained within the 95% confidence interval for the mean.

Things get trickier if we want to compare the means of two conditions or more (Schenker & Gentleman 2001).[3] In certain situations, statistical analysis is conducted by comparing the confidence intervals of the estimates to determine if there is any overlap. When the confidence intervals do not overlap, this is interpreted as indicating a statistically significant difference (as shown in Figure 6.6.3). It is generally accepted that non-overlapping confidence intervals signify statistical significance, but it's important to note that the reverse is not always true for overlapping confidence intervals (as depicted in Figure 6.6.3). For instance, what about the case where the confidence intervals overlap one another but don't contain the means for the other group? In this case, the answer depends on the relative variability of the two variables, and there is no general answer. To obtain a more precise assessment, an alternative method involves calculating the ratio or difference between the two estimates and constructing a test or confidence interval based on that particular statistic.

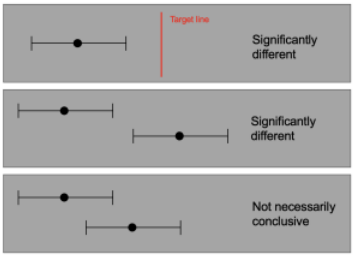


Figure 6.6.3. Using confidence intervals for making comparisons. The two top images show non-overlapping confidence intervals which can be statistically significant. The bottom image shows that overlapping confidence intervals do not always indicate a difference that is not statistically significant

While some academics suggest avoiding the "eyeball test" for overlapping confidence intervals (e.g., Poldrack, 2023), academics like Geoff Cummings are a strong advocate for using confidence intervals instead of NHST. [4]

$$d = \frac{M_1 - M_2}{S_{pooled}}$$

where M_1 and M_2 are the means of the two groups, and S_{pooled} is the pooled standard deviation (which is a combination of the standard deviations for the two samples, weighted by their sample sizes). Note that this is very similar in spirit to the t statistic – the main difference is that the denominator in the t statistic is based on the standard error of the mean, whereas the denominator in Cohen's d is based on the standard deviation of the data. This means that while the t statistic will grow as the sample size gets larger, the value of Cohen's d will remain the same.

Figure 6.6.4 shows that the two distributions are quite well separated, though still overlapping, highlighting the fact that even when there is a very large effect size for the difference between two groups, there will be individuals from each group that are more like the other group.

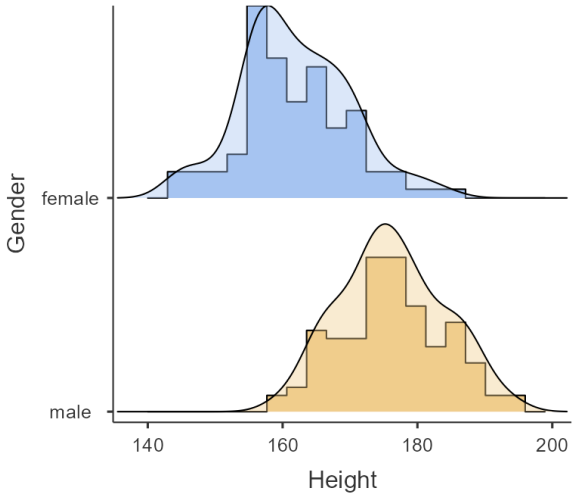


Figure 6.6.4 Histogram with density plots for male and female heights in the NHANES dataset, showing distinct but also clearly overlapping distributions. Screenshot from the jamovi program

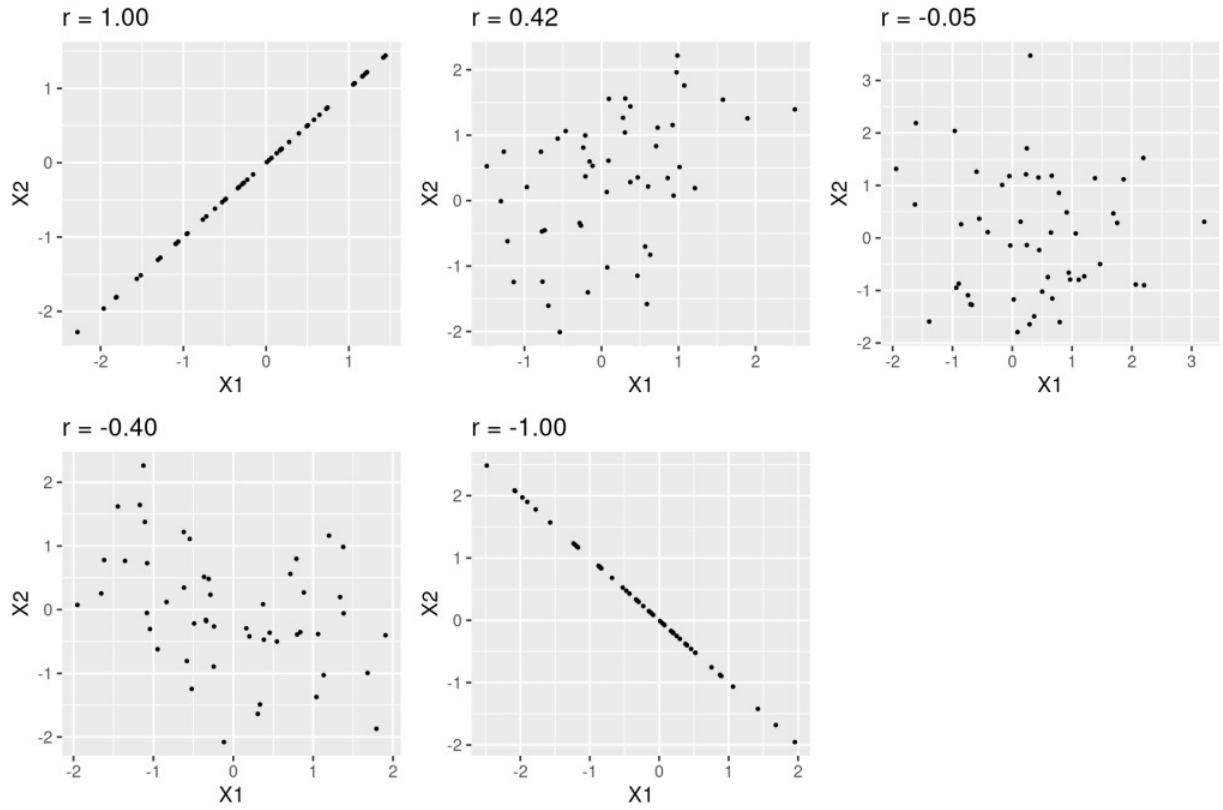


Figure 6.6.5. Examples of various levels of Pearson's r. Image by Poldrack, licenced under CC BY-NC 4.0

Figure 6.6.6 shows an example of how power changes as a function of these factors.

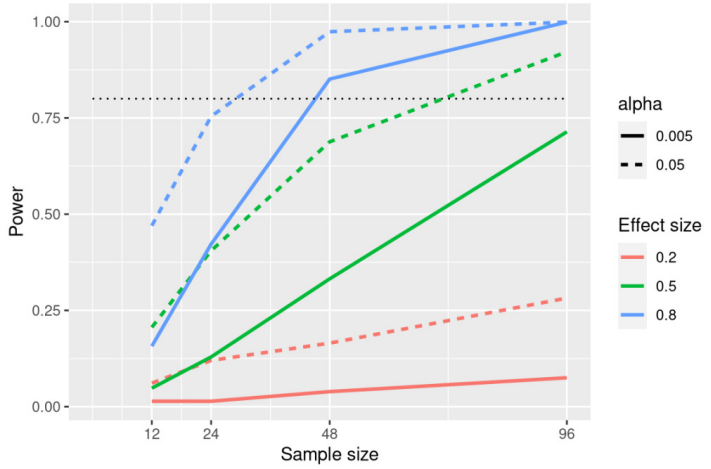


Figure 6.6.6. Results from power simulation, showing power as a function of sample size, with effect sizes shown as different colours, and alpha shown as line type. The standard criterion of 80 per cent power is shown by the dotted black line. Image by Poldrack, licensed under CC BY-NC 4.0

This simulation shows us that even with a sample size of 96, we will have relatively little power to find a small effect (d=0.2) with α=0.005. This means that a study designed to do this would be *futile* – that is, it is almost guaranteed to find nothing even if a true effect of that size exists.

There are at least two important reasons to care about statistical power. First, if you are a researcher, you probably don't want to spend your time doing futile experiments. Running an underpowered study is essentially futile because it means that there is a very low likelihood that one will find an effect, even if it exists. Second, it turns out that any positive findings that come from an underpowered study are more likely to be false compared to a well-powered study.

Power Analysis

Fortunately, there are tools available that allow us to determine the statistical power of an experiment. The most common use of these tools is in planning an experiment (i.e., *a priori* power analysis), when we would like to determine how large our sample needs to be to have sufficient power to find our effect of interest. We can also use power analysis to test for sensitivity. In order words, a priori power analysis answers the question, "How many participants do I need to detect a given effect size?" and sensitivity power analysis answers the question, "What effect sizes can I detect with a given sample size?"

In jamovi, a module called jpower allows users to conduct power analysis when conducting an independent samples t test, paired samples t test and one sample t test. This module is a good start – however, if you need another software that can accommodate other statistical tests, *G*Power* is one of the most commonly used tools for power analysis. You can find the latest version using this link.

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

1. Neyman, J. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 236*(767), 333–80. doi.org/10.1098/rsta.1937.0005 ↵
2. shiny.rit.albany.edu/stat/confidence/ ↵
3. Schenker, N., & Gentleman J. F. 2001. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, *55*(3), 182–86. www.jstor.org/stable/2685796 ↵
4. Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7-29. doi.org/10.1177/0956797613504966 ↵
5. Sullivan, G. M., & Feinn, R. (2012). Using effect size-or why the p value Is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. doi.org/10.4300/JGME-D-12-00156.1 ↵
6. Cohen, J. (1994). The earth is round (p< 0.05). *American Psychologist*, *49*(12), 997. ↵
7. Wakefield, A. J. (1999). MMR vaccination and autism. *The Lancet*, *354*(9182), 949–950. https://doi.org/10.1016/S0140-6736(05)75696-8 ↵

CHAPTER OVERVIEW

Chapter 7: The General Linear Model

Learning Objectives

After reading this chapter, you should be able to:

- describe the concept of the general linear model and provide examples of its application
- describe the concept of linear regression and apply it to a dataset.

7.1: General Linear Model
7.2: Modelling Continuous Relationships
7.3: Comparing Means
7.4: Working with Categorical Outcomes
7.5: Introduction to Multivariate Statistical Modelling

7.1: General Linear Model

> *If you should say to a mathematical statistician that you have discovered that linear multiple regression analysis and the analysis of variance (and covariance) are identical systems, he would mutter something like, "Of course – general linear model," and you might have trouble maintaining his attention. If you should say this to a typical psychologist, you would be met with incredulity, or worse. Yet it is true, and in its truth lie possibilities for more relevant and therefore more powerful exploitation of research data. (Cohen 1968)*[1]

The above quote is from Cohen (1968) when he introduced psychology to the General Linear Models (GLM) using the language of regression analysis. So, what is GLM exactly? Before we delve into this topic, let's have a history lesson.

A bit of history – Pearson vs Fisher

There are three important statisticians that I would like to introduce to you – Francis Galton, Karl Pearson and Ronald Ayman Fisher. Some of these names have been mentioned previously in the book. Francis Galton introduced the statistical concepts of correlation and regression and was credited as the first person to apply statistical methods to study human differences and intelligence. Karl Pearson, a protege of Galton, built on his mentor's work and popularised various statistical analyses such as chi-squared tests. Both Galton and Pearson's work played a big role in the eugenic movements and was criticised as scientific racism.[2]

Fisher, on the other hand, is an experimentalist who primarily worked on agriculture. He proposed the concepts of *mean generalised* and *analysis of variance* (mostly known as ANOVA) using Galton's ideas on regression and Pearson's ideas on probability distribution. Fisher also published a journal article that criticised one of Pearson's formulas while introducing the idea of degrees of freedom (Salsburg, 2001).[3] Due to his work, RA Fisher is credited as the founder of modern statistics.

This brief history lesson provides background information on why people think that regression-based analysis and group-differences analysis (such as t-tests and ANOVAs) are different Regression analysis was popularised by figures like Galton and Pearson to explore *natural variations.* A couple of decades later, Fisher developed the analysis of variance and analysis of covariance to study artificial or controlled variations due to his experimentations in agriculture (Cohen, 1968).

Over the years, misconceptions about ANOVAs, t-tests, and regression analysis began to emerge. Some individuals held the mistaken belief that ANOVAs could establish causation while thinking that regression analysis could not. There was also a misconception that regression analysis was only suitable for numerical variables, while ANOVAs were somehow more appropriate for experiments. However, there are no arguments that the statistical methods are inherently related to each other. Some statisticians even suggest that ANOVA is *just a special case of the GLM.* Some psychology educators (me included), think that this division creates unnecessary stress for students as they are led to believe that the statistical analyses they learn are distinct and disconnected.

In saying that however, I do acknowledge that learning about ANOVAs and t-tests still has its place. For instance, a commentary on Twitter argued that learning about ANOVA "forces an understanding of and respect for degrees of freedom" and "treats experiments like actual effing experiments". They also argued that "people trained in ANOVA can correctly use regression to analyze experiments. People trained in regression but not ANOVA mostly can not, in my experience", (Brewer, 2023).[4]

Nevertheless, I believe that there are more advantages to teaching statistics using the GLM approach because it's easier for students to transition into more advanced topics when there's a framework that binds all statistical concepts together. Therefore, this framework will be the basis of the current chapter.

What is the General Linear Model?

Remember that early in the book we described the basic model of statistics:

$$data = model + error$$

Where our general goal is to find the model that minimises the error, subject to some other constraints (such as keeping the model relatively simple so that we can generalise beyond our specific dataset).

You have already seen the general linear model (GLM) in the earlier chapters where we modelled height in the NHANES dataset as a function of age. As you can see in Table 7.1.1, nearly all of the statistical analyses you will encounter in a psychology statistics

course can be framed in terms of the GLM or an extension of it.

A general linear model is one in which the model for the outcome variable (which is often referred to as *Y*) is composed of a *linear combination* of predictors (which is often referred to as *X*) that are each multiplied by a weight (which is often referred to as the Greek letter beta – β), which determines the relative contribution of that predictor variable to the model prediction.

Let's try a more intuitive explanation. A general linear model is like a recipe for making predictions. Imagine you're trying to predict something, like the price of a house. In this model, you have a bunch of ingredients (predictors), like the size of the house, the number of bedrooms, and so on. Each ingredient is given a number (the beta value), which tells you how important that ingredient is for making the prediction. The bigger the beta, the more impact that *ingredient* has on the prediction.

So, you take the size of the house, multiply it by its beta, and then do the same for the number of bedrooms and all the other *ingredients*. You then add all these numbers together, which gives you the final prediction. GLM helps you figure out the best combination of these *ingredients* to make the most accurate prediction.

Table 7.1.1. The differences between the GLM equation and the statistical tests procedure that are commonly taught in undergraduate statistics courses in psychology, adapted from Fife, 2022[5] and used under a CC BY-SA licence

Procedure	GLM Equation	Interpretation
one sample t test	$y = b_0$	b0 (the intercept) is the value we are testing against
independent-sample t test	$y = b_0 + b_1 \times Treatment$	b0 (the intercept) is the mean of the control group and b1 is the difference between treatment and control groups
within sample t test	$Time_2 Time_1 = B_0$	b0 (the intercept) is the average difference from Time 1 to Time 2
ANOVA	$y = b_0 + b_1 Treatment_A \times Treatment_B$	b0 (the intercept) is the mean of the control group, b1 is the difference between Treatment A and the control, and b2 is the difference between Treatment B and the control.
ANCOVA	$y = b_0 + b_1 Covariate + b_2 \times Treatment$	b0 (the intercept) is the mean of the control group, b1 is the slope of the covariate, and b2 is the difference between the Treatment and the control group.
Factorial ANOVA	$y = b_0 + b_1 \times Treatment + b_2 \times Female + b_3 \times Female \times Treatment$	b0 (the intercept) is the mean of the men in the control group, b1 is the difference between Treatment and control, b2 is the difference between Males and Females, and b3 is the difference between females in the treatment group and males in the control group.

Why is it important to learn about the GLM? Because it eliminates the need to commit complex decision trees with intricate rules to memory when determining which statistical model to use, as depicted in the image above. This simplification is incredibly beneficial. Essentially, all you need to know is which variables you aim to predict (i.e., your outcome variable) and which variables you use for prediction (such as IQ, SES, gender, or treatment/control status).

Group Differences Versus Predictors

When discussing different categories (e.g., treatment vs. control, males vs. females, freshmen vs. seniors), we often express our interest in estimating *differences between these groups*. For example, males and females *differ* in their hand grip strength. However, you could express group differences by stating that group membership *predicts* scores on the outcome variable. For example, gender *predicts* hand grip strength.

From a mathematical standpoint, there's absolutely no distinction between estimating group differences and predicting an outcome. It's purely a matter of terminology. Some may object to the use of this language, arguing that "it's not appropriate to refer to group membership as a predictor."

Well, in practical terms, it doesn't make a mathematical difference, so why engage in semantic disputes?

Moreover, adopting a consistent terminology can simplify the decision-making process in statistical analysis. Once more, it's simply a matter of distinguishing which variables act as predictors and which one serves as the outcome.

In the next few sections, we will apply the GLM framework using different questions. We will start with assessing group differences.

Chapter attribution

This chapter contains material taken and adapted from *The Order of the Statistical Jedi* by Dustin Fife, used under a CC BY-SA 4.0 licence.

---

1. Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*(6, Pt.1), 426-443. doi.org/10.1037/h0026714 ↵
2. Nobles, M., Womack, C., Wonkam, A. & Wathuti, E. (2022, June 8). Science must overcome its racist legacy: *Nature*'s guest editors speak [Editorial]. *Nature*. https://www.nature.com/articles/d41586-022-01527-z ↵
3. Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Macmillan. ↵
4. Brewer, N. [@noelTbrewer]. (2023, August 14). *Yes and militant about it. ANOVA forces an understanding of and respect for degrees of freedoms (and familywise error), treats* [Post]. X. twitter.com/noelTbrewer/status/1690844693951631360 ↵
5. Fife, J. (2022). *The Order of the Statistical Jedi: Responsibilities, routines, and rituals*. QuantPysch. quantpsych.net/stats_modeling/the-general-linear-model.html ↵

7.2: Modelling Continuous Relationships

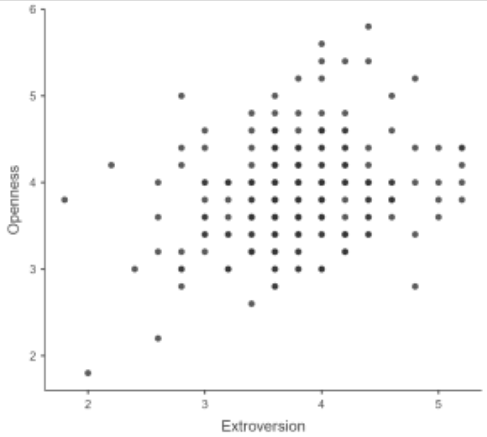


Figure 7.2.1. Scatterplot of extraversion and openness to experience

Figure 7.2.2 shows us the results:

Correlation Matrix

Correlation Matrix

		Extroversion	Openness
Extroversion	Pearson's r	—	
	p-value	—	
Openness	Pearson's r	0.28	—
	p-value	< .001	—

Figure 7.2.2. Correlation matrix for extraversion and openness to experience (screenshot from jamovi)

The correlation value of 0.28 between extroversion and openness to experience seems to indicate a reasonably moderate positive relationship between the two. The p-value above shows that the likelihood of an *r* value this extreme or more is quite low under the null hypothesis, so we would reject the null hypothesis of $r = 0$. Note that this test assumes that both variables are normally distributed.

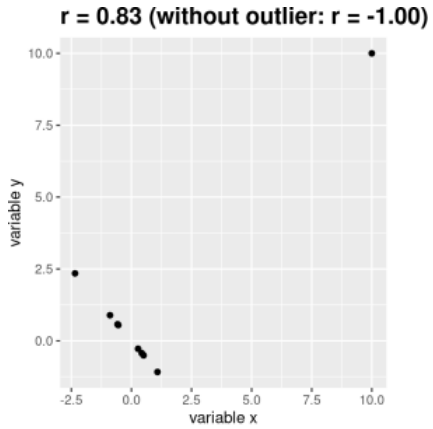


Figure 7.2.3. A simulated example of the effects of outliers on correlation. Without the outlier, the remainder of the data points have a perfect negative correlation, but the single outlier changes the correlation value to be strongly positive. Image by Poldrack, licensed under CC BY-NC 4.0

One way to address outliers is to compute the correlation on the ranks of the data after ordering them, rather than on the data themselves; this is known as the Spearman correlation. Whereas the Pearson correlation for the example above is 0.28, the Spearman correlation is 0.25, showing that the rank correlation reduces the effect of the outlier and reflects the negative relationship between the majority of the data points. Getting the Spearman correlation is really easy in jamovi, you just click this as an additional option for your results.

Correlation and Causation

When we say that one thing *causes* another, what do we mean? There is a long history in philosophy of discussion about the meaning of causality, but in statistics, one way that we commonly think of causation is in terms of experimental control. That is, if we think that factor X causes factor Y, then manipulating the value of X should also change the value of Y.

In medicine, there is a set of ideas known as *Koch's postulates* which have historically been used to determine whether a particular organism causes a disease. The basic idea is that the organism should be present in people with the disease, and not present in those without it – thus, a treatment that eliminates the organism should also eliminate the disease. Further, infecting someone with the organism should cause them to contract the disease. An example of this was seen in the work of Dr. Barry Marshall, who had a hypothesis that stomach ulcers were caused by a bacterium (*Helicobacter pylori*). To demonstrate this, he infected himself with the bacterium, and soon thereafter developed severe inflammation in his stomach. He then treated himself with an antibiotic, and his stomach soon recovered. He later won the Nobel Prize in Medicine for this work.

Often we would like to test causal hypotheses but we can't actually do an experiment, either because it's impossible ("What is the relationship between human carbon emissions and the earth's climate?") or unethical ("What are the effects of severe neglect on child brain development?"). However, we can still collect data that might be relevant to those questions. For example, we can potentially collect data from children who have been neglected as well as those who have not, and we can then ask whether their brain development differs.

Let's say that we did such an analysis, and found that neglected children had poorer brain development than non-neglected children. Would this demonstrate that neglect *causes* poorer brain development? No. Whenever we observe a statistical association between two variables, it is certainly possible that one of those two variables causes the other. However, it is also possible that both of the variables are being influenced by a third variable; in this example, it could be that child neglect is associated with family stress, which could also cause poorer brain development through less intellectual engagement, food stress, or many other possible avenues. The point is that a correlation between two variables generally tells us that something is *probably* causing something else, but it doesn't tell us what is causing what.

Figure 7.2.4 shows the causal relationships between study time and two variables that we think should be affected by it: exam grades and exam finishing times.

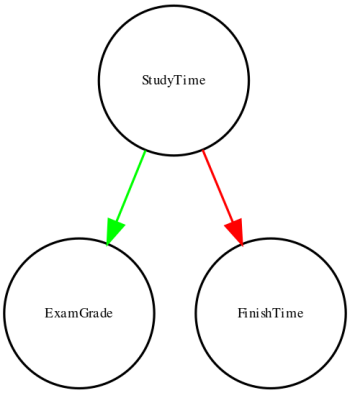


Figure 7.2.4. A graph showing causal relationships between three variables: study time, exam grades, and exam finishing time. A green arrow represents a positive relationship (i.e. more study time causes exam grades to increase), and a red arrow represents a negative relationship (i.e. more study time causes faster completion of the exam). Image by Poldrack, licensed under CC BY-NC 4.0

However, in reality, the effects on finishing time and grades are not due directly to the amount of time spent studying, but rather to the amount of knowledge that the student gains by studying. We would usually say that knowledge is a *latent* variable – that is, we can't measure it directly but we can see it reflected in variables that we can measure (like grades and finishing times). Figure 7.2.5 shows this:

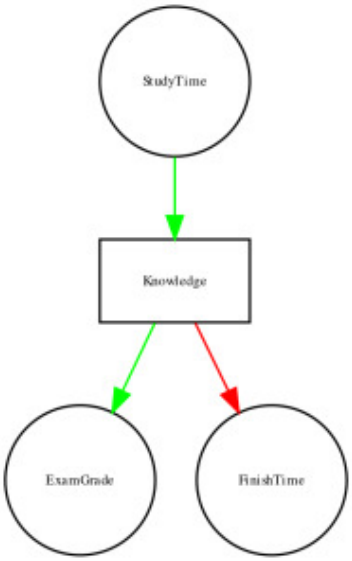


Figure 7.2.5: A graph showing the same causal relationships as above, but now also showing the latent variable (knowledge) using a square box. Image by Poldrack, licensed under CC BY-NC 4.0

$$y = x * \beta_x \beta_0 \epsilon$$

The βx"> β_x value tells us how much we would expect y to change given a one-unit change in x. The intercept β_0 is an overall offset, which tells us what value we would expect y to have when $x = 0$; you may remember from our early modelling discussion that this is important to model the overall magnitude of the data, even if xX"> never actually attains a value of zero. The error term ϵ∈"> refers to whatever is left over once the model has been fit; we often refer to these as the *residuals* from the model. If we want to know how to predict y (which we call $\hat{y}$) after we estimate the beta values, then we can drop the error term:

$$\hat{y} = x * \hat{\beta}_x + \hat{\beta}_0$$

Figure 7.2.6. The linear regression solution for extroversion and openness to experience is shown in the solid line

The value of the intercept is equivalent to the predicted value of the y variable when the x variable is equal to zero. The value of beta is equal to the slope of the line – that is, how much y changes for a unit change in x. This is shown schematically in the dashed lines, which show the degree of increase in openness to experience for a single unit increase in extroversion.

The relation between correlation and regression

There is a close relationship between correlation coefficients and regression coefficients. Remember that Pearson's correlation coefficient is computed as the ratio of the covariance and the product of the standard deviations of x and y:

$$\hat{r} = \frac{covariance_{xy}}{s_x * s_y}$$

whereas the regression beta for x is computed as:

$$\hat{\beta}_x = \frac{covariance_{xy}}{s_x * s_x}$$

Based on these two equations, we can derive the relationship between $\hat{r}$and latex]\hat{\beta}[/latex]:

$$covariance_{xy} = \hat{r} * s_x * s_y$$

$$\hat{\beta}_x = \frac{\hat{r} * s_x * s_y}{s_x * s_x} = r * \frac{s_y}{s_x}$$

That is, the regression slope is equal to the correlation value multiplied by the ratio of standard deviations of y and x. One thing this tells us is that when the standard deviations of x and y are the same (e.g. when the data have been converted to Z scores), then the correlation estimate is equal to the regression slope estimate.

Regression to the Mean

The concept of *regression to the mean* was one of Galton's essential contributions to science, and it remains a critical point to understand when we interpret the results of experimental data analyses. Let's say that we want to study the effects of a reading intervention on the performance of poor readers. To test our hypothesis, we might go into a school and recruit those individuals in the bottom 25% of the distribution on some reading test, administer the intervention, and then examine their performance on the test after the intervention. Let's say that the intervention actually has no effect, such that reading scores for each individual are simply independent samples from a normal distribution. Results from a computer simulation of this hypothetical experiment are presented in Table 7.2.1.

Table 7.2.1. Reading scores for Test 1 (which is lower, because it was the basis for selecting the students) and Test 2 (which is higher because it was not related to Test 1).

test	score
Test 1	88
Test 2	101

If we look at the difference between the mean test performance at the first and second test, it appears that the intervention has helped these students substantially, as their scores have gone up by more than ten points on the test! However, we know that in fact the students didn't improve at all, since in both cases the scores were simply selected from a random normal distribution. What has happened is that some students scored badly on the first test simply due to random chance. If we select just those subjects on the basis of their first test scores, they are guaranteed to move back towards the mean of the entire group on the second test, even if there is no effect of training. This is the reason that we always need an untreated *control group* in order to interpret any changes in performance due to an intervention; otherwise, we are likely to be tricked by regression to the mean. In addition, the participants need to be randomly assigned to the control or treatment group, so that there won't be any systematic differences between the groups (on average).

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

7.3: Comparing Means

Comparing a Mean to a Target Value

A straightforward question we could ask is whether a specific mean is higher or lower than a target value. For example, let's consider testing if the average diastolic blood pressure in adults from the NHANES dataset is greater than 80, a threshold for hypertension set by the American College of Cardiology. Imagine we randomly selected 250 adults from the dataset to explore this.

We can answer this question using Student's t-test, which you have already encountered earlier in the book. We will refer to the mean as $\bar{x}$ and the hypothesised population mean as μ. The t-test for a single mean is:

$$one\ sample\ t\ test = \frac{\bar{X} - \mu}{SEM}$$

where SEM (as you may remember from the chapter on sampling) can be calculated by using the following formula: $\frac{standard\ deviation}{\sqrt{n}}$.

In essence, the t =-statistic asks how large the deviation of the sample mean from the hypothesised quantity is with respect to the sampling variability of the mean.

To conduct one sample's t-test in jamovi, go to Analyses > Exploration > but BPDiaAve into the dependent variables. Set the test value at 80. As you learned from the previous chapters, we also want jamovi to give us effect size and descriptives.

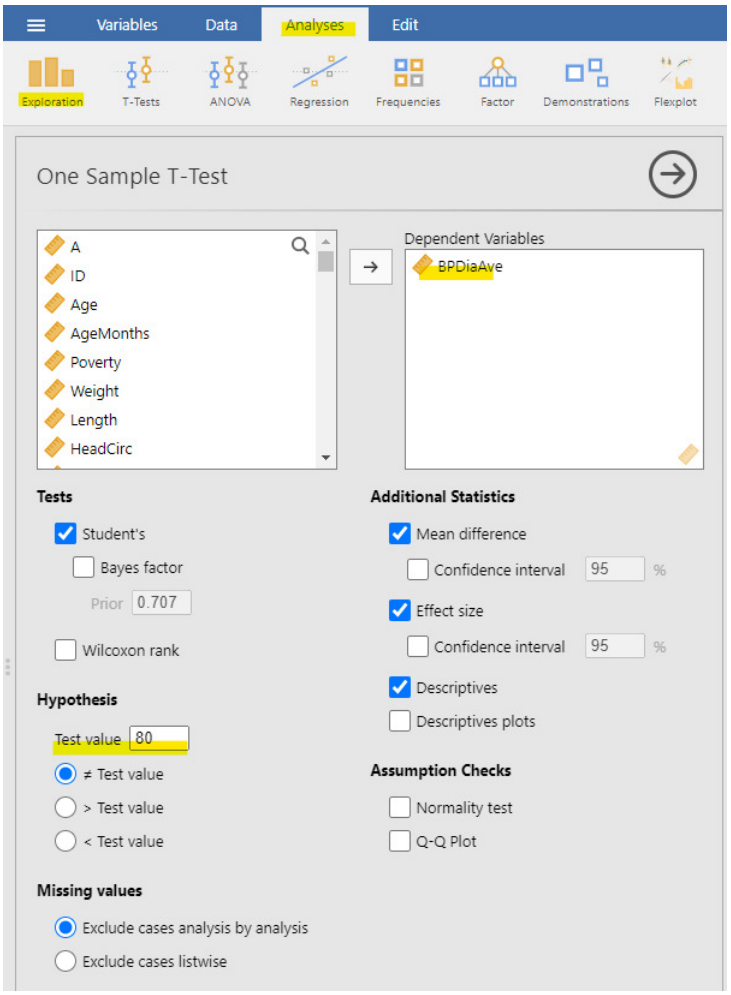


Figure 7.3.1. How to conduct one sample t-test in jamovi

This information reveals that the mean diastolic blood pressure in the dataset (70) is significantly lower than 80. Our test to check if it's above 80 is not even close to being statistically significant. It's important to remember that a large p-value doesn't offer evidence in support of the null hypothesis because we initially assumed the null hypothesis to be true.

Comparing Two Means

A more common statistical question often revolves around whether there's a difference between the averages of two different groups. For example, let's say we want to find out if regular marijuana smokers consume more alcohol during the day than non-regular smokers. We have the following hypothesis – smoking marijuana is linked to increased alcohol consumption (H_A).

We can explore this question using the NHANES dataset. We take a sample of 5% from the dataset and investigate if the amount of alcohol consumed per year is linked to regular marijuana use. In Figure 7.3.2, you can see these data visually presented with a box plot. It's evident that those who regularly use marijuana are also more likely to consume alcohol during the day.

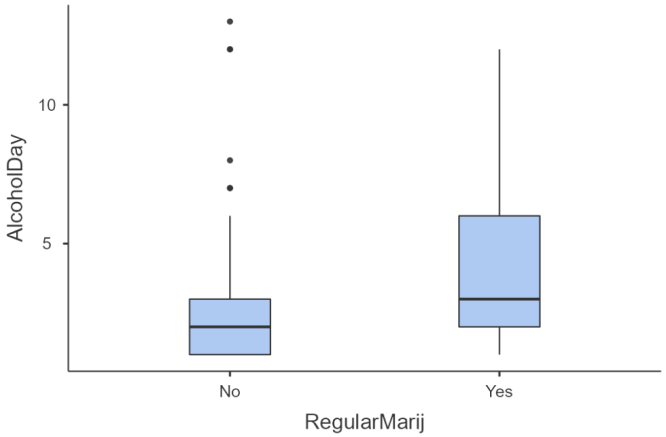


Figure 7.2.2. Box plot showing the data for the amount of alcohol drank during the day by regular marijuana use

We can also conduct the Student's t-test to assess differences between two groups of independent observations, as we discussed in an earlier chapter. As a recap, we assess the mean differences using the t-distribution. To calculate the degrees of freedom for this test, we will use the Welch test given that the group sample size differs (e.g., non-regular smokers, n = 79 versus regular smokers, n = 28). We also use the Welch test if our data violates the assumption of homogeneity of variances. We can check this in jamovi by selecting the "Homogeneity test" under Assumption Checks.

To perform the independent t-test in jamovi, follow these steps: Go to Analyses > Exploration > Place "AlcoholDay" in the dependent variables and "RegularMarij" in the Grouping Variable. As you've learned from previous chapters, we also want jamovi to provide us with effect size and descriptives. In this particular scenario, we began with a specific hypothesis that smoking marijuana is linked to increased alcohol consumption, so we'll use a one-tailed test. Under Hypothesis, select "Group 1 < Group 2," taking into account that the grouping in the "RegularMarij" variable is: 1 = No | 2 = Yes.

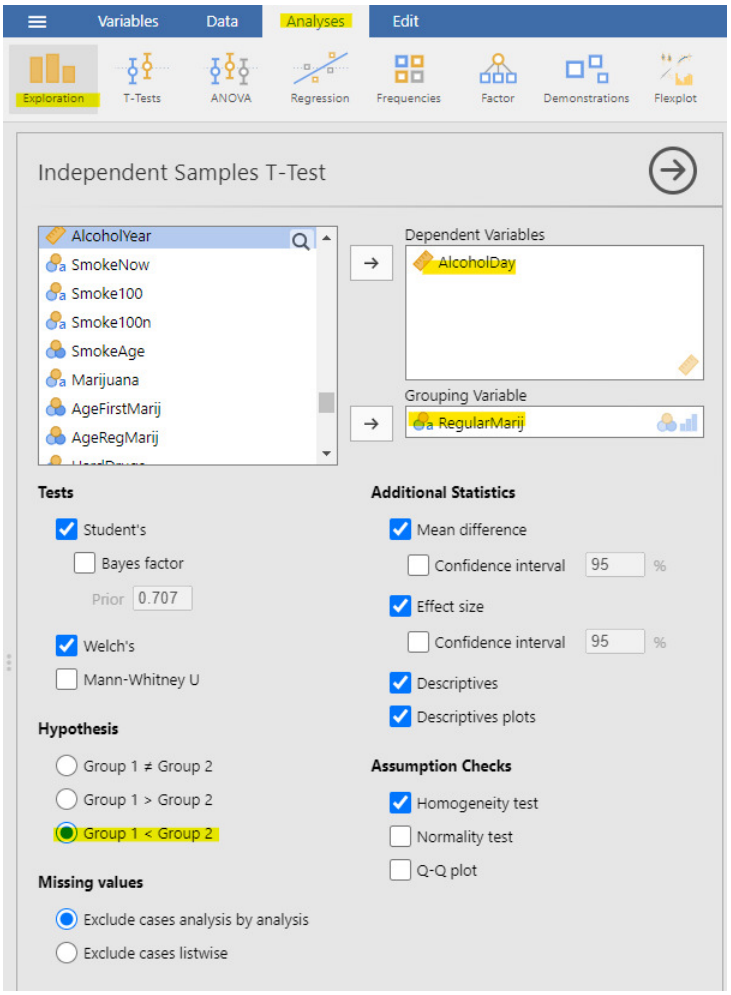


Figure 7.3.3 How to conduct independent samples t-test in jamovi

Here are the results from jamovi. We observe a statistically significant difference between the groups, as we hypothesised. Individuals who smoke marijuana are more likely to consume larger amounts of alcohol during the day.

Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference		Effect Size
AlcoholDay	Student's t	−2.66 [a]	105	0.004	−1.49	0.561	Cohen's d	−0.585
	Welch's t	−2.45	41.3	0.009	−1.49	0.610	Cohen's d	−0.560

Note. H_a $\mu_{No} < \mu_{Yes}$

[a] Levene's test is significant ($p < .05$), suggesting a violation of the assumption of equal variances

[0]

Group Descriptives

	Group	N	Mean	Median	SD	SE
AlcoholDay	No	79	2.72	2.00	2.42	0.273
	Yes	28	4.21	3.00	2.88	0.545

Plots

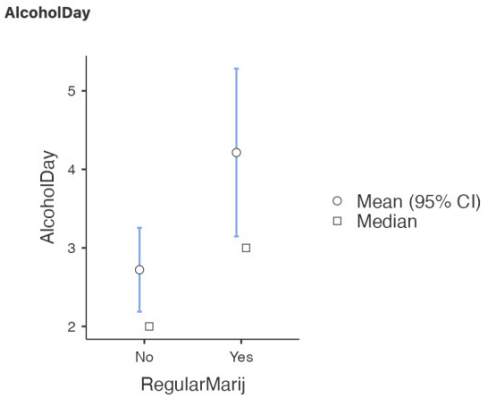


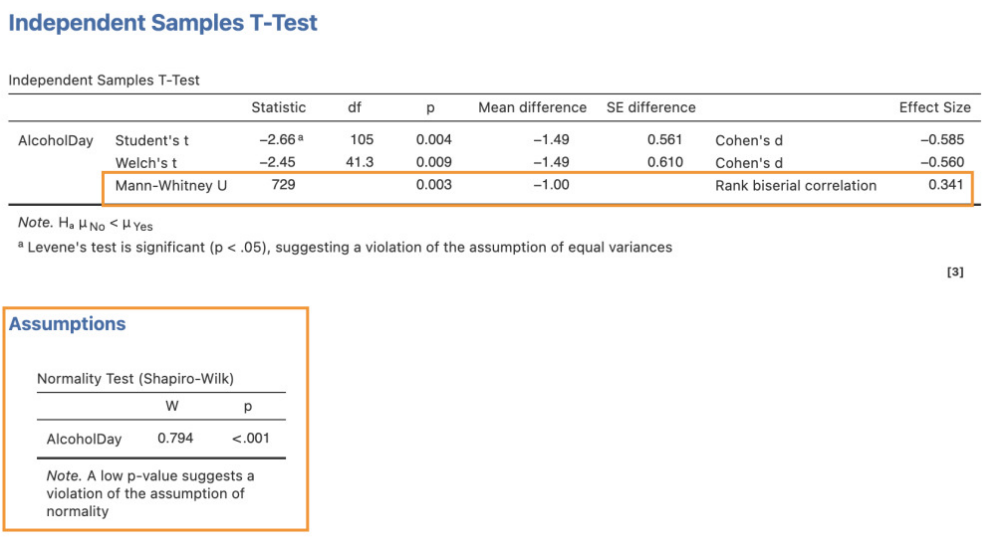
Figure 7.3.4. Results for the independent samples t-test in jamovi

Non-Parametric Independent t-Test: Mann-Whitney U

The t-test relies on the assumption that the data comes from populations with normal distributions. When dealing with small sample sizes, it can be challenging to rigorously assess this assumption. Instead of assuming that our data was sampled from normal populations, we can use the non-parametric Mann-Whitney test to assess differences between the two groups. Most statistical software can provide this test.

In jamovi, you can find this option within the independent samples t-test window. To check if our data violates the normality assumption, click on the “Normality test” under Assumption Checks. After performing the Shapiro-Wilk test, it appears that our data indeed violates the normality assumption, as indicated by the significant p-value.

Using the Mann-Whitney U test, we obtained a p-value of 0.003, which remains statistically significant.



Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference		Effect Size
AlcoholDay	Student's t	−2.66 [a]	105	0.004	−1.49	0.561	Cohen's d	−0.585
	Welch's t	−2.45	41.3	0.009	−1.49	0.610	Cohen's d	−0.560
	Mann-Whitney U	729		0.003	−1.00		Rank biserial correlation	0.341

Note. H_a $\mu_{No} < \mu_{Yes}$

[a] Levene's test is significant (p < .05), suggesting a violation of the assumption of equal variances

[3]

Assumptions

Normality Test (Shapiro-Wilk)

	W	p
AlcoholDay	0.794	<.001

Note. A low p-value suggests a violation of the assumption of normality

Figure 7.3.5. Results for Mann-Whitney U test and normality assumption testing in jamovi

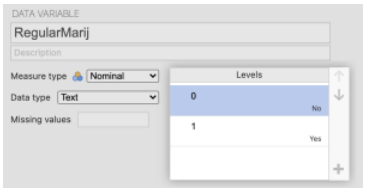
The t-Test as a Linear Model

The t-test is often presented as a specialised tool for comparing means, but it can also be viewed as an application of the GLM. In this case, the model would look like this:

$$AlcoholDayConsumption = \beta \times Marijuana_{regular} + \beta_0$$

Dummy coding

Since regular use of marijuana is a binary variable, we need to assign dummy coding to the levels of the variable. We will use 0 for non-regular users and 1 for regular users. We do this by going into double-clicking the variable you want to dummy code. This will open the data variable tab and type 0 for those who said no, and 1 for those who said yes.



7.3.6. Dummy coding in jamovi

In that case, β1 is simply the difference in means between the two groups, and β0 is the mean for the group that was coded as zero. We can fit this model using the general linear model function in our statistical software. To do this in jamovi, you have to install the module named ***gamlj***. As you can see in Figure 7.3.7 below, it will give the same t statistic above (without the Welch correction).

General Linear Model

Model Info

Info	
Estimate	Linear model fit by OLS
Call	AlcoholDay ~ 1 + RegularMarij
R-squared	0.0632
Adj. R-squared	0.0543

Model Results

ANOVA Omnibus tests

	SS	df	F	p	η^2p
Model	46.1	1	7.09	0.009	0.063
RegularMarij	46.1	1	7.09	0.009	0.063
Residuals	682.6	105			
Total	728.7	106			

Fixed Effects Parameter Estimates

				95% Confidence Interval					
Names	Effect	Estimate	SE	Lower	Upper	β	df	t	p
(Intercept)	(Intercept)	3.47	0.280	2.912	4.02	0.000	105	12.37	<.001
RegularMarij1	1 - 0	1.49	0.561	0.381	2.60	0.569	105	2.66	0.009

Plots

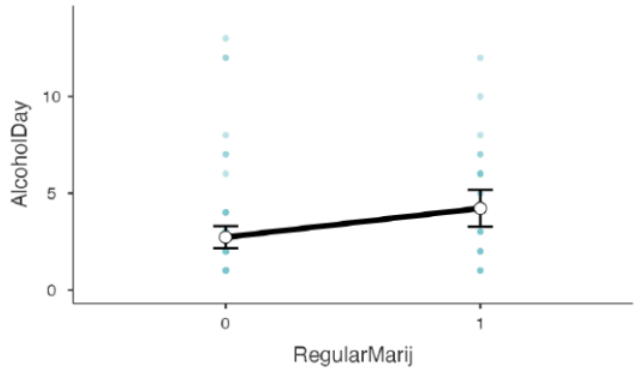


Figure 7.3.7. Results of the General Linear Model in jamovi

Comparing Paired Observations

In experimental research, we often use *within-subjects* designs, in which we compare the same person on multiple measurements. The measurements that come from this kind of design are often referred to as *repeated measures*. For example, in the NHANES dataset blood pressure was measured three times. Let's say that we are interested in testing whether there is a difference in mean systolic blood pressure between the first and second measurements across individuals in our sample.

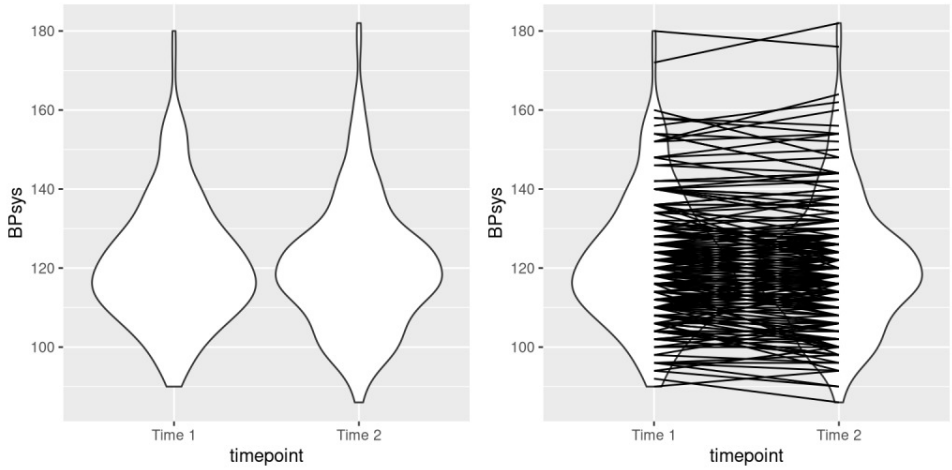


Figure 7.3.8. Left: Violin plot of systolic blood pressure on first and second recording, from NHANES. Right: Same violin plot with lines connecting the two data points for each individual

We see that there does not seem to be much of a difference in mean blood pressure (about one point) between the first and second measurements. First let's test for a difference using an independent samples t-test, which ignores the fact that pairs of data points come from the the same individuals.

This analysis shows no significant difference. However, this analysis is inappropriate since it assumes that the two samples are independent, when in fact they are not, since the data come from the same individuals. We can plot the data with a line for each individual to show this (see the right panel in Figure 7.3.8).

In this analysis, what we really care about is whether the blood pressure for each person changed in a systematic way between the two measurements, a common strategy is to use a *paired t-test*, which is equivalent to a one-sample t-test for whether the mean difference between the measurements within each person is zero. We can compute this using our statistical software, telling it that the data points are paired. With this analysis, we see that there is in fact a significant difference between the two measurements.

Comparing More than Two Means

Often we want to compare more than two means to determine whether any of them differ from one another. Let's say that we are analysing data from a clinical trial to see the efficacy of drugs in improving mood. In the study, volunteers are randomized to one of three conditions: anxifree, joyzepam or placebo. Our hypothesis is to see whether there is a significant difference in mood improvement between these three conditions. For this scenario, let's use sample data from jamovi's data library titled "Clinical Trial". Let's create a box plot for each drug with mood gain as our outcome variable (see Figure 7.3.9):

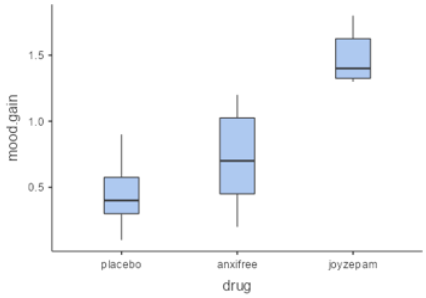


Figure 7.3.9. Box plots showing mood gain for three different groups in our clinical trial

Just from looking at the box plots above, there did seem to be differences between the groups. However, let's see if these differences are statistically significant.

$$MS_{model} = \frac{SS_{model}}{df_{model}} = \frac{SS_{model}}{k-1}$$

$$MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{SS_{error}}{N-k}$$

Where k is the number group means we have computed.

With ANOVA, we want to test whether the variance accounted for by the model is greater than what we would expect by chance, under the null hypothesis of no differences between means. Instead of the t-distribution, we use another theoretical distribution that describes how ratios of sums of squares are distributed under the null hypothesis: The *F* distribution (see Figure 7.3.10). This distribution has two degrees of freedom, which correspond to the degrees of freedom for the numerator (which in this case is the model), and the denominator (which in this case is the error).

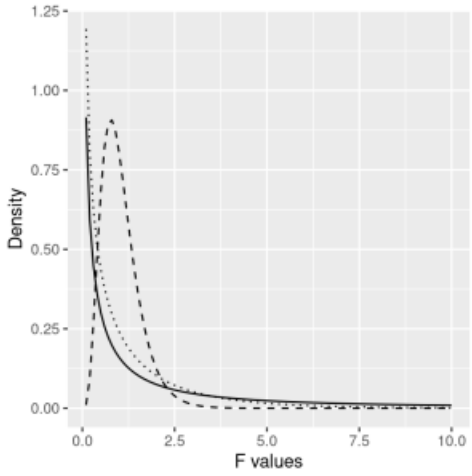


Figure 7.3.10. F distributions under the null hypothesis, for different values of degrees of freedom. Image by Poldrack, under CC BY-NC 4.0.

Figure 7.3.11. Results for ANOVA in jamovi

Remember that the hypothesis that we started out wanting to test was whether there was any difference between any of the conditions; we refer to this as an *omnibus* hypothesis test, and it is the test that is provided by the F statistic. In this case, we see that the F test is significant (p-value is < 0.001), consistent with our impression that there did seem to be differences between the groups. However, the output does not specifically inform us which of the drugs significantly differs from the placebo and by how much. If we believe that one of the drugs is not significantly different from the placebo, would it not make more sense to opt for the placebo?

We can ask jamovi to provide more tests for us. Post-hoc tests can be conducted to delve deeper into the differences between each drug and the placebo. These tests can offer a more granular understanding of the comparative effects. By utilising additional post-hoc tests in jamovi, we can enhance the precision of our analysis and make more informed decisions regarding the potential efficacy of each drug compared to the placebo.

Why not Multiple t-Tests?

Let's use the example we have above: anxifree, joyzepam and placebo. We might think of doing three separate t-tests: comparing anxifree to joyzepam, anxifree to placebo and joyzepam to placebo.

But, we don't do multiple t-tests because it increases the chance of making a Type I error. If I did three separate t-tests, set my alpha (Type I error rate) at 5% for each, and knew for sure there's actually no effect, each test has a 5% chance of making a Type I error. But since we're doing three tests, our overall error rate becomes 14.3%, not the 5% we set alpha at.

With more tests, it gets riskier:

- 1 test: 5%
- 2 tests: 9.8%
- 3 tests: 14.3%
- 4 tests: 18.6%
- 5 tests: 22.6%

- 10 tests: 40.1%
- 20 tests: 64.1%

So, doing 10 tests could have a 40% chance of showing a false positive (saying there's an effect when there isn't). To avoid this, we use one-way ANOVA as one test to see if there's a difference overall. We can also do things to control our error rate. Check out this xkcd comic for a good visual explanation.

Post-Hoc and Planned Comparisons

Post-Hoc Comparisons

Sometimes, we want to know not just if there's a difference overall (which the F-statistic tells us), but where exactly the differences are between groups. To figure that out, we use planned contrasts when we have specific ideas we want to test or post-hoc comparisons when we don't have specific ideas. It's important to mention that you only do these comparisons if the *ombinus F-statistic is statistically significant.* There's no point in looking at differences between groups if the test says there are no differences between the groups!

Here are some details about post-hoc comparisons:

- No correction: This doesn't correct for errors at all, like doing separate t-tests for each group. It's not recommended because it can mess up our error rate (as discussed above).
- **Tukey:** This is a common one. It controls errors well but isn't as strict as Bonferroni. The p-values are smaller than unadjusted but not as big as Bonferroni.
- **Scheffe:** It's complicated, and I don't use it much.
- **Bonferroni:** This is super conservative, good if you don't have many comparisons or really want to control errors. It multiplies your p-value by the number of comparisons.
- **Holm:** Like Bonferroni but adjusts p-values sequentially, making it less strict.

Remember, if you're doing Welch's F-test (unequal variances) or Kruskal-Wallis test (non-normal distribution), use the Games-Howell or DSCF pairwise comparisons, respectively.

Planned Comparisons

If you already have specific ideas about differences between groups before analyzing your data, you'd use planned contrasts. You can find these in the ANOVA setup as a drop-down menu. Just a heads up, you can't do planned contrasts with Welch's F-test or Kruskal-Wallis test.

Even though there are six contrasts in jamovi, you usually only do one. Here they are for explanation:

- Deviation: Compares each category (except the first) to the overall effect. The order is alphabetical or numerical. Placebo is considered the first category (because I have manually put this in the first level).
- **Simple:** Compares each category to the first. The order is alphabetical or numerical. Placebo is considered the first.
- **Difference:** Each category (except the first) is compared to the mean effect of all previous categories.
- **Helmert:** Each category (except the last) is compared to the mean effect of all subsequent categories.
- **Repeated:** Each category is compared to the last.
- **Polynomial:** Tests trends in the data. It looks at the n-1th degree based on the number of groups. For example, with 3 groups, it tests linear (1) and quadratic (2) trends. If there were 5 groups, it would test linear (1), quadratic (2), cubic (3), and quartic (4) trends. Note: Your factor levels must be ordinal for a polynomial contrast to make sense.

Running ANOVA as a GLM

We can also just run ANOVA as a GLM using the methods above. Using GLM, jamovi will also provide the ANOVA omnibus tests and you will see that the F test is identical to the ANOVA method. GLM will also provide us with the result of a t-test for each of the conditions, which basically tells us whether each of the conditions separately differs from placebo; it appears that Drug 2 (joyzepam) does whereas Drug 1 (anxifree) does not. However, keep in mind that if we wanted to interpret these tests, we would need to correct the p-values to account for the fact that we have done multiple hypothesis tests (otherwise, we are inflating our error).

General Linear Model

Model Info

Info	
Estimate	Linear model fit by OLS
Call	mood.gain ~ 1 + drug
R-squared	0.71
Adj. R-squared	0.67

[3]

Model Results

ANOVA Omnibus tests

	SS	df	F	p
Model	3.45	2	18.61	< .001
drug	3.45	2	18.61	< .001
Residuals	1.39	15		
Total	4.85	17		

Fixed Effects Parameter Estimates

				95% Confidence Interval					
Names	Effect	Estimate	SE	Lower	Upper	β	df	t	p
(Intercept)	(Intercept)	0.88	0.07	0.73	1.04	0.00	15	12.30	< .001
drug1	1 - 0	0.27	0.18	−0.11	0.64	0.50	15	1.52	0.150
drug2	2 - 0	1.03	0.18	0.66	1.41	1.94	15	5.88	< .001

Figure 7.3.12. Results of the General Linear Model in jamovi

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.

7.4: Working with Categorical Outcomes

The analyses I've presented so far are geared towards cases where the outcome variable is numeric. Dealing with a categorical outcome variable introduces additional complexity. In such situations, when both the outcome and predictor variables are categorical, we typically employ a χ^2 test. On the other hand, when the predictor is numeric and the outcome is categorical, we might opt for a logistic or multinomial logistic regression.

It's important to note that these analyses don't conform to the general linear model framework. Instead, they fall under a different category of models known as ***generalised linear models***.

The exciting part is, that regardless of the statistical method you're using – whether it's a t-test, ANOVA, regression, or other advanced techniques like mixed models, random forests, or generalised linear models – the process for fitting and visualising the model, as well as computing model estimates, remains consistent when you utilise flexplot:

1. Fit the model model = lm(y~x, data=data)
2. Visualise the model visualise(model)
3. Compute estimates for the model estimates(model)

This uniformity was intentional and designed for simplicity. Even as we delve into more complex statistical topics after the probability chapter, the process continues to follow the same steps:

1. Fit the model
2. Visualise the model with the visualise command.
3. Compute the estimates with the estimates command.

Beyond the probability chapter, we'll introduce a few additional steps, including fitting an alternative model, visualising both models and performing model comparisons. However, the core process remains consistent, regardless of the nature of the independent and dependent variables, whether they are numeric, categorical, or involve various groupings.

This streamlined approach greatly simplifies the analytical process.

Chapter attribution

This chapter contains material taken and adapted from *The Order of the Statistical Jedi* by Dustin Fife, used under a CC BY-SA 4.0 licence.

7.5: Introduction to Multivariate Statistical Modelling

Determining what constitutes a multivariate analysis can be a tricky question, and the answer can vary depending on who you ask. Technically, the term “multivariate” signifies the involvement of multiple variables, implying that any analysis with more than one variable could be considered a multivariate analysis.

However, I’ve noticed that people tend to use “multivariate” in one of two distinct ways:

1. When conducting an analysis involving multiple dependent variables. In statistical jargon, multivariate often pertains to analyses where researchers investigate multiple dependent variables. These scenarios call for the application of techniques like Multivariate Analysis of Variance (MANOVA), factor analysis, principal component analysis, structural equation modelling, and canonical correlations. Personally, I find many of these analytical methods somewhat outdated and not particularly useful. They often lack a clear theoretical basis, blur the lines between exploratory and confirmatory research, and can be challenging to interpret. This isn’t the focus of this chapter.
2. When performing an analysis that incorporates multiple independent variables. Most people, except for statisticians with a strong historical background, use “multivariate” to describe situations involving multiple independent variables. This is precisely what I mean when I refer to multivariate analysis in this textbook. Does this make me a “mutt-breed” statistician? Perhaps, but sometimes practicality outweighs the need for strict authenticity.

So, to clarify, this chapter (and those following it) deals with scenarios where we employ multiple predictor variables to model a single outcome variable. Hooray!

Now, let’s explore the reasons for using multivariate Generalised Linear Models (GLMs):

1. To study interaction effects: Occasionally, variables “interact,” meaning their impact depends on other variables. For instance, the level of annoyance (the outcome variable) I feel about attending department meetings (predictor variable #1) might depend on whether there’s food served (predictor variable #2). I might be more willing to attend meetings if they offer baklava and pizza, but not so much without these incentives.
2. To control for uninteresting factors: Suppose you know that people who are depressed tend to have poor social lives, but your primary focus is on studying depression’s unique influence on health, not social functioning. In this case, you’d want to “control” for social functioning, effectively isolating the effect of depression on health.
3. To improve predictions: In short, the more variables you include, the better your predictions become. Therefore, if you’re aiming to predict the next world wood-chopping champion, you can add more predictors, such as bicep circumference, years of experience, and beard length, to enhance the accuracy of your predictions.

Before we dive into a discussion about each of these reasons, let’s take a brief intermission to enjoy some illustrative visuals. Why not, right? Pictures have their charm. However, to avoid overwhelming my imaginary editor, I’ll integrate these captivating visuals into our data analysis process.

Chapter attribution

This chapter contains material taken and adapted from *The Order of the Statistical Jedi* by Dustin Fife, used under a CC BY-SA 4.0 licence.

CHAPTER OVERVIEW

Chapter 8: Putting it all Together

Learning Objectives

After reading this chapter, you should be able to:

- practice statistical modelling from start to finish.

In this chapter, we will bring together everything that we have learned, by applying our knowledge to a practical example. James and colleagues (2015) wondered if doing a visual task, like playing Tetris, while a memory is reconsolidating could interrupt the memory storage and make the intrusive memories happen less often. They argued that people who played Tetris after remembering the traumatic event would see a reduction in those intrusive memories. Just playing Tetris without remembering the trauma or remembering the trauma without playing Tetris wouldn't have the same effect. We will use this study to show how one would go about analysing an experimental dataset from start to finish.

8.1: Practical steps to Statistical Modelling

8.1: Practical steps to Statistical Modelling

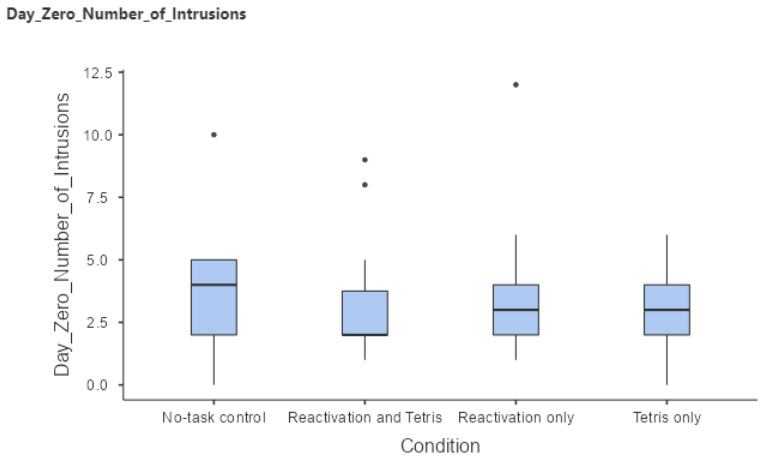


Figure 8.1.1. Box plot of a number of intrusive memories on Day 0 before the experimental task by the different conditions

The boxplot above shows that the intrusive thoughts are relatively similar for all conditions. Remember, we expect that all groups should have the same amount of bothersome memories during the first 24 hours since this is before any changes (Day 0). This is just to ensure that all groups started with a relatively similar baseline.

However, what we really want to test is the effect of the experimental manipulation. In particular, we want to examine whether there is a significant difference between the conditions on the number of memory intrusions in the seven days following the experimental task. We will use the variable named *Day_One_to_Seven_Number_of_Intrusions* and visualize them to see if there are any outliers:

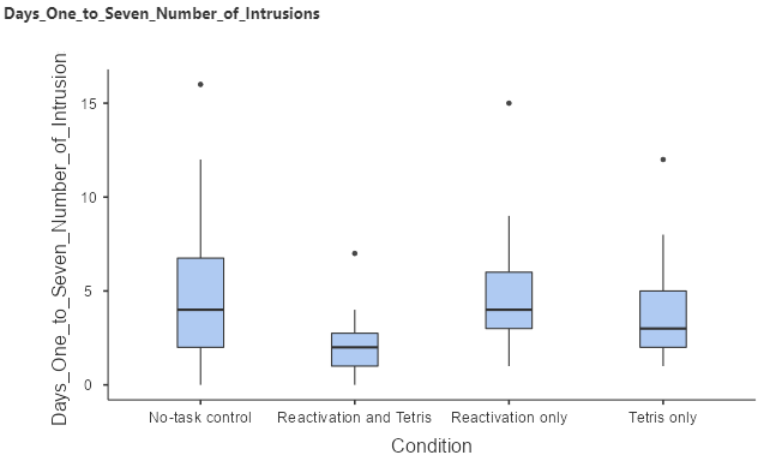


Figure 8.1.2. Box plot of a number of intrusive memories on Day 1 to 7 after the experimental task by the different conditions

Box plots are useful to see the shape of the distributions, as shown in Figure 8.1.2. Those data look fairly reasonable – there are a couple of outliers (indicated by the dots outside of the box plots), but they don't seem to be extreme. We can also see that the distributions seem to differ a bit in their variance, with the reactivation and Tetris showing somewhat less variability than the other groups, while the no-task control has the most variability. This means that any analyses that assume the variances are equal across groups might be inappropriate. Fortunately, the statistical model that we plan to use is fairly robust to this.

Step 4. Determine the Appropriate Model

There are several questions that we need to ask in order to determine the appropriate statistical model for our analysis.

Figure 8.1.3. Results of the General Linear Model (jamovi screenshot)

Note that the software automatically generated dummy variables that correspond to three of the four conditions, leaving the no-task control without a dummy variable. This means that the intercept represents the mean of the no-task control condition, and the other

three variables model the difference between the means for each of those conditions and the mean for the no-task control condition. No-task control condition was chosen as the unmodeled baseline variable simply because it is first in alphabetical order.

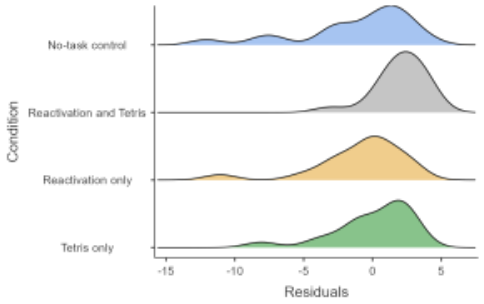


Figure 8.1.4. Distribution of residuals for each condition

Another important assumption of the statistical tests that we apply to linear models is that the residuals from the model are normally distributed. It is a common misconception that linear models require that the *data* are normally distributed, but this is not the case; the only requirement for the statistics to be correct is that the residual errors are normally distributed. The right panel of Figure 8.1.5 shows a Q-Q (quantile-quantile) plot, which plots the residuals against their expected values based on their quantiles in the normal distribution. If the residuals are normally distributed then the data points should fall along the dashed line – in this case, the plot doesn't look the best. What we want is for the residuals (denoted by the dots) to be tightly packed around a line (in other words, linear). However, given that this model is also relatively robust to violations of normality, we will go ahead and continue with our analysis.[3]

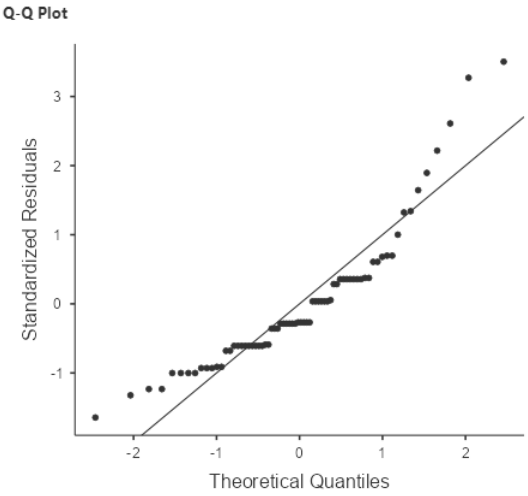


Figure 8.1.5. Q-Q plot of actual residual values against theoretical residual values

Figure 8.1.6. Fixed effects parameters estimates table for the dummy coded variables (jamovi screenshot)

From the table above, we can see that the frequency of intrusive memories for participants under the no-task control and reactivation-only conditions was significantly different from the reactivation task with the Tetris condition.

Post-Hoc Comparisons

For the following analysis, we will differ from the original paper to show you how you would conduct the analysis if they did not provide a specific hypothesis.

Because we are doing several comparisons, we also need to correct those comparisons, which is accomplished using a procedure known as the Tukey method, which can be requested by going into the Post Hoc Tests, putting the condition into the variable window and checking Tukey under correction.

Post Hoc Tests

Post Hoc Comparisons - Condition

Comparison							
Condition		Condition	Difference	SE	t	df	P_{tukey}
No-task control	-	Reactivation and Tetris	3.22	1.06	3.04	68.00	0.017
No-task control	-	Reactivation only	0.28	1.06	0.26	68.00	0.994
No-task control	-	Tetris only	1.22	1.06	1.15	68.00	0.657
Reactivation and Tetris	-	Reactivation only	−2.94	1.06	−2.78	68.00	0.034
Reactivation and Tetris	-	Tetris only	−2.00	1.06	−1.89	68.00	0.242
Reactivation only	-	Tetris only	0.94	1.06	0.89	68.00	0.809

Figure 8.1.7. Post-Hoc Tests between the different conditions (jamovi screenshot)

The column titled Ptukey in the rightmost column shows us which of the groups differ from one another, using a method that adjusts for the number of comparisons being performed. Anything below the p-value of .05 is significantly different from one another. This shows that the pairing of no-task control and reactivation and Tetris as well as reactivation and Tetris and reactivation only were the only pairs that significantly differ from one another.

What about Possible Confounds?

If we look more closely at the James et al. paper, we will see that they also collected data on attention paid to the film. Let's plot this data on a bar plot for each condition.

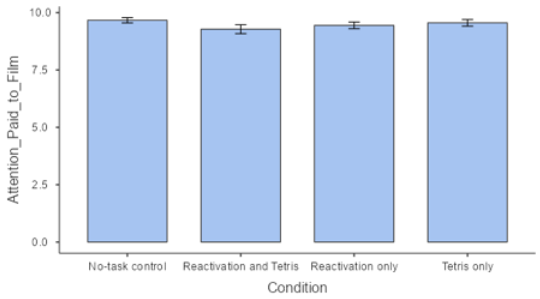


Figure 8.1.8. Barplot of attention given to the film per condition

Looking at the data it seems that the rates were consistent across the conditions. If the data is quite different across groups, then we may be concerned that these differences could have affected the results of the intrusive memory outcomes. In our case, this is not an issue. However, it is also good to check potential confounding variables that may be affecting your data.

Getting Help

Whenever one is analysing real data, it's useful to check your analysis plan with a trained statistician, as there are many potential problems that could arise in real data. In fact, it's best to speak to a statistician before you even start the project, as their advice regarding the design or implementation of the study could save you major headaches down the road. Most universities have statistical consulting offices that offer free assistance to members of the university community. Understanding the content of this book won't prevent you from needing their help at some point, but it will help you have a more informed conversation with them and better understand the advice that they offer.

1. James, E. L., Lau-Zhu, A., Tickle, H., Horsch, A., & Holmes, E. A. (2015). Playing the computer game Tetris prior to viewing traumatic film material and subsequent intrusive memories: Examining proactive interference. *Journal of Behavior Therapy and*

Experimental Psychiatry, 53, 25-33. https://doi.org/10.1016/j.jbtep.2015.11.004 ↵

2. This example came from OpenStatsLab. For more practical exercises such as this one, visit: https://sites.google.com/view/openstatslab/about ↵
3. Some may argue that these violations suggest that we should not fit the GLM in our data. This is fine – we can instead conduct a Generalised Linear Model if you are concerned about these violations. Another option is to conduct the non-parametric equivalent of ANOVA, which is the Kruskal-Wallis test. ↵

CHAPTER OVERVIEW

Chapter 9: Beyond Research and Statistics

Learning Objectives

After reading this chapter, you should be able to:

- have a critical lens on psychological research.

In this chapter, we will take a step back and critique “objectivity” and “value neutrality” that permeate the ideas in psychological research. We will also discuss their potential implications on research and practice when left unaddressed. Lastly, we will discuss the scientific reform movement and alternative frameworks of knowing.

9.1: Beyond Research and Statistics

9.1: Beyond Research and Statistics

> *All methodologies, even the most obvious ones, have their limits.*
>
> *– Paul Feyerabend in Against Method*[1]

Scientific objectivity is a key characteristic of different elements of science. It embodies the notion that the assertions, methodologies, and outcomes of science, and even the scientists themselves, should not be, or strive not to be, swayed by specific viewpoints, value assessments, communal prejudice, or personal interests, among other relevant factors. Objectivity is often seen as an ideal state in scientific investigation, a compelling reason for appreciating scientific knowledge, and the foundation of science's authority in society.

The field of psychology has earned its reputation by emulating the natural sciences. This perspective of psychology presumes that its theories and methods are impartial and devoid of values (the concept we refer to as ***value neutrality***) and that our investigations about the world are devoid of preconceived notions, vested interests, and subjective interpretations.

However, there are key debates revolving around the concepts of objectivity and value neutrality in research. Can we truly be devoid of preconceived notions and biases while conducting our scientific inquiry?

What is Objectivity and Value Neutrality?

Objectivity is an important concept in science. When we label something as objective, we are expressing its significance and our approval. Objectivity is not a binary concept; rather, it exists on a spectrum. Claims, methods, results, and scientists can exhibit varying degrees of objectivity, and, all else being equal, greater objectivity is generally considered preferable. The term "objective" often carries a distinct rhetorical weight. The widespread admiration for science and its authoritative position in public life largely derives from the perception that science is objective, or at least more so than other modes of inquiry. Therefore, a comprehensive understanding of scientific objectivity is essential for grasping the essence of science and its societal role.

There are different viewpoints on how we can conceptualise objectivity and in this chapter, we will briefly discuss objectivity as 1) faithfulness to the facts, 2) value-free or value-neutral and, 3) freedom from personal biases.[2]

Objectivity as Faithfulness to the Facts

Often, people attribute **objectivity as faithfulness to the facts.** The philosophical basis for this understanding of objectivity lies in the belief that there are factual elements existing "out there" in the world, and it is the responsibility of scientists to uncover, scrutinise, and organise these facts. The term "objective" is thus linked to success; if a statement is objective, it correctly describes some aspect of the world. According to this standpoint, the objectivity of science is determined by its proficiency in identifying and generalising facts, distancing itself from the standpoint of the individual scientist.

Many philosophers argue that the relationship between observation and theory is complex, with influences running both ways. Thomas S. Kuhn (1970) presented a lasting criticism in his work titled, *The Structure of Scientific Revolutions*.[3]

Kuhn's analysis is based on the idea that scientists approach research problems through the lens of a paradigm, encompassing relevant problems, axioms, methodological presuppositions, and techniques. He supported this with historical examples, highlighting that scientific progress occurs within a guiding paradigm that influences individual scientists and community standards in everyday science.

Can observations challenge such a paradigm and advocate for a different one? Kuhn famously emphasizes that observations are "theory-laden" (Hanson, 1958),[4] influenced by a body of theoretical assumptions that shape their perception and conceptualisation.

Objectivity as Value-Free or Neutral-Free

An alternative view of objectivity as faithfulness to the facts is the viewpoint that objectivity is **value-free or value-neutral**. If the aim of science is to generate empirical knowledge, and if disputes involving value judgments cannot be resolved through empirical methods, then values cannot have a place in science. However, is this possible? Let's look at a classic example that relates to what we have been learning about in this book – fitting a mathematical function to a dataset.

Fitting a mathematical function to a dataset involves making a choice for the researcher. They can opt for a complex function, which may complicate the relationship between variables but results in a more accurate fit to the data. Alternatively, they can propose a simpler relationship that is less accurate. Both simplicity and accuracy are crucial cognitive values, and balancing them requires careful consideration. However, philosophers of science often view the presence of values in this context as acceptable. Cognitive values, also known as “epistemic” or “constitutive” values, such as predictive accuracy, scope, unification, explanatory power, simplicity, and coherence with other accepted theories, are considered indicative of the truth of a theory. Consequently, they offer reasons for favouring one theory over another.

In most perspectives, the objectivity and authority of science are generally unaffected by cognitive values, only by non-cognitive or contextual values. These contextual values encompass moral, personal, social, political, and cultural aspects like pleasure, justice, equality, conservation of the natural environment, and diversity. Improper use of such values has historically led to severe consequences, as seen in instances where contextual values influenced scientific agendas with intolerant and oppressive outcomes. For example, during the Third Reich, certain branches of physics were condemned due to the Jewish background of their inventors, and in the Soviet Union, biologist Nikolai Vavilov faced harsh consequences for theories conflicting with Marxist-Leninist ideology. Both regimes sought to align science with political convictions, resulting in disastrous effects.

Less dramatic but perhaps more common are cases where research is biased towards the interests of sponsors, like tobacco companies, food manufacturers, and pharmaceutical firms (e.g., Reiss 2010).[5] This preference bias violates conventional research standards to achieve a specific result and is clearly harmful from an epistemic perspective. Particularly for critical issues such as drug approval or the consequences of human-induced global warming, it is desirable for research scientists to assess theories without being influenced by such considerations. This concept is encapsulated in the value-free ideal, which suggests that scientists should minimize the impact of contextual values on scientific reasoning, particularly in gathering evidence and assessing/accepting scientific theories.

To be **value-free**, scientific objectivity is marked by the absence of contextual values and freedom from cognitive biases. However, for value-freedom to be a reasonable ideal, it must be attainable to some degree. In other words, it must not be completely unattainable. Instead, some people call for **value-neutrality**. Value-neutrality asserts that scientists can, at least in principle, gather evidence and assess/accept theories without making contextual value judgments. Unlike the value-free ideal, the value-neutral thesis is not normative; it addresses whether scientists’ judgments can be, or could possibly be, free of contextual values.

Objectivity as Freedom from Personal Biases

According to this perspective, science is considered objective when personal biases are absent from scientific reasoning or can be eliminated through a social process. Common ways to achieve this objectivity include measurement and quantification. Measured and quantified values are verified against a standard, like stating the height of the Eiffel Tower in meters. This truth is relative to a standard unit and conventions about instrument use, making it independent of the person measuring.

Measurement provides some independence of perspective. For instance, yesterday’s weather in Durham, UK might be considered “really hot” by a typical North Eastern Brit and “very cold” by an average Mexican, yet both would agree it was 21°C. However, measurement doesn’t offer a completely neutral perspective or free us from presuppositions. Measurement instruments interact with the environment, so results are influenced by both the properties of the environment being measured and the instrument used, providing a perspectival view of the world.

However, measurements do not result in a completely unbiased view. Measurement instruments interact with the environment, offering a perspectival view (cf. Giere 2006)[6]. Interpreting measurement results is also crucial. For example, early thermometry, according to Hasok Chang (2004)[7], relied on a “principle of minimalist overdetermination” to find a reliable thermometer with minimal assumptions. However, even reliable procedures can be influenced by the purposes of the scientists involved, especially in the social sciences where normative assumptions, i.e., values, often play a role.

Julian Reiss (2008, 2013)[8][9] argues that economic indicators, like consumer price inflation and gross domestic product, are value-laden. For instance, consumer-price indices assume ethical positions regarding consumer preferences, and national income measures make value-laden assumptions about market exchange. Furthermore, beyond measuring and quantifying characteristics, we use statistics to describe relationships between quantities and make inferences in scientific work. We should know from this book that statistics is certainly vulnerable to personal biases.

Feyerabend's Arguments Against Rationality and Objectivity of the Scientific Method

In the 1970s, Paul Feyerabend became well-known for his criticisms of the scientific method and is considered an important science philosopher. Feyerabend challenged the rationality and objectivity of the scientific method. Feyerabend argued against the "tyranny" of rational methods, stating it hinders science from serving society. He valued diverse, even idiosyncratic perspectives, rejecting the idea that freedom from personal "bias" is beneficial.

Feyerabend's criticism of rational methods starts with the claim that strict rules like the value-free ideal stifle an open exchange of ideas and hinder scientific creativity. In his most famous work, *Against Method* (originally published in 1975), he explores the historical clash between the Catholic Church and Galileo, illustrating that groundbreaking scientific progress often involves violating traditional rules. Feyerabend's "Anything goes" dictum rejects the notion that rational methods can fully capture the irrational ways science deepens understanding.

He argues against an objective, value-free, and method-bound view of science, stating it limits our perspective, creativity, and humanity. Feyerabend sees traditional forms of inquiry, like Chinese medicine, on par with Western counterparts. He criticises appeals to "objective" standards as tools for bolstering Western intellectual authority.

Feyerabend contends that personal perspectives and biases can be beneficial for science. He suggests that scientific research should be accountable to society, advocating for democratic institutions and laymen's involvement in setting research agendas and ethical standards.

Feyerabend supports epistemic pluralism, accepting diverse approaches to knowledge acquisition. Instead of a narrow ideal of objectivity, he promotes a science that respects the diversity of values and traditions, harkening back to its role during the scientific revolution and the Enlightenment as a liberating force against oppression.

Alternative Form of Objectivity: Objectivity as a Feature of Scientific Communities and Their Practices

The following section argues an alternative form of objectivity – objectivity as a feature of scientific communities and their practices. This view of objectivity rejects the idea that objectivity is about correspondence between theories and the world or an individual's reasoning practices. Instead, they assess the objectivity of a collective of studies and the methods guiding scientific research. Three perspectives are discussed: reproducibility and the meta-analytic perspective; feminist and standpoint epistemology; and the incorporation of indigenous knowledge.

Reproducibility and the Meta-Analytic Perspective

In times of crises, such as the replication crisis, the collective perspective becomes crucial. Large-scale replication projects reveal the lack of trustworthiness in findings across various fields. Replicability has long been argued to provide evidence of freedom from biases and scientific artifacts and therefore establish the reliability of the result.

When replication failures in a discipline are notably significant (as we have seen in the discipline of Psychology), it may be inferred that the published literature lacks objectivity – at the very least, the discipline fails to instil confidence that its discoveries surpass mere artifacts of the researchers' endeavours. Conversely, when observed effects can be replicated in subsequent experiments, a form of objectivity is attained that extends beyond the concepts of freedom from personal bias, mechanical objectivity, and subject-independent measurement.

This is what Freese and Peterson (2018)[10] call **statistical objectivity**. It is rooted in the perspective that even the most meticulous and diligent researchers cannot achieve complete objectivity independently. The term "objectivity" instead pertains to a collection or population of studies, with **meta-analysis** (a formal method for aggregating the results from a range of studies) as the "apex of objectivity" (Freese & Peterson 2018). Specifically, combining studies from different researchers may offer evidence of systematic bias and questionable research practices in the published literature. The diagnostic function of meta-analysis in identifying deviations from objectivity is bolstered by statistical techniques such as the funnel plot and the p-curve (Simonsohn et al., 2014).[11] However, it is important to acknowledge that meta-analyses are still vulnerable to biases as authors may choose not to share details about their methods such as specification choices regarding the exact method for computing effect sizes, selection choices for weighting factors, not providing raw statistics used and the script used for their analyses (López-Nicolás et al., 2022).[12]

In addition to its epistemic aspect, research on statistical objectivity also carries an activist dimension: methodologists encourage researchers to publicly share essential parts of their research before commencing data analysis and to enhance transparency in their methods and data sources. For instance, it is hypothesized that the replicability (and hence objectivity) of science will improve by making all data accessible online, preregistering experiments, and adopting the registered reports model for journal articles (i.e., the

journal decides on publication before data collection based on the significance of the proposed research and the experimental design). The rationale is that transparency regarding the dataset and experimental design facilitates the replication of an experiment and the evaluation of its methodological quality. Furthermore, committing to a data analysis plan in advance is expected to reduce the occurrence of questionable research practices and attempts to tailor data to hypotheses rather than making accurate predictions.

Feminist Epistemology

Feminist perspectives challenge traditional notions of objectivity – there are various viewpoints on this matter but in this chapter, we will focus on **feminist epistemology**. Feminist epistemology explores how sex and gender impact scientific knowledge, often rejecting the value-free ideal. Specifically, feminist epistemology underscores the epistemic dangers arising from systematically excluding women from the scientific community and overlooking women as subjects of study. Notable instances include the disregard for the female orgasm in biology, the exclusive testing of medical drugs on male participants, the concentration on male specimens when examining the social behaviour of primates, and the explanation of human mating patterns through imaginary neolithic societies.

Frequently, though not always, feminist epistemologists move beyond highlighting what they see as androcentric bias and completely reject the value-free ideal, focusing on the social and moral responsibility of scientific inquiry. They aim to demonstrate that a science infused with values can still meet crucial criteria for being epistemically reliable and objective. A prominent example of such efforts is Longino's (1990)[13] contextual empiricism. She supports Popper's emphasis on "the objectivity of scientific statements lies in the fact that they can be inter-subjectively tested" (1934 [2002], p. 22)[14], but in contrast to Popper, she views scientific knowledge as fundamentally a social product. Therefore, our understanding of scientific objectivity must directly involve the social process that generates knowledge.

Indigenous Knowledge

> *In indigenous ways of knowing, we understand a thing only when we understand it with all four aspects of our being: mind, body, emotion, and spirit. I came to understand quite sharply when I began my training as a scientist that science privileges only one, possibly two, of those ways of knowing: mind and body.*
>
> *– Robin Wall Kimmerer in Braiding Sweetgrass: Indigenous Wisdom, Scientific Knowledge and the Teachings of Plants*

Recognising the importance of diverse knowledge systems, there is a growing acknowledgment of **Indigenous knowledge** as a valuable perspective in scientific discourse. Indigenous knowledge, rooted in the wisdom of Indigenous communities, offers unique insights into ecosystems, biodiversity, and sustainable practices. Incorporating Indigenous knowledge in scientific research is seen as a step toward a more inclusive and holistic understanding of the world. For instance, in indigenous knowledge, nature is commonly depicted as an intricate and interconnected system where each part relies on the others. Humanity is seen as an inherent component of nature, and some indigenous scholars even describe the human-nature relationship as symbiotic. Kimmerer, an Indigenous scholar and botanist, wrote in her book:

> *With a long, long history of cultural use, sweetgrass has apparently become dependent on humans to create 'disturbance' that stimulates its compensatory growth. Human participates in a symbiosis in which sweetgrass provides its fragrant blades to the people and people, by harvesting, create the conditions for sweetgrass to flourish. (164)*

From this text, Kimmerer (2013) notes that sweetgrass relies on human-created disturbance for its growth, establishing a symbiosis where people benefit from the fragrant blades while contributing to sweetgrass flourishing through harvesting. This quote embodies the contextuality that science (as explained above) wants to avoid. In other words, contextuality is the enemy of objectivity. Modern knowledge flourishes through abstract formulations and exists separately from people's lives. In contrast, Indigenous knowledge is deeply connected and harmonious with the lives of the people who created it.

Unlike modern knowledge, which claims superiority based on universal validity, local knowledge is constrained by space and time, and influenced by contextual and moral factors. Importantly, it cannot be detached from broader moral or normative purposes. To achieve universality and validity, knowledge must be dissociated from a larger epistemic framework that ties it to normative and social objectives. Context is local, anchoring technical knowledge to a specific social group in a particular setting at a specific time.

Conclusion

In conclusion, the exploration of scientific objectivity reveals its nuanced and multifaceted nature. Beyond the foundational emphasis on faithfulness to facts, the chapter critically examines the value-free ideal, challenges in achieving freedom from personal biases, and Paul Feyerabend's call for epistemic pluralism. Transitioning to alternative forms of objectivity, the collective perspectives of reproducibility, meta-analysis, feminist epistemology, and Indigenous knowledge underscore the need for a more inclusive and reflective scientific approach. The chapter advocates for ongoing dialogue and a reevaluation of traditional norms, recognising that diverse lenses enrich our understanding of the intricate interplay between knowledge, context, and societal values.

1. Feyerabend, P. K. (2010). *Against method* (4th ed.). Verso Books. ↵
2. Reiss, J., & Sprenger, J. (2020). Scientific objectivity. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy archive* (Winter 2020 ed.). https://plato.stanford.edu/archives/win2020/entries/scientific-objectivity/ ↵
3. Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). University of Chicago Press. ↵
4. Hanson, N. R. (1958). *Patterns of discovery: An inquiry into the conceptual foundations of science*. Cambridge University Press. ↵
5. Reiss, J. (2010). In favour of a Millian proposal to reform biomedical research. *Synthese, 177*(3), 427–447. https://doi.org/10.1007/s11229-010-9790-7 ↵
6. Giere, R. N. (2006). *Scientific perspectivism*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226292144.001.0001 ↵
7. Chang, H. (2004). *Inventing temperature: measurement and scientific progress.* Oxford University Press. doi.org.10.1093/0195171276.001.0001 ↵
8. Reiss, J. (2008). *Error in economics: The methodology of evidence-based economics.* Routledge. ↵
9. Reiss, J. (2013). *Philosophy of economics: A contemporary introduction.* Routledge. ↵
10. Freese, J., & Peterson, D. (2018). The emergence of statistical objectivity: Changing ideas of epistemic vice and virtue in science. *Sociological Theory, 36*(3), 289–313. doi.org/10.1177/0735275118794987 ↵
11. Simonsohn, U., Nelson, L. D., & Simmons, P. P. (2014). P-Curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547. doi.org/10.1037/a0033242 ↵
12. López-Nicolás, R., López-López, J. A., Rubio-Aparicio, M., & Sánchez-Meca, J. (2022). A meta-review of transparency and reproducibility-related reporting practices in published meta-analyses on clinical psychological interventions (2000-2020). *Behavior Research Methods, 54*(1), 334-349. https://doi.org/10.3758/s13428-021-01644-z ↵
13. Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press. ↵
14. Popper, K. (1934/1968). *The logic of scientific discovery* (2nd ed.). Harper and Row. (Originally published in 1934 as Logik der Forschung by Julius Springer) ↵

Index

D

dire

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: A Contemporary Approach to Research and Statistics in Psychology (Somoray)

Webpages: 67

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC 4.0: 85.1% (57 pages)
- Undeclared: 14.9% (10 pages)

By Page

- A Contemporary Approach to Research and Statistics in Psychology (Somoray) - *CC BY-NC 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - Acknowledgements - *CC BY-NC 4.0*
    - Acknowledgement of Country - *CC BY-NC 4.0*
    - About the Book - *CC BY-NC 4.0*
    - About the Author - *CC BY-NC 4.0*
    - Why do we Need Another Book About Research and Statistics? - *CC BY-NC 4.0*
  - Chapter 1: Research and Statistical Thinking in Everyday Life - *CC BY-NC 4.0*
    - 1.1: What can Statistics do for us? - *CC BY-NC 4.0*
    - 1.2: If I Apply Statistical Thinking all the Time then why do I find it Difficult? - *CC BY-NC 4.0*
    - 1.3: Big Ideas in Statistics - *CC BY-NC 4.0*
    - 1.4: Data is/are - *CC BY-NC 4.0*
  - Chapter 2: Working with jamovi - *CC BY-NC 4.0*
    - 2.1: Why jamovi? - *CC BY-NC 4.0*
    - 2.2: Getting Started with jamovi - *CC BY-NC 4.0*
    - 2.3: Analyses - *CC BY-NC 4.0*
    - 2.4: The Spreadsheet - *CC BY-NC 4.0*
    - 2.5: Loading Data in jamovi - *CC BY-NC 4.0*
    - 2.6: Installing add-on Modules into jamovi - *CC BY-NC 4.0*
  - Chapter 3: Brief Review of Research Methods - *CC BY-NC 4.0*
    - 3.1: How do we Measure Variables in Psychology? - *CC BY-NC 4.0*
    - 3.2: Introduction to Psychological Measurement - *CC BY-NC 4.0*
    - 3.3: What Makes a Good Measure? - *CC BY-NC 4.0*
    - 3.4: Some Complexities - *CC BY-NC 4.0*
    - 3.5: The Role of Variables - Predictors and Outcomes - *CC BY-NC 4.0*
    - 3.6: Research Design I- Experimental Designs - *CC BY-NC 4.0*
    - 3.7: Research Design II- Non-Experimental Designs - *CC BY-NC 4.0*
  - Chapter 4: The Replication Crisis - *CC BY-NC 4.0*
    - 4.1: How we Think Science Should Work - *CC BY-NC 4.0*
    - 4.2: Reasons for Non-Replication - *CC BY-NC 4.0*
    - 4.3: What can we do About it? - *CC BY-NC 4.0*
  - Chapter 5: Aggregation - *CC BY-NC 4.0*
    - 5.1: Why Summarise Data? - *CC BY-NC 4.0*
    - 5.2: Summarising Data Using Tables - *CC BY-NC 4.0*
    - 5.3: Summarising Data Using Graphs - *CC BY-NC 4.0*
    - 5.4: The Middle of the Data - *CC BY-NC 4.0*
    - 5.5: Variability - How Spread Out are the Values? - *CC BY-NC 4.0*
    - 5.6: Z Scores - *CC BY-NC 4.0*
  - Chapter 6: Modelling Variations - *CC BY-NC 4.0*
    - 6.1: A Simple Model - *CC BY-NC 4.0*
    - 6.2: Statistical Modelling Using a Single Number - *CC BY-NC 4.0*
    - 6.3: Sampling and Sampling Error - *CC BY-NC 4.0*
    - 6.4: The Central Limit Theorem - *CC BY-NC 4.0*
    - 6.5: Null Hypothesis Testing - *CC BY-NC 4.0*
    - 6.6: Quantifying Effects - *CC BY-NC 4.0*
  - Chapter 7: The General Linear Model - *CC BY-NC 4.0*
    - 7.1: General Linear Model - *CC BY-NC 4.0*
    - 7.2: Modelling Continuous Relationships - *CC BY-NC 4.0*
    - 7.3: Comparing Means - *CC BY-NC 4.0*
    - 7.4: Working with Categorical Outcomes - *CC BY-NC 4.0*
    - 7.5: Introduction to Multivariate Statistical Modelling - *CC BY-NC 4.0*

- Chapter 8: Putting it all Together - *CC BY-NC 4.0*
  - 8.1: Practical steps to Statistical Modelling - *CC BY-NC 4.0*
- Chapter 9: Beyond Research and Statistics - *CC BY-NC 4.0*
  - 9.1: Beyond Research and Statistics - *CC BY-NC 4.0*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - 20: Accessibility Statement - *CC BY-NC 4.0*
  - Chapter 11: Versioning History - *CC BY-NC 4.0*
  - Chapter 10: Review Statement - *CC BY-NC 4.0*
  - Detailed Licensing - *Undeclared*

20: Accessibility Statement

2

Accessibility Overview

A contemporary approach to research and statistics in psychology has been designed with accessibility in mind.

In addition to the web/HTML version, this book is available in several file formats, including PDF and EPUB (for eReaders).

Look for the “Download this book” drop-down menu to select the file type you want.

Accessibility Standards

The web version of this resource is designed to meet Web Content Accessibility Guidelines 2.0.

Let us Know if you are Having Problems Accessing this Book

We are always looking for how we can make our resources more accessible. If you are having problems accessing this resource, please contact us to let us know so we can fix the issue.

Please include the following information:

- the location of the problem by providing the book title, a web address or page description
- a description of the problem
- the computer, software, browser, and any assistive technology you are using that can help us diagnose and solve your issue
  - e.g., Windows 10, Google Chrome (Version 65.0.3325.181), NVDA screenreader.

You can contact us via:

- Web form: Submit a question to JCU Library

This statement was last updated on: 7 February 2023

Chapter 11: Versioning History

1

This page provides a record of changes made to this book. Each set of edits is acknowledged with a 0.1 increase in the version number. The exported/downloadable files for this book reflect the most recent version.

Version	Date	Change	Details
1.0	28 February 2023	Front matter and first four chapters published in accordance with OER Collective grant timeline requirements.	Chapter 1: What can statistics do for us? Chapter 2: If I apply statistical thinking all the time then why do I find it difficult? Chapter 3: Big ideas in statistics Chapter 4: Data is/are
2.0	15 August 2024	Five additional chapters published in August 2024.	Chapter 5: Aggregation Chapter 6: Modelling variations Chapter 7: The general linear model Chapter 8: Putting it all together Chapter 9: Beyond research and statistics

Chapter 10: Review Statement

2

A Contemporary Approach to Research and Statistics in Psychology was produced with support from the Open Educational Resources Collective initiative of the Council of Australian University Librarians.

This book has been peer-reviewed by a subject expert.

The review was structured around the needs of the intended audience of the book, and covered:

- subject matter
- accuracy
- presentation of diverse perspectives
- longevity of the text
- clarity of the writing
- structure
- consistency
- user experience.

The author wishes to thank Dr Janine Lurie, psychology lecturer at James Cook University, Australia, for taking the time to thoroughly review the book.

Detailed Licensing

Overview

Title: A Contemporary Approach to Research and Statistics in Psychology (Somoray)

Webpages: 67

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC 4.0: 85.1% (57 pages)
- Undeclared: 14.9% (10 pages)

By Page

- A Contemporary Approach to Research and Statistics in Psychology (Somoray) - *CC BY-NC 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - Acknowledgements - *CC BY-NC 4.0*
    - Acknowledgement of Country - *CC BY-NC 4.0*
    - About the Book - *CC BY-NC 4.0*
    - About the Author - *CC BY-NC 4.0*
    - Why do we Need Another Book About Research and Statistics? - *CC BY-NC 4.0*
  - Chapter 1: Research and Statistical Thinking in Everyday Life - *CC BY-NC 4.0*
    - 1.1: What can Statistics do for us? - *CC BY-NC 4.0*
    - 1.2: If I Apply Statistical Thinking all the Time then why do I find it Difficult? - *CC BY-NC 4.0*
    - 1.3: Big Ideas in Statistics - *CC BY-NC 4.0*
    - 1.4: Data is/are - *CC BY-NC 4.0*
  - Chapter 2: Working with jamovi - *CC BY-NC 4.0*
    - 2.1: Why jamovi? - *CC BY-NC 4.0*
    - 2.2: Getting Started with jamovi - *CC BY-NC 4.0*
    - 2.3: Analyses - *CC BY-NC 4.0*
    - 2.4: The Spreadsheet - *CC BY-NC 4.0*
    - 2.5: Loading Data in jamovi - *CC BY-NC 4.0*
    - 2.6: Installing add-on Modules into jamovi - *CC BY-NC 4.0*
  - Chapter 3: Brief Review of Research Methods - *CC BY-NC 4.0*
    - 3.1: How do we Measure Variables in Psychology? - *CC BY-NC 4.0*
    - 3.2: Introduction to Psychological Measurement - *CC BY-NC 4.0*
    - 3.3: What Makes a Good Measure? - *CC BY-NC 4.0*
    - 3.4: Some Complexities - *CC BY-NC 4.0*
    - 3.5: The Role of Variables - Predictors and Outcomes - *CC BY-NC 4.0*
    - 3.6: Research Design I- Experimental Designs - *CC BY-NC 4.0*
    - 3.7: Research Design II- Non-Experimental Designs - *CC BY-NC 4.0*
  - Chapter 4: The Replication Crisis - *CC BY-NC 4.0*
    - 4.1: How we Think Science Should Work - *CC BY-NC 4.0*
    - 4.2: Reasons for Non-Replication - *CC BY-NC 4.0*
    - 4.3: What can we do About it? - *CC BY-NC 4.0*
  - Chapter 5: Aggregation - *CC BY-NC 4.0*
    - 5.1: Why Summarise Data? - *CC BY-NC 4.0*
    - 5.2: Summarising Data Using Tables - *CC BY-NC 4.0*
    - 5.3: Summarising Data Using Graphs - *CC BY-NC 4.0*
    - 5.4: The Middle of the Data - *CC BY-NC 4.0*
    - 5.5: Variability - How Spread Out are the Values? - *CC BY-NC 4.0*
    - 5.6: Z Scores - *CC BY-NC 4.0*
  - Chapter 6: Modelling Variations - *CC BY-NC 4.0*
    - 6.1: A Simple Model - *CC BY-NC 4.0*
    - 6.2: Statistical Modelling Using a Single Number - *CC BY-NC 4.0*
    - 6.3: Sampling and Sampling Error - *CC BY-NC 4.0*
    - 6.4: The Central Limit Theorem - *CC BY-NC 4.0*
    - 6.5: Null Hypothesis Testing - *CC BY-NC 4.0*
    - 6.6: Quantifying Effects - *CC BY-NC 4.0*
  - Chapter 7: The General Linear Model - *CC BY-NC 4.0*
    - 7.1: General Linear Model - *CC BY-NC 4.0*
    - 7.2: Modelling Continuous Relationships - *CC BY-NC 4.0*
    - 7.3: Comparing Means - *CC BY-NC 4.0*
    - 7.4: Working with Categorical Outcomes - *CC BY-NC 4.0*
    - 7.5: Introduction to Multivariate Statistical Modelling - *CC BY-NC 4.0*

- Chapter 8: Putting it all Together - *CC BY-NC 4.0*
  - 8.1: Practical steps to Statistical Modelling - *CC BY-NC 4.0*
- Chapter 9: Beyond Research and Statistics - *CC BY-NC 4.0*
  - 9.1: Beyond Research and Statistics - *CC BY-NC 4.0*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - 20: Accessibility Statement - *CC BY-NC 4.0*
  - Chapter 11: Versioning History - *CC BY-NC 4.0*
  - Chapter 10: Review Statement - *CC BY-NC 4.0*
  - Detailed Licensing - *Undeclared*