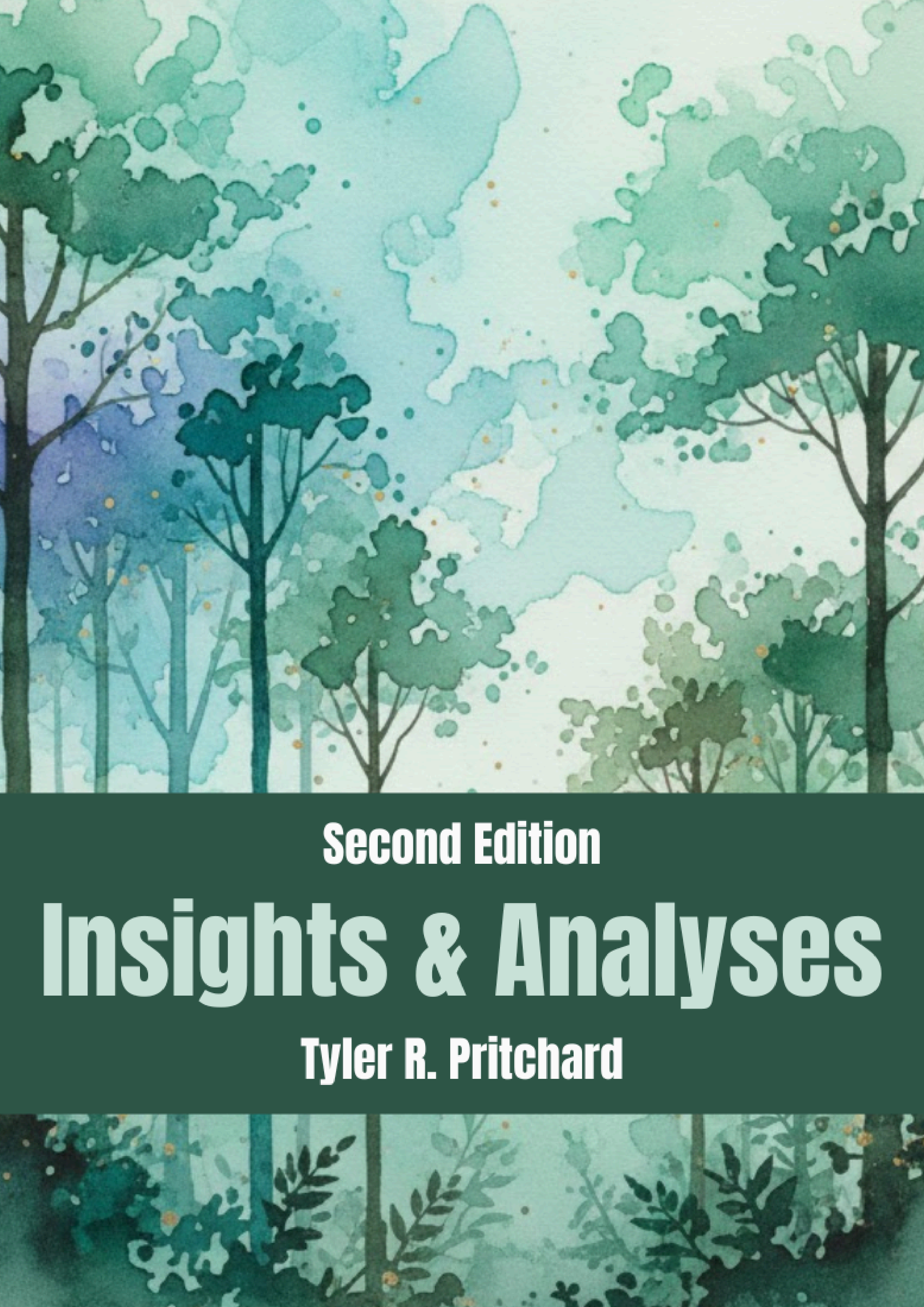


Second Edition

Insights & Analyses

Tyler R. Pritchard

For my favorite people: Kacey, Nathan, and Janine (in order of height, from shortest to tallest—in 2026).

Table of Contents

Insights and Analyses

This book is intended for educational and scholarly purposes. While every effort has been made to ensure accuracy, the author assumes no responsibility for errors or omissions, or for any consequences arising from the use of the information contained herein.

0.1 Additional Resources

Additional resources for this book may be found at: https://tylerpritchard.netlify.app/insightsandanalyses

Second edition
Printed online only

1 Introduction

About this Book

This book is a companion to Psychology 3950 (Research Methods and Statistics III) at Grenfell Campus, Memorial University. Additionally, this book may serve to support Grenfell Honours students when completing their relative analysis. While most example analysis are done using R, many of the key concepts are applicable to SPSS, JASP, or SAS. PSYC 3950 students can refer back to the three lab workbooks for additional support conducting these analyses in SPSS. Additional, non-R resources may be added at a later date. This book is a **working document**. I will make adjustments, add content, or fix errors as needed. This book is free to use. While you do not require permission, I ask you appropriately cite this document:

Pritchard, T. R. (2025). *Insights and analyses: A course companion (2nd edition)*. https://insightsandanalyses.netlify.app/

Additional Resources

Additional resources for this book may be found at: https://tylerpritchard/netlify/app/insightsandanalyses

Report errors, recommendations, or concerns to tpritchard@mun.ca.

About the Course

From the university calendar:

> PSYC 3950 Research Methods and Data Analysis in Psychology III will cover advanced research methods, including survey methods, and supporting statistical concepts and techniques. Designs will

> include single factor designs and multi-factor designs with both random and fixed factors. Supporting statistical concepts will include analysis of variance (ANOVA) from a linear model perspective, statistical power, and multiple regression, including model building. There may be a general introduction to multivariate statistical techniques. Ethical issues in research will be discussed in detail. Students will be required to design and carry out at least one research project from the design to the writeup stage, including an ethics review.
>
> — Grenfell Calendar

About Dr. Pritchard

Dr. Pritchard is currently a TTA at Memorial University (Grenfell Campus) and director of the Suicide and Health-Related Outcomes in Rural Environments (SHORE) Lab. Please visit the SHORE Lab website for more information.

So Tyler doesn't forget and for your own knowledge. The following color palette may look nice!

- Deep Pine Green (#2F4F4F)
- Sage Green (#8FAE94)
- Mint Green (#CFE3D4)
- Sky Blue (#A7C7E7)
- Soft Cream (#F4F1E9)

2 The Scientific Method

This chapter will provide a foundation for chapters to come. It introduces the scientific process, from theory to dissemination.

However, we must acknowledge that there are many ways to conduct science. Indeed, Kuhn (1962) detailed the existence of *scientific paradigms*–widely accepts models on how to 'science'. Without veering into philosophy of science, please recognize that this books depicts one such paradigm–albeit a popular one in psychology.

Think about it. How do we gain knowledge?

Imagine the following:

- Your grandfather says that 'kids these days spend too much time on the internet.'
- An X (formerly, Twitter) user posts that "teens are becoming dumber and sadder because of excess social media use."
- A peer-reviewed scientific article suggests potential small positive effects of social media use on depressive symptoms.

Who would you trust more? What impacts this decision? How does your own biases impact your decision? What would you suggest to someone asking for advice on monitoring their social media use?

We all have biases, flaws, and are prone to making errors. There were likely numerous times where you and a friend, partner, or family member disagreed about a situation–*you saw things differently*. As you have learned in other classes, our minds use heuristics to ease the cognitive load of having to process vast amounts of information. While it can speed things along for us and do a reasonably effective job in navigating our complex world, it can sometimes lead us astray.

As an example, consider someone with a cognitive bias to interpret their friends' actions in a negative way–as if often seen in those with depression. The person may text a friend asking to "hang out tonight?". The person could interpret the lack of response many ways such as 1) the friend is busy doing something else, 2) the friend's phone has died, 3) the 'friend' actually doesn't like them. In reality, the friend is at a family event and forgot their phone at home. Unfortunately for the person in this example, they conclude that option 3 is correct, without a doubt. In fact, this person always believes that they annoy their friends or are a burden to others. Importantly, these thoughts may not be grounded in 'reality'; the friends may perfectly enjoy the company of the individual in question. While this example may feel extreme, we all exhibit some cognitive biases to some degree in our lives.

As scientists, it's imperative to reflect on our biases and how that may impact how we collect and interpret information, and draw conclusions based on the information. Some researchers implement specific methods primarily to counter their potential biases (e.g., experimenter bias) (Strickland & Suben, 2012) and arrive at some 'truth'. Others draw on these biases in a reflexive manner, acknowledging that one can never completely separate from their own experiences and biases, and use this to strengthen their understanding of complex topics (e.g., reflexive thematic analysis) (Braun & Clarke, 2019). Although there are many 'ways' to do science, they typically use a systematic approach to generating an argument or idea, and then planning and implementing a method to test the verisimilitude (i.e., truthfulness) of the idea. For the purposes of this course, we will adhere to a commonly employed scientific method in psychology. Namely, null hypothesis significance testing. Very briefly, our research process will consist of:

1. **Generating hypotheses**
2. **Designing a study**
3. **Collecting data**
4. **Analyzing data**

5. **Disseminating results**

Think about it. Can we be bias free?

- What parts of psychology most interest you? Clinical, developmental, social, cognitive, etc.?
- Why do you want to study a specific topic in psychology?
- How does your background and, potential, biases impact this decision?
- How might these biases impact how we view a topic? For example, how does a view that 'all suicides can and should be prevented' impact how someone studies suicide?

2.1 Generating Hypotheses

Before we continue, its important to distinguish some common terms used in psychological research.

2.1.1 Theory

Quite broadly, we start with a theory. **A theory is a set of ideas or statements that explain how phenomena–things you observe in the world–work.** You encounter theories and apply them all the time. When you throw a ball to your friend, you do so with with an understanding that the Earth is bending space time and will cause the ball to accelerate downwards (i.e., gravity and theory of general relativity). When you go to the grocery store, you believe that you and the other customers are generally good people who have rights and responsibilities that they will abide by (i.e., social contract theory).

Formally, you have encounter many psychological theories in your students. For example, do these names ring a bell? Pavlov (theory of learning)? Eriksson (psychosocial theory)? Piaget (theory of cognitive development)? Freud (psychoanalytic theory)? Greenberg (terror management theory)? Deci and Ryan (self-determination theory)? I could go on (and on)! One commonality of the theories developed by these individuals is that they propose an explanation for some facet of psychological functioning. Do we need so many theories?

Definition

A theory is a set of ideas or statements that explain how phenomena–things you observe in the world–work.

There are myriad theories in psychology. In fact, theoretical pluralism is often viewed as a strength and necessity in our field. Human behavior is so complex that we need a diverse set of theories to explain different behaviors in different contexts. Some of the psychological theories seek to explain a specific component of psychology. For example, consider a theory seeking to explain suicide: the interpersonal psychological theory of suicide (Van Orden et al., 2010). Other theories may explain human behavior more broadly such as theories of learning.

Furthermore, sometimes theories contradict each other in their explanations. For example, one study may propose that suicide is caused solely by X, but another solely by Y. This marks the importance of research to elucidate which theory best reflect real world phenomenon.

Regardless, let's return to the interpersonal theory of suicide. The following set of statements may be derived from the theory:

Theory Example

Theory: Feeling like a burden is detrimental and indicates that the person perceives that they take more from relationships than they provide. Thinking about killing oneself (i.e., suicidal ideations) will occur when one believes that they are a burden to others. Importantly, the perception of being a burden is more important than whether that person is actually a burden.

2.1.2 Hypotheses

From a theory, we can derive a hypothesis (or the plural, hypotheses)– **a specific statement or prediction about something that will happen** When we throw the ball up in the air, we predict it will come down. When we are shopping for groceries, we predict that we won't be assaulted or robbed, and that all customers will pay for their goods.

Definition

A hypotheses is a specific statement that predicts something that will happen.

Going back to our theory of suicide, we can derive a hypothesis:

Hypothesis Example

Hypothesis: individuals who are induced with thoughts of burdensomeness (x) will have more thoughts of suicide (y) than those who are not induced thoughts of burdensomeness.

Thus:

$$x \rightarrow y$$

or more specifically:

$$Burdensomeness \rightarrow SuicidalThoughts$$

Why care about hypotheses? Hypotheses are foundational designing research studies.

2.1.3 What makes a quality hypothesis?

Not all hypotheses are equal. Although not an exhaustive list, there are several features of *higher quality* hypotheses. Typically, high quality hypothesis are testable, falsifiable, clear and precise, simple, and derived from relevant theory or observations. Let's explore these in detail.

2.1.3.0.1 1. Testable

A good hypothesis can be tested empirically. That is, you can design an experiment that can feasibly collect the data required to test it. Considering the example hypothesis above using the interpersonal theory of suicide, are we **able** to design a study to test it?

We *could* randomly assign people to one of two groups: one group is provided (false) information that indicate their family and friends believe that they are a burden. The other is provided more neutral information about their family and friends' beliefs about them. Then, we can measure people's suicidal ideation levels after receiving that information. What should we observe based on the above hypothesis?

> Think about it.
>
> What results would you expect based on the hypothesis?

While this study is *technically* possible, it would not be ethically possible. Ethically, we cannot *try* to psychologically or physically harm people–definitely not with the goal of inducing suicidal ideations. The research method we use to test a hypothesis must be **both practical and ethical**.

> Think about it.
>
> How could we design a study to ethically test the above hypothesis?

Instead, in our suicide example we may seek out two naturally occurring groups: one of individuals who *already* perceive themselves as a burden to others and another who do not. We could then measure and compare their suicidal ideations.

2.1.3.0.2 2. Falsifiability

A good hypothesis (and theory) able to be proven false; they are **falsifiable**. As a simple example, imagine we predict that x causes y. If we conduct a study and the results are that x occurs, but y does not, we have logical evidence that our hypothesis is false.

Indeed, Popper (1959) proposed that a scientist's goal should be to prove their theories false. Said another way, if we are advocates of truth through science then we should bravely try to prove ourselves wrong–a scary proposition. However, consider two possible research findings and their implications. The first is that our hypothesis *is not* supported by our research finding and that, as a result, our theories are wrong. In this case, we must discard or revise our theory. This is progress! The second is that our hypothesis *is* supported by our research finding and that, as a result, our theories may be correct. In this case, we must continue to try to prove our self wrong. The more we *can't* prove ourselves wrong, the more evidence that our ideas are correct. Indeed, it takes courage to try to prove yourself wrong. As humans we like to be right. However, many great thinkers and scientists have approached science with a 'prove myself wrong' mindset. When describing Einstein's theory, Popper (1959) writes about the risky nature of Einstein's experiments:

> Now the impressive thing about this case is the risk involved in a prediction of this kind. If observation shows that the predicted effect is definitely absent, then the theory is simply refuted.
>
> — [Popper]

2.1.3.0.3 3. Clarity and Precision

A good hypothesis has clear and precise definitions. Researchers must clearly *operationally define* the variables of interest. Researchers need to clearly and concretely explain how they conceptualize and measure a construct/variable of interest. For example, what do we mean by 'suicidal ideations' or 'believing they are a burden to others'? Researchers must **precisely** explain how they plan to measure the variables included in a hypothesis.

Per the American Psychological Association, an operational definition is:

> a description of something in terms of the operations (procedures, actions, or processes) by which it could be observed and measured. For example, the operational definition of anxiety could be in terms

> of a test score, withdrawal from a situation, or activation of the sympathetic nervous system. The process of creating an operational definition is known as operationalizationThis is a quote.
>
> — APA, 2018

Furthermore, the proposed relationships between variables should be clear. Consider the relationships between x and y. Do we expect a positive relationship, wherein higher scores on x are associated with higher scores on y? Or a negative a relationship, where higher scores on x are associated with lower scores on y?

One major goal of clarity and precision is to allow others to know exactly what constructs we are testing and the patterns we expect to observe in the data. A second major goal is that it allow others to attempt to replicate our findings.

2.1.3.0.4 4. Simplicity

A good hypothesis is as **simple as needed, but no more.** Thus, hypotheses should offer the simplest explanation possible that does not oversimplify the complexity of humans. Sometimes oversimplification (i.e., incorrectly proposing that a complex phenomenon is cause by something simple) occurs in psychological theories. sometimes this is because psychological theories are, by definition, psycho-centric and may neglect the biological, physiological, socio-cultural, and systemic factors that impact our cognition, affect, and behavior.

As a historical example, consider Watson (1913) who rejected the relevance of consciousness on understanding human learning and, rather, that behaviorism and conditioning were all that psychologists needed to consider. Per Watson (1913), "The time seems to have come when psychology must discard all reference to consciousness." We now know that myriad other factors beyond behaviorism impact learning and behavior.

2.1.3.0.5 5. Theory-derived

The last feature of a good hypothesis is that it is grounded in theory or empirical observations. It is useful to review the literature to gauge what is already known about the topic of interest and current state of knowledge. It's would likely not be a good feeling or use of your time to think of a great research idea and begin to plan a study, only to find out that the idea has already been studied with unfavorable results. Note: replications are important and will be discussed later. Regardless, do your background homework. Before coming up with a hypothesis, you must review the literature to determine what we know and what we don't–the latter is where you come in, ambitious scholar!

2.1.3.1 Statistical Hypotheses

Translating a word-based hypothesis into a statistical hypothesis is a critical step in psychological research, as it allows researchers to formally test their predictions using statistical analyses. A word-based hypothesis, or a **conceptual hypothesis**, is often a statement that predicts a relationship between two or more variables. For example, a researcher might hypothesize that "higher levels of social support reduce the risk of being diagnosed with depression one year later." Or, as another example, consider the following hypothesis from a former honours student from studying various forms of connectedness and their relationship to suicidal ideations. Each construct was operationally defined and the relationships was specified:

All forms of connectedness will be uniquely and negatively associated with suicidal ideations.

While both examples of conceptualized hypotheses are clear and meaningful in everyday language, they would need to be translated into a form that you can take your subsequent numbers and data, and get a numerical response–was the hypothesis supported?

Definition

A conceptual hypothesis is often a statement that predicts a relationship between two or more variables.

The are many ways to translate a conceptual hypothesis into a statistical hypothesis. Here is one common method. We will use another example of a researcher who is interested in the links between social support and depression.

The first step is to ensure you have an adequate conceptual hypothesis. Have you expressed it as testable prediction? For example, the researcher may write that "Students with higher social support will report lower levels of depression." The second step is to operationally define the variables. The researcher has two variables: social support and depression. The researcher may operationally define them as:

1. **Social support**
   - Definition: perceptions of help received from others. Types include tangible support (e.g., providing transportation), emotional support (e.g., listening empathically), and informational support (e.g., providing advice).
   - Measurement: Received Support Scale (Krause & Borawski-Clark, 1995)
2. **Depression**
   - Definition: an array of cognitive, emotional, and behavioral symptoms marked by negative emotions and lack of engagement and enjoyment in activities.
   - Measurement: Score on Beck Depression Inventory (BDI) (Beck et al., 1996)

The third step is to determine the appropriate statistical test based on the operationalization of the variables of interest and the proposed relationships/effects of the conceptual hypothesis. Many chapters in this book are dedicated to a specific analysis that can be used to test various hypotheses. In the example above, the researcher would propose using a correlation or regression analysis. The fourth and last step is to express your statistical hypothesis. In this book we will primarily express these as the null (H_0) and alternative (H_A) hypotheses (more to come). Our researcher may express:

$$H_0 : \rho = 0$$

AND

$$H_A : \rho \neq 0$$

Where ρ represents the population correlation between social support and depression.

This process of translating a conceptual hypothesis into a statistical hypothesis allows researchers to test their ideas statistical using the collected data. Additionally, it ensures that psychological research is grounded in empirical evidence and provides clear criteria for evaluating the validity and reliability of their findings. Statistical hypotheses will help with designing our study.

2.2 Designing a Study

After we have developed a suitable hypothesis, we can begin to plan out our study. Essentially, we want to develop the methods to test the hypotheses.

Researchers outline a proposed research plan that typically includes details about the 1) participants, 2) measures , 3) and procedure of a study. The **Method** section, which usually depicts these three key areas of a study, is essential for explaining how the study was conducted and ensuring it can be replicated.

It typically begins with a **Participants** subsection. Here, a description of the participants, including relevant demographic details such as age, gender, and ethnicity. The sample size is justified, typically using some form of a power analysis (see a later chapter). Additionally, any inclusion and exclusion criteria used to select participants are detailed (e.g., "we excluded individuals with a diagnosed mental disorder because…"). The participants section also describes how participants were recruited (e.g., through advertisements, schools, or online platforms) and whether any compensation was provided.

The following is a participants section from a former honours student:

Sample Size Determination

A power analysis was conducted using previously established effect sizes. Specifically, the lower limit of confidence intervals of existing effect sizes marking the association between pain and suicidal ideations, hopelessness and suicidal ideations, and connectedness and suicidal *ideations were used. When confidence intervals were not presented, halving the effect size was used to account for publication bias. The resulting analysis indicated that a total of 75 participants were needed to achieve a power of .80.*

Participants

A total of 100 individuals clicked the survey link. However, data from 42 of these individuals were excluded from the study because they failed to complete the survey in full, resulting in missing data. Thus, the current study included a total of 58 participants who were members of the general public. Participants were required to be at least 19 years of age or a university/college student (i.e., mature minor), and fluent in English. All participants who did not meet these requirements were excluded from the study. Participants' ages ranged from 18 to 53 years (M = 24, SD = 9.76). Five (8%) participants were men, 44 (76%) were women, nine were non-binary or preferred to not answer (16%). Forty-three participants were White (74%), 17 were Indigenous (29%), and six (9%) were individuals who identified as other minorities or preferred to not answer. Due to these low frequencies and to ensure anonymity, race and ethnicity demographic information for these individuals are not reported.

Next is the **Materials** section. This is sometimes referred to as the **Measures** section. Here the materials and measures used in the study are outlined. This includes any tools or equipment used such as computers or specialized software. Additionally, descriptions of questionnaires, surveys, or psychological tests used to collect data are included here, along with details about their psychometric properties (e.g., reliability and validity). If the study involves specific stimuli (e.g., images, sounds, or videos), these are described as well. For example, if the study included pre-recorded videos showing someone in a fake therapy session, they would be described in detail. Additionally, measures can be provided as supplementary material (i.e., people reading your work can easily access the materials, unless there are some copyright or ethical concerns).

The following is a measures section from a former honours student (note that the appendices were in her thesis and are not in this book):

Measures

Demographics. *A questionnaire was constructed to collect demographic information. Participants were asked standard demographic questions including age, race/ethnicity, and gender identity. See Appendix A for a full list of demographic questions.*

Suicidal Ideation. *The Depressive Symptom Index Suicidality Subscale (DSI-SS; Metalsky & Joiner, 1997) was used to measure suicidal ideation. The DSI-SS is free to use without asking for permission. The DSI-SS is a 4-item self-report questionnaire used to measure the frequency and intensity of suicidal ideations within the past two weeks. Items were rated on a scale ranging from 0 to 3 with higher scores indicating greater intensity of suicidal ideation. For example, participants were asked about the intensity of their suicidal ideations on a scale ranging from 0 (I am not having impulses to kill myself) to 3 (in all situations I have impulses to kill myself). Joiner et al. (2002) suggested the DSI-SS was a valid and reliable measure of suicidal ideations with a Cronbach's alpha coefficient of 0.90. The reliability of the current data is acceptable (Cronbach's alpha = 0.92). See Appendix B for the DSI-SS.*

Connectedness. *The Watts Connectedness Scale (WCS; Watts et al., 2022) was used to measure various types of connectedness. Permission to use the WCS for this study was obtained (see Appendix C). The WCS is a 19-item self-report questionnaire used to measure connectedness with three subtypes: connectedness to self, connectedness to others, and connectedness to the world, within the past two weeks. Items were rated on a visual analogue scale ranging from 0 (not at all) to 100 (entirely). Watts et al. (2022) suggested the WCS was a valid and reliable measure of connectedness. The reliability of the current data is acceptable (Cronbach's alpha = 0.89). See Appendix D for the WCS.*

Pain. *The Psychache Scale (Holden et al., 2001) was used to measure pain. Permission to use the Psychache Scale for this study was obtained (see Appendix E). The Psychache Scale is a 13-item self-report questionnaire used to measure psychological pain referred to as psychache. Items were rated on a 5-point Likert-scale ranging from either 1 (never) to 5 (always) or from 1 (strongly disagree) to 5 (strongly agree). Holden et* *al. (2001) suggested that the Psychache Scale was a valid and reliable measure of psychological pain. The reliability of the current data is acceptable (Cronbach's alpha = 0.93). See Appendix F for the Psychache Scale.*

Hopelessness. *The General Hopelessness Scale (GHS; Drinkwater et al., 2023) was used to measure hopelessness. Permission to use the GHS for this study was obtained (see Appendix G). The GHS is a 22-item self-report questionnaire used to measure hopelessness. Items were rated on a 7-point Likert-scale ranging from 1 (strongly disagree) to 7 (strongly agree). Drinkwater et al. (2023) suggested that the GHS was a valid and reliable measure of hopelessness. The reliability of the current data is acceptable (Cronbach's alpha = 0.78). See Appendix H for the GHS scale.*

Next is the **Procedure** section. Here, the complete procedure of the study is detailed, which offers a step-by-step account of what participants did during the study. This includes the instructions they received, the tasks they performed, and any experimental conditions they were assigned to (e.g., control or experimental groups). The procedure also details how participants were assigned to these conditions (e.g., random assignment) and the duration of each session. If deception was used, a description of the debriefing process is included.

The following is a procedure section from a former honours student:

Procedure *An online questionnaire was constructed using Qualtrics to measure pain, hopelessness, various types of connectedness, and suicidal ideations. The questionnaire was a self-report measure that included questions derived from a variety of scales. Participants were recruited using posters displayed around Grenfell Campus, as well as an online advertisement that was shared on [STUDENT'S] social media pages, the Grenfell Psychology Participant pool website, and the psychology major/minors page on Brightspace. Posters included information about how the study will examine the relationship between suicidal ideations and various types of connectedness. Participation in the study was voluntary and anonymous. A QR code and link to access the online questionnaire was also included in the poster.*

Once the study QR code was scanned or link was accessed, an online informed consent form was provided to participants to be completed prior to the distribution of the questionnaire. This form indicated who the researchers are, the purpose of the study, what tasks are required, how long it will take, potential risks and benefits, anonymity and confidentiality of participants and data, the right to withdrawal, and researcher contact information. After the online informed consent form was reviewed, participants provided consent to participate by clicking a button that stated, 'I consent to participate in this study.' Participants who did not consent could close their web browser. After clicking this button, participants were directed to a set of online questionnaires measuring pain, hopelessness, connectedness, and suicidal ideations. Respective instructions were provided at the start of each questionnaire.

Upon completion of the questionnaires, participants were directed to an end of study form that indicated how to receive 0.5% course credit for participating psychology courses or how to be entered to win a $20 gift card from their choice of Walmart, Amazon, or Tim Hortons. Specifically, interested participants were directed to an independent Qualtrics survey requesting their email, which served as an entry in the draw. Thus, participants' data is not linked to their name or email addresses. The end of study form also included mental health resources for Grenfell students (i.e., Counselling and Psychological Services contact information) and the public (i.e., Bridge the GAPP, the Warm Line, and NL Mental Health Crisis Line contact information). The study was approved by the Health Research Ethics Authority (2024.215).

While these are the core aspects of the method section, it is common to include other sections.

The **analytic plan** section may describe in detail the proposed statistical analyses of the research. Part of others being able to replicate your findings means they must have a sound understanding of your exact analytic plan. For example, how (if you did) did you remove outliers? Why did you choose a specific level of statistical significance? What type of regression and variable entry method did you use? These details help you justify and others understand your analytic choices.

Unfortunately, these is reticence among psychologists to share these details out of fear they have done something wrong or incorrectly. Indeed, a study by Houtkoop et al. (2018) asked psychology researchers about barriers to data sharing. Their results suggested that concluded that 38% of researchers feared someone would discover errors in their analyses or data, and 77% (!!) feared that others might conduct a different type of analysis that would expose their own conclusions as invalid. My recommendation is to be open and honest about your research and, thus, willing to share the information.

Additionally, it is typically required to present the ethical approval of your study. This section can briefly explain whether the study received approval from an ethics committee or institutional review board (IRB) (sometimes called a Research Ethics Board [REB]).

2.3 Collecting Data

The next step, after ethical approval, the goal is to carry forward your study. You will collect data in accordance with your REB-approved design until you have reached your *a priori* sample size–the sample size you determined that you would need to have adequate statistical power.

2.4 Analyzing Data

An integral part of research is conducting the appropriate statistical analyses. In essence, we have an hypotheses (i.e., idea/prediction) about how the data should fit together (e.g., x and y are correlated; $x \leftrightarrow y$). Analyses allow us to model the data (i.e., force some structure to it) to determine how well it fits with our hypotheses. The main goal of this e-text is to outline some commonly employed statistical analyses used in psychological research. As such, the chapters that follow may explain and, through examples, complete statistical analyses.

2.5 Disseminating Results

What do you do after you've conducted your analyses and came to a conclusion whether your hypotheses were supported? Importantly, we should communicate to other researchers and the general public exactly what we did, what we found, and what are the practical applications/ meanings in an honest and transparent way. This can be through public forums, academic journals, or as registered reports. Ideally, we can accumulate enough evidence to support our theories and, ultimately, how accurately we explain psychological phenomenon.

I strongly recommend reading Chambers (2019) to help uncover some of the darker sides of our current publication structure.

2.6 Conclusion

This chapter introduced and explained the key considerations of a scientific method, which seeks to take a theory that explains phenomenon and test it empirically. This broad method is applicable across a range of disciplines and specific research areas. Psychology is no exception. The focus will be on statistical analyses–those used to test specific hypotheses–related to the PSYC 2925, 2950, and 3950 courses at Grenfell Campus, Memorial University of Newfoundland.

2.7 Practice Questions

1. Identify a theory in psychology you would be interested in testing.
2. Derive a hypothesis from this theory.
3. Design a hypothetical study to test the hypotheses.

- Who are the participants?
- What materials do you need?
- What procedure would you follow?

    - Ensure people reading your study design could attempt to directly replicate your results.

4. What parties would be interested in knowing the results of the study?
5. How would you communicate your results?

3 Variables

> Two behaviorists meet on the street. One asks the other, "You're fine. How am I feeling today?
>
> — Unkown

A **variable** is a characteristic or attribute that can vary or take on different values. These values can be measured, observed, or manipulated in a study. For example, we can measure anxiety symptoms, observe aggressive behaviors on the playground, or manipulate the type of interventions a individual gets. As researchers we are particularly interested in variables because they are directly referred to in our hypotheses. We can use variables to examine relationships or make comparisons between them, and draw conclusions about the phenomena we are are studying.

Consider the following hypotheses (see Section 2 for a refresher):

1. Children who are more **anxious** are more likely to engage in **aggressive behaviors**.
2. Children enrolled in a **novel intervention** will have less **anxiety symptoms** when compared to those in the **standard/typical intervention**.

Based on these hypotheses, we have three key constructs to measure: anxiety, aggressive behaviors, and intervention type. While there are myriad ways to measure these things, the following is table provides an example that can helps us turn constructs into some quantitative/numerical representation:

Variable	Possible Measurement Approach	Value Range	Example
Depression	Beck Depression inventory	0-63	Johnny scored 24
Aggressive behaviors	Numbers of times a child pushes, punches, kicks, spits, or throws objects at someone during an individual recess period	0 - infinity	Sally engaged in 4 aggressive behaviors during Monday's recess period
Learning Intervention	Clinician randomly assigns kids to each group	0 - Typical/old intervention; 1 - Novel/new intervention	Karrie gets randomly assigned to receive the typical intervention (coded as 0 in data)

Testing

Importantly, different variables have different characteristic. For example, if we ask 100 people their ages, we can easily calculated their average age using the arithmetic mean. However, if we ask them their favorite color, we can't calculate the arithmetic mean. Or related to the example above, we could calculate the number of aggressive behaviors a child exhibits per recess period, which would give us a number (e.g., 4). Or, we could observe and classify children into a category of 'aggressive' or 'not aggressive.' Each will have pros and cons. Regardless, the way you measure a construct will limit the ways in which you can analyse the data. We will review four major classifications of variables–often remembered using the acronym **NOIR**.

3.1 NOIR

Nominal, ordinal, interval, and ratio (NOIR) are four levels of measurement that describe the nature of the values that a variable can take. These levels of measurement are hierarchical, with each level including all the characteristics of the levels below it. For example, a ratio variables carries all the characteristic of nominal, ordinal, and interval (and then some!). Here's a brief explanation of each:

3.1.1 Nominal Level

Nominal variables involve categories without any inherent order or ranking. Examples include gender, where categories are mutually exclusive, yet there is no inherent order or ranking among them. In nominal measurement, the focus is on classifying items into discrete categories. Typically nominal variables are analysed using frequencies. We can determine *how many* individuals endorse or identify within a specific, mutually exclusive category. However, they also used in more complex analyses such as chi-square. To visualize nominal variables, we may use bar graphs. For example, we ask 100 Grenfell students "what is your favorite color?" and get the following data.

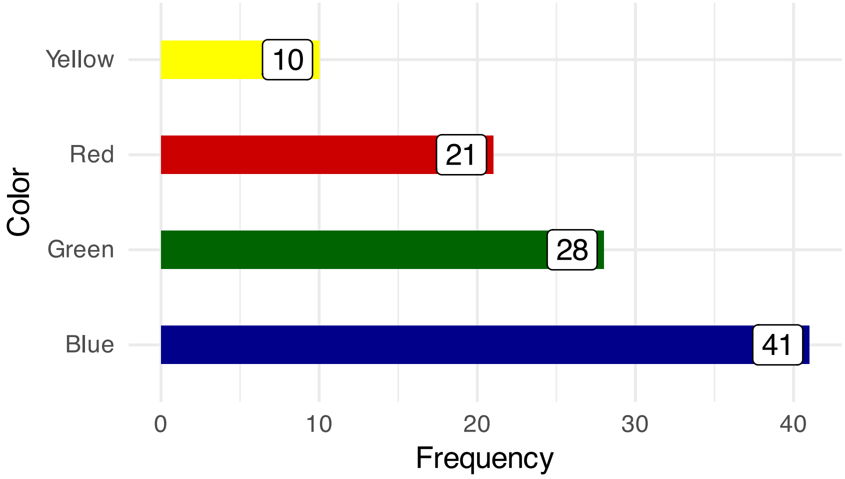


Figure 1: Students' favourite colors.

As an example, consider the following: individuals may be randomly assigned to one of two groups. One group receives a drug and the other a placebo [a nominal variable]. Sometimes researchers use labels to represent one's value on a nominal variable. Others may use a number to represent the value, with a "0" used to represent one group and "1" used to represent another (see table below). Note: using a number to represent a nominal variable does not create some numerical meaning to the variable. Also, the number choice is arbitrary–0 and 1 could easily be 44 and −166684. Continuing with the example, researchers may then determine the impact of the drug versus the placebo on the severity of psycho-pathological symptoms [not a nominal variable].

Person	Intervention Label	Intervention Number
1	Placebo	0
2	Drug	1
3	Placebo	0
...	...	...
N	Drug	1

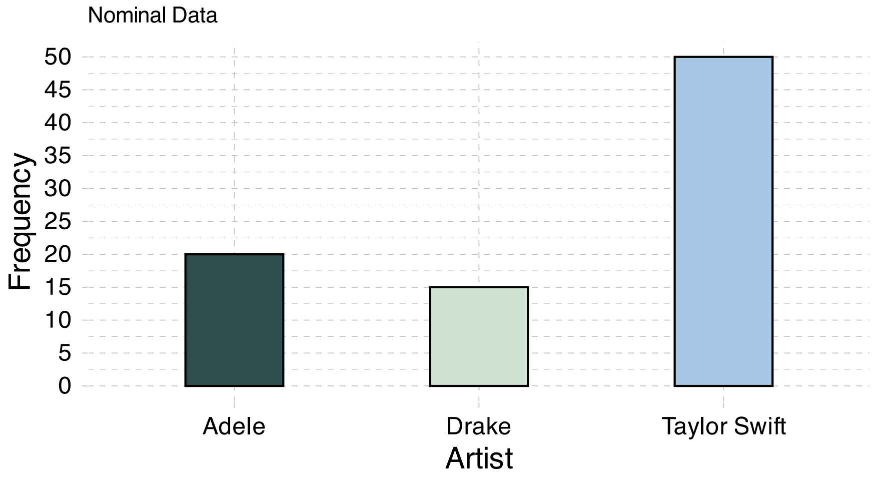
As another example, perhaps we measure Grenfell students' favorite musician. We collect data from a sample of 75 students. We can calculate frequencies of this nominal variable. We may represent the data the following table, which has both text labels (e.g., "Taylor Swift"), and numerical, "dummy-coded" labels. Here, a score of 0 indicates that Taylor is the favorite artist, 1 indicates Adele, and 2 indicates Drake. I have placed a "?" in one cell.

Individual/Participant	Artist Label	Artist Number
1	Taylor	0
2	Adele	1
3	Taylor Swift	0
4	Drake	2
...	...	...
75	?	1

Think about it

Based on the above table, who is person 75′s favorite artist?

We can also represent the final data as a graph. For example, a bar graph can be used to show the frequencies of the groups:



There is no inherent order here.

3.1.2 Ordinal Level

Ordinal variables possess a meaningful order or ranking of categories. However, the intervals between categories are not consistent or meaningful. That is the relative ranking is meaningful (e.g., category x comes before category y, which comes before category z). However, the differences between these categories are not uniform (the difference between category x and y is not necessarily the same as the difference between category y and z).

For example, consider a typical Likert-style scale (pronounced LICK-ert and named after its developer: American psychologist Rensis Likert). The difference between **strongly agree** and **agree** is not necessarily the same difference between **agree** and **neither agree not disagree**, regardless of the numbers you may assign to them. Or, as another example, consider educational levels of employees at Grenfell (e.g., high school diploma, bachelor's degree, master's degree, PhD). While you might be able to rank them, the differences between the categories is not equal for each level (i.e., the educational difference between a BA and an MA is not necessarily the same difference between a MA and a PhD).

The figure below shows both nominal and ordinal data. There is no inherent order for artists. You could impose some sort of order, such as alphabetical, but it is likely unrelated to the research question of interest.

However, students typically progress sequentially in their education (i.e., their is an order): first comes bachelor's, second masters, third PhD. Importantly, both can be represented as bar graphs.

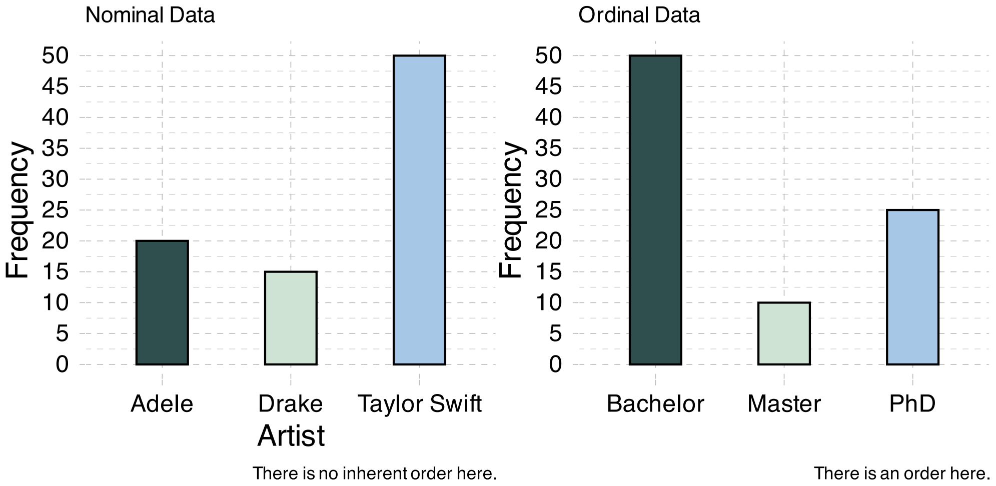


Figure 2: Nominal and ordinal data.

3.1.3 Interval Level

Interval variables maintain a meaningful order, and there are consistent intervals between values. However, these variables lack a true zero point–**zero does not represent the absence of the measured quantity**. Examples include temperature measured in Celsius or Fahrenheit; when it's zero degrees out, it does not mean there is *no* temperature. Also, 20 degrees Celsius isn't twice as much temperature as 10 degrees. Another example would be IQ scores: an IQ score of zero does not exist. Furthermore, an IQ of 120 isn't *"twice as smart"* as someone with an IQ of 60.

In interval measurement, researchers focus on both the order and the equal intervals between values. The difference between n values is equal for each ordered pair. Consider four ordered value:

$$a, b, c, d$$

Interval values have the property such that the difference between a and b is the same as the difference between b and c, which is the same as the difference between c and d:

$$a - b = b - c = c - d$$

3.1.4 Ratio Level

Ratio variables exhibit a meaningful order, consistent intervals between values, and a true zero point. In this level of measurement, a score of zero represents the absence of the measured quantity. Examples include height, weight, income, and age. Someone 120cm tall is twice as tall as someone who is 60cm tall. Someone who is 50 is twice as old as someone who is 25. The quantity of time since they were born is approximately twice as big. Ratio measurement allows for meaningful ratios and absolute distinctions between values.

The following table may be helpful, adapted from Nunnally & Bernstein (1994), who adapted it from Stevens (1951):

Scale	Operations	Transformations	Statistics	Examples
Nominal	Equal/not equal	So many	Frequency; mode	Gender; political party; employment status
Ordinal	Greater/less than	Monotonically increasing	Median; percentiles	SES (low, middle, high); Likert-style items
Interval	Equality of intervals	General linear	Arithmetic mean; variance	Temperature
Ratio	Equality of ratios	Multiplicative	Geometric mean	Height; weight

In addition, we can describe variables in terms of their function in our models. Note that these are independent of which type within NOIR that they are. We will focus on experimental and regression models.

3.2 Experimental Variables

An experiment has a typical structure, where the researcher **manipulates** one variables and observes another. Participants are typically assigned to the manipulated variable through a process called **random assignment**.

In random assignment, each participant has an equal probability of being assigned either value of the manipulated variable. Thus, there will be no systematic differences between the group of participant who received one value of the manipulated variable or the either.

Definition

Random assignment is a research method used to place participants into different experimental groups (such as a treatment group or a control group) using a random process like flipping a coin or using a random number generator.

Participants are then observed/measures on a second variable. Any differences in the observed variable are then attributed to the initial manipulation. There are two main types of variables in experimental psychological research.

3.2.1 Independent Variable (IV)

Independent variables (IVs) are variables that are manipulated or controlled by the researcher. It is the variable that is hypothesized to cause a change in the dependent variable. For example, in an experiment investigating the effects of a new teaching method on student performance, a researcher may design two teaching methods. Students are randomly assigned to one of the two conditions. The teaching method would be the IV.

The IV is often broken into two major groupings/conditions: the control group/condition and the experimental group/condition. A control group is the baseline that receives no treatment or a placebo, while the experimental group gets the actual intervention being tested. If the researcher has successfully randomly assignment participants to a IV condition, then both groups should be relatively similar on all other variables (e.g., age, gender, income), allowing researchers to isolate the effect of the IV and compare results to see if the treatment caused a change in the dependent variable.

Think about it

While the IV is often broken into a control and experimental condition, it could be more than two conditions. For example, a study by The MTA Cooperative Group (1999) –one of the longest and comprehensive treatment studies for ADHD–had four groups: 1) medication management; 2) intensive behavioral treatment; 3) the two combined; or 4) standard community care (treatments by community providers).

Importantly, *experimental* variables are mutually exclusive from which type of NOIR variable it is. An independent variable could, theoretically, be nominal, ordinal, interval, or ratio.

3.2.2 Dependent Variable (DV)

Dependent variables are variables that are measured or observed without some form of manipulation. Typically, in the context of experimental research, dependent variables are believed to differ based on or because of the independent variable. It *depends* on the independent variable.

Consider an example wherein researchers want to know if the different teaching methods lead to different student outcomes (e..g, better grades). The researcher could manipulate which students receive which interventions (the IV) and measure the student outcomes (the DV).

Importantly, and again, this is mutually exclusive from our NOIR variables. A dependent variable could, theoretically, be nominal, ordinal, interval, or ratio.

Another type of experiment is call a quasi experiment. This is a study that evaluates an intervention's impact but *lacks random assignment* to treatment or control groups. It resembles a true experiments but using existing groups–like different classrooms or communities–for comparison instead. For example, consider our experiment about different teaching interventions. Perhaps the researcher cannot randomly assign kids to receive one method or the other and, instead, must select a school that has a specifically trained educator to implement the teaching method. Here, the kids are not randomly assigned: there may be systematic differences about the classroom that receives the new versus old method.

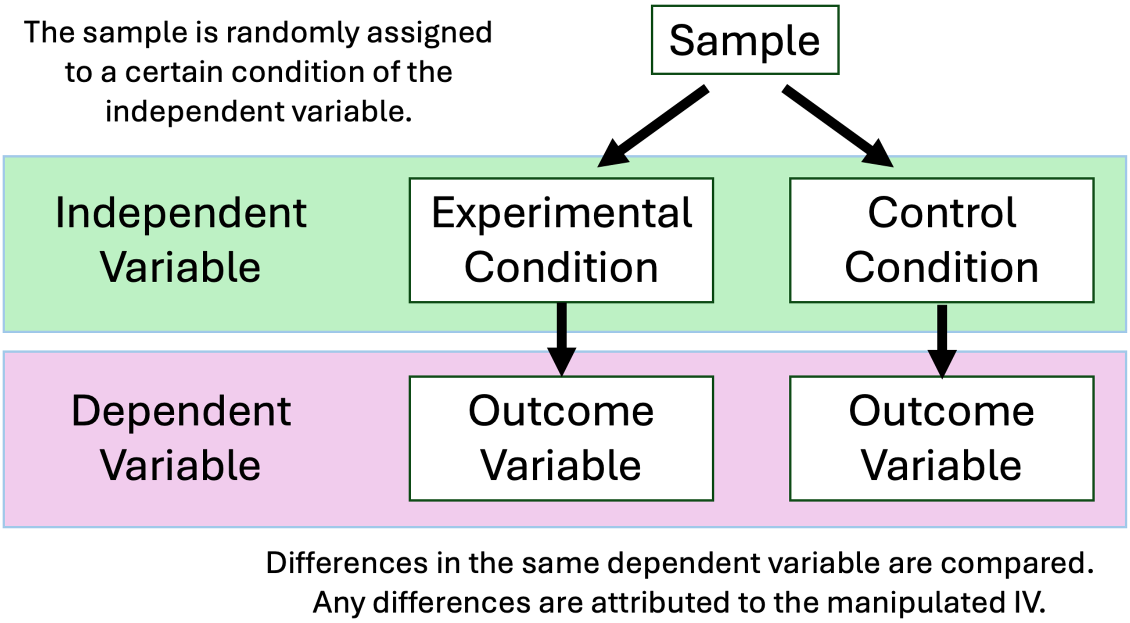

The sample is randomly assigned to a certain condition of the independent variable.
Sample
Independent Variable
Experimental Condition
Control Condition
Dependent Variable
Outcome Variable
Outcome Variable
Differences in the same dependent variable are compared. Any differences are attributed to the manipulated IV.

Figure 3: The experimental method.

3.3 Other Considerations

Researchers also consider and control for extraneous variables, which are variables that are not the focus of the study but could potentially influence the results. Controlling for these variables helps ensure that any observed effects or associations can be attributed to the manipulation of the IV.

3.3.1 Extraneous Variables

Extraneous variables are any variables other than the IV that may influence the results of an experiment. These variables are unwanted or unplanned factors that can introduce variability into the study, making it difficult to determine the true effect of the IV on the DV.

For example, if a researcher is investigating the effect of a new teaching method on student performance, extraneous variables could include the students' prior knowledge, motivation, or even the time of day the experiment is conducted.

3.3.2 Confounding Variables

Confounding variables are a specific type of extraneous variable that systematically varies with the IV and has a causal relationship with the DV. In other words, confounding variables can lead to a false interpretation of the relationship between the IV and DV. The researcher may believe that the IV had an impact on the DV, but it was actually a confounding variable.

Confounding variables can obscure the true effects of the independent variable, making it challenging to attribute changes in the DV solely to the manipulated IV.

For example, consider a study that examines birth order and likelihood of having autism. The researcher may conclude that the youngest sibling (quasi-experimental IV; researcher cannot manipulate birth order) is more likely to develop autism (DV). However, it may be that younger siblings' mothers and father are older when during prenatal development: both being risk factors for developing autism (Wu et al., 2017). Thus, maternal and paternal age are better explanations for the development of autism compared to birth order, which is related to maternal and paternal age.

3.4 Regression Variables

Variables in regression and correlation studies are not typically called independent and dependent variables. First, regression typically uses a **predictor variable** or set of predictor variable that are used to make predictions about another variable. Predictor variables are sometimes referred to as independent variables, analogous to an independent variable in experiments, but in correlational studies, it's about relationships and not necessarily causation. Last, the criterion variable is the outcome or behavior that is being predicted or explained. It is sometimes referred to as the outcome variables or the dependent variable.

For example, by knowing x, how well could I predict someone's score on y? If our hypothesis and subsequent model is accurate to the real world phenomenon, the predictor variables will do a good job. However, rarely can we perfectly predict the criterion variable, so there will be some error in our predictions. The typical structure is depicted as (a familiar sight, I hope):

$$y_i = x_i + e_i$$

Or, as another example, by knowing someone's degree of social disconnect, how well can I predict the severity of their current suicidal ideations?

$$ideations_i = disconnect_i + e_i$$

3.5 Conclusion

Variables in psychological research are key elements that researchers manipulate, measure, and/or analyze to gain a better understanding of psychological phenomena. Your theory and subsequent hypothesis will determine your variables of interest. How you operationalize your variables determines how you should measure them. How you measure them determines what your resulting data will be. Your data will determine the types of analyses you can do. The types of analyses you do determine the conclusion you can draw. Thus, it is imperative to effectively plan your research Methods to ensure that you can answer your research question and hypotheses.

3.6 Practice Questions

Identify the type of the following variables (NOIR):

1. The order of finishing for the participants in a race.
2. The numerical value representing the income level of individuals in a particular household.
3. Temperature difference between two consecutive days.
4. The preferred mode of transportation chosen by respondents.
5. Number of hours a student spends studying for the exam.
6. Participant gender,
7. Customer satisfaction levels on a scale from 1 to 5.
8. What are the IV and DV in the following experiment?:

A study investigates the impact of sleep duration on memory retention in college students. Participants are randomly assigned to either a group with regular sleep patterns (7-8 hours per night) or a group with restricted sleep (4-5 hours per night). Memory performance is assessed through a standardized memory test administered the following day.

9. Identify some confounding variables for the previous study.

3.7 Answers

1. Ordinal
2. Ratio
3. Interval
4. Nominal
5. Ratio
6. Nominal
7. Ordinal
8. IV = sleep; DV = standardized memory test

9. Stress level; individual variability in required sleep (i.e., what if someone in the restricted group only needs 5 hours); caffeine use prior to testing; sleep quality (e.g., total REM sleep for the night); etc.

4 Types of Statistics

This chapter will introduce two classifications of statistics: descriptive and inferential. While both are commonly used in psychology, they each have different use cases. By the end of this chapter, you will have a sound understanding of both.

4.1 Descriptive Statistics

Descriptive statistics involve the use of numerical and graphical methods to *summarize and present data* in a meaningful way. Descriptive statistics focus on describing and summarizing the main features of a variable or data-set. **The primary goal is to simplify large amounts of data**. This can help researchers, clinicians, and the public understand it. Some commonly used descriptive statistics include measures of central tendency and variability.

4.1.1 Central Tendency

Central tendency has to do with the typical or average score for a variable of interest. There are three main ways to calculate the central tendency of a variable: the mean, median, and mode. Prior to explaining each, let's imagine we measure the age of five university students' and get the following data:

Person	Age
1	20
2	19

Person	Age
3	23
4	22
5	19

4.1.1.1 Mean (Average)

The mean is one way we can understand the 'average' score of participants. The mean is the sum of all values divided by the number of observations:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

For our hypothetical data above:

$$\bar{x} = \frac{20 + 19 + 23 + 22 + 19}{5} = 20.6$$

There are several benefits of using the mean. The first benefit is that the mean takes into account every data point in the data-set, making it sensitive to changes in any value. For example, if the first person measured was 25 years old, and not 20, the new mean would be 21.6–an increase of 1. The second benefit is that the mean has convenient mathematical properties, making it suitable for various statistical analyses.

The major downside to using the mean is that it is highly sensitive to extreme values (outliers). Having a few outliers can drastically skew the result. Let's add one extreme value to our hypothetical data: a person of age 68. Our new mean would be:

$$\bar{x} = \frac{20 + 19 + 23 + 22 + 19 + 68}{6} = 28.5$$

You can see how drastic the change is (20.6 compared to 28.5). Or, as another example, imagine that you are interested in determining the average household income in Corner Brook, NL. You sample 10 houses and get the following data:

Household	Income
1	48000
2	52000
3	25000
4	101000
5	23000
6	55000
7	45000
8	100000
9	23000
10	950000

Notice that the 10th household you sample has a substantially higher income ($950,000). Indeed, when including household 10, the mean is 142200 Indeed, the mean is higher than all but one of our household incomes.

4.1.1.2 Median

The median is the middle value of a variable in a data-set when the scores are ordered numerically from least to greatest. It represents the 50th percentile of the data, meaning half of the values fall below it and half above. If the data-set has an even number of observations then there will be two middle values and the median will be calculated by taking the mean of the two middle values.

The main benefit of the median is that it is not influenced by extreme values, making it a robust measure of central tendency, especially for skewed distributions or those with outliers.

The main drawback is that it only considers the order of values, ignoring the actual numerical differences between them, which means it doesn't capture information about the magnitude of differences in scores in the data-set.

When calculating the median by hand, it's helpful to first order the values from least to greatest For example, let's order our 5 individuals, who's ages we collected:

Person	Age
2	19
5	19
1	20
4	22
3	23

Here, with an odd number of value, calculating the median is simple. It's the middle value (i.e., the 3rd of 5). The median is 20.

However, suppose we sampled another person, who's age was 68. What would the new median be?

Answer

The new median would be 21.

Notice how the median was less influenced by the 68-year-old outlier than the mean. The median changed by 1 when we included the outlier, while the mean changed by 7.9.

4.1.1.3 Mode

The mode is the most frequently occurring value of a variable in a data-set.

There are benefits and downsides to using the mode. It is particularly useful for nominal data, where there is no inherent order to the values, and it is also easy to understand and calculate. However, a data-set may have no mode. If each values occurs the same amount of times (e.g., each value only occurs once), one should use an alternative form of central tenancy. Furthermore, is that it is possible to have multiple modes. In this case, the distribution is described as multimodal (e.g., consider the scores on a variable of $4, 4, 6, 8, 8$; there are two modes). Last, the mode may not provide an accurate representation of the center of a distribution, especially for variables with skewed range of values.

Using our age example, what would be the mode?

Person	Age
1	20
2	19
3	23
4	22
5	19

Answer

The mode is 19.

4.1.2 Variability

Variability refers to how spread out or dispersed the values of a variable in a data-set are. It gives an indication of how much the data points differ from each other and from the central tendency (e.g., the mean or median). High variability means the data points are spread widely apart, while low variability indicates they are clustered closely together.

The following shows the variability in three variables:

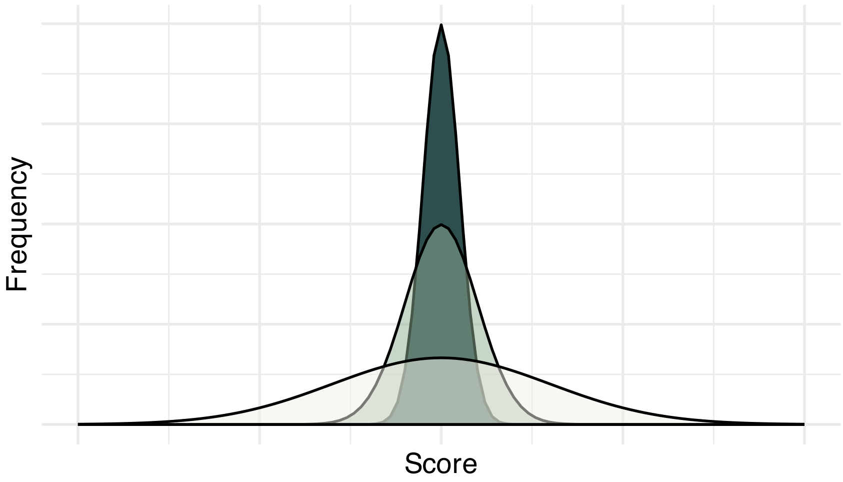


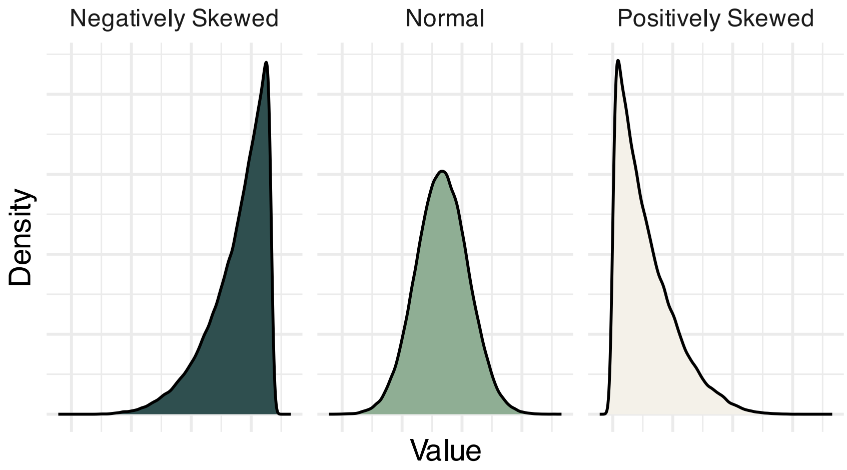
Figure 4: Three distributions with difference variability.

In the above, you can tell that the light green distribution is more dispersed than the medium green distribution, which is more dispersed than the dark green distribution. The dark green group is more closely grouped towards the center and mean of their group, whereas the others are more spread out. Said another way, the average distance between any score and their group's mean is much smaller for the dark green group compared to the medium green, which is smaller than the light green group.

4.1.2.1 Range

The range represents the difference between the minimum and maximum values of a variable in a data-set. It is a quick way to determine the variability in scores and allows us to identify outliers or other extreme scores.

However, range tells nothing about the type of distribution of the data you are describing. A range cannot tell us if data are skewed. For example, all three of following distributions of data have the same range:



For now, though, let's return to our age data:

Person	Age
1	20
2	19
3	23
4	22
5	19

For our hypothetical age data above, the min value is 19 and the max value is 23. Thus, the range is:

$$Range = 23 - 19 = 4$$

4.1.2.2 Variance

Variance quantifies the spread of data points in a data-set. It is the average squared differences from the mean. It indicates how much individual data points deviate from the mean, with a higher variance reflecting greater variability.

Variance is widely used to assess variability, identify trends, and predict outcomes. For instance, in psychological research, variance helps measure individual differences or treatment effects across groups.

A benefit of using variance is that it considers all data points in the data. A drawback is that variance is sensitive to outliers, which can distort the results. This is particularly problematic for small data sets. A second drawback is that, since variance is expressed in squared units, it can be difficult to interpret in practical terms. To address this, the square root of variance, known as the standard deviation, is often used for better interpretability.

Variance is calculated as:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Where:

- σ^2 is variance.
- $\sum$ is the sum.
- n is the number of data points in the data-set.
- x_i represents each individual data point.
- $\bar{x}$ is the mean of the data-set.

The above is the population variance. The sample variance, which you will likely use in your psychological research is:

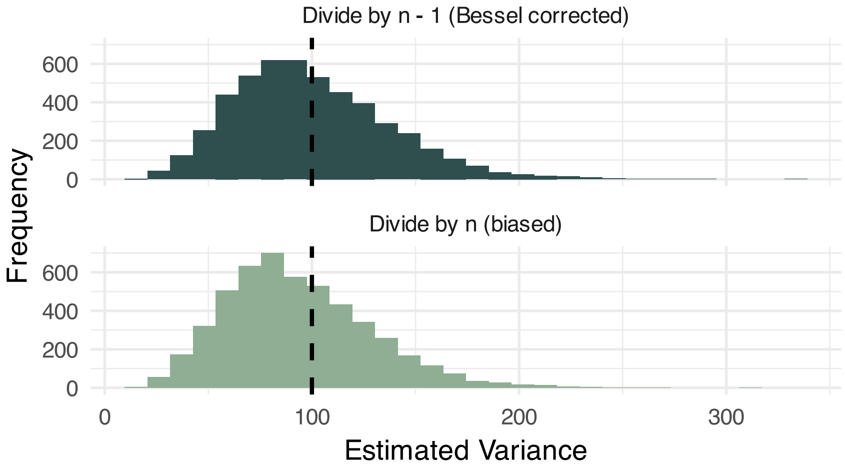
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

You may notice that the sample version uses $n-1$ instead of n. This is known as Bessel's correction.

💡 Bessel's Correction

Bessel's correction adjusts the denominator because sample means are calculated from the data, sample deviations tend to be smaller. We typically don't *know* the true mean and, thus, have to estimate it using our sample. When we do this we underestimate the true spread of the data. Dividing by a slightly smaller number (n-1) makes the variance estimate slightly larger, correcting for this bias.

The following figures show the estimate of variance in a sample when using and not using Bessel's correction. The vertical black line is the *true* population variance. Notice that the distribution is more closely centered around the true variance when using Bessel's correction.



For our data age data, the variance is:

$$s^2 = \frac{(20-20.6)^2 + (19-20.6)^2 + (23-20.6)^2 + (22-20.6)^2 + (19-20.6)^2}{5-1}$$

$$= \frac{13.2}{4} = 3.3$$

4.1.2.3 Standard Deviation

Standard deviation is used to quantify the amount of variation or dispersion in a variable of a data-set. It is the square root of variance, making it a more interpretable measure–it is expressed in the same units as the data. Standard deviation is useful to help researchers understand how spread out scores are around the mean, such as in assessing the variability of test scores or behavioral responses. A low standard deviation indicates that data points are close to the mean, while a high standard deviation shows that scores have more variability.

The key benefit of standard deviation is that it's intuitive and interpretable, especially compared to variance. It gives a clear sense of how much data points deviate from the average. A drawback is that, like variance, standard deviation is sensitive to outliers, which can inflate its value (i.e., make it higher). A second concern is that may not fully capture variability in skewed data-sets or those with non-normal distributions.

The formula for population and sample standard deviation can easily be calculated as the square root of population and sample variance, respectively. Or:

Population:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

and using Bessel's correction for samples:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

For our data:

$$s = \sqrt{\frac{(20-20.6)^2 + (19-20.6)^2 + (23-20.6)^2 + (22-20.6)^2 + (19-20.6)^2}{5-1}} =$$

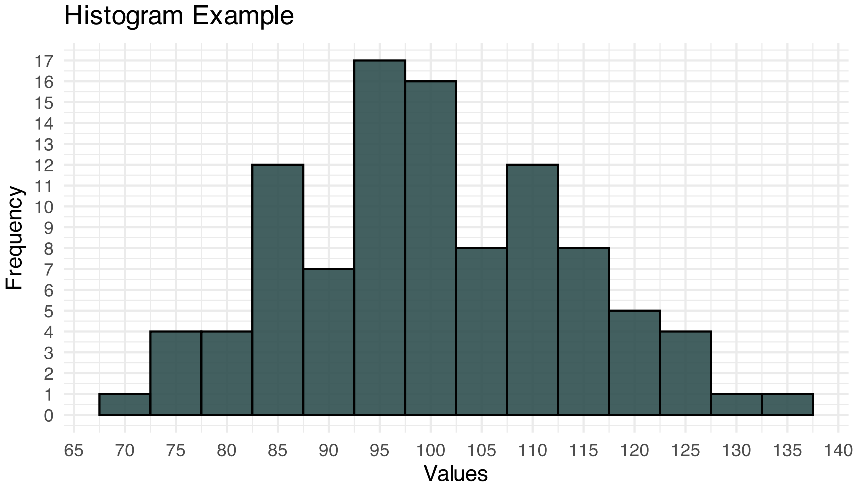
$$\sqrt{\frac{13.2}{4}} = \sqrt{3.3} = 1.82$$

The utility of means, variance, and standard deviation cannot be overstated (note: covariance is also imperative, but covered in a later chapter). As you will learn, they are the foundation of many of our statistical analyses. For now, let's consider some commonly used graphs and figures to represent data.

4.2 Graphs and Figures

4.2.1 Histograms

A histogram is often used to visually represent of the distribution of one (1) non-categorical variable (not a 'N' in NOIR) in a data-set. It shows the frequency of different ranges of values. Imagine we want to plot the distribution of IQ scores of 100 Grenfell students.



To read a histogram, start by examining the x-axis, which represents the range of values in the data-set. The data is divided into intervals or **bins** along the x-axis, and the y-axis displays the **frequency** or count of observations within each bin.

The height of each bar corresponds to the number of data points falling within that bin. A key aspect is the width of the bins, as it influences the visual interpretation. A narrower bin width can reveal finer details in the distribution, while a broader bin width may smooth out fluctuations. As an example, consider the following three histograms, all with the same data, but differing bin widths. The frequency of each bin is represented above the respective bin:

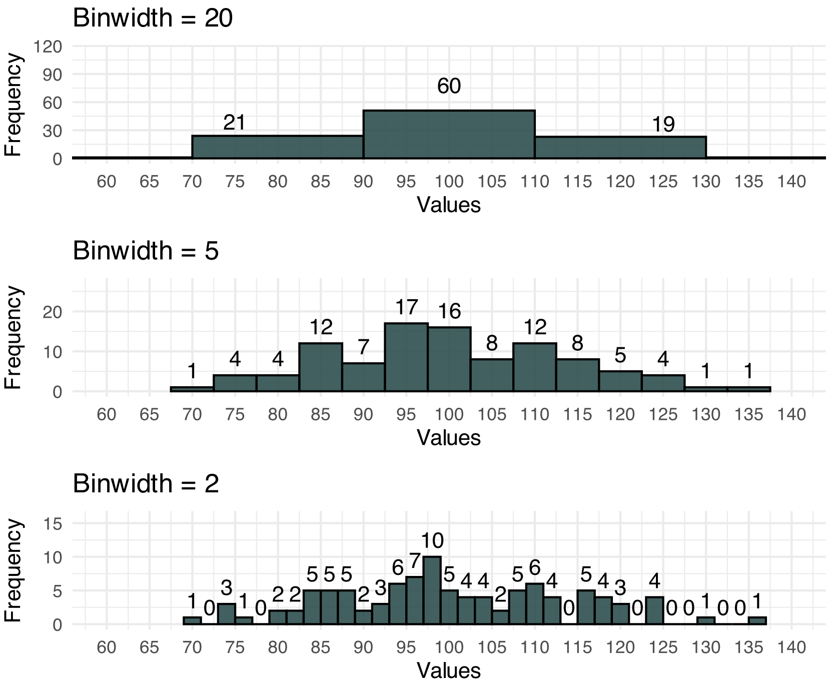
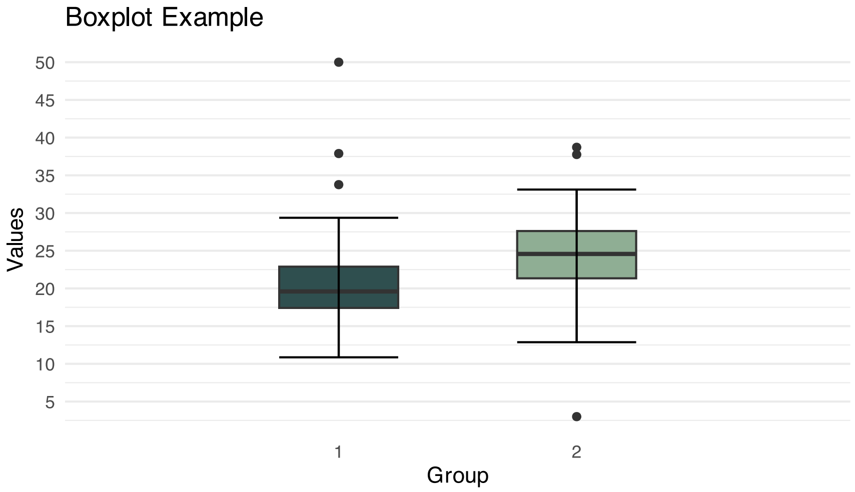


Figure 5: Impact of binwidth on histogram.

Additionally, the overall shape of the histogram indicates the data's central tendency, spread, and symmetry. Peaks and valleys highlight regions of higher or lower frequency, and the tails provide information about outliers or other extreme values.

4.2.2 Box plots

Box plots are graphical summaries of the distribution of a variable in a data-set, including the median, quartiles, and potential outliers. The following box plot represents data from two groups. The left box represents one group, the right box the the other.



4.2.2.1 Interpreting a Boxplot

First, the box–in our figure there are two colored boxes–represents the middle 50% of the data, known as the interquartile range (IQR). The lower (Q1) and upper (Q3) edges of the box correspond to the 25th and 75th percentiles, respectively.

Second, the line inside the box represents the median. You have already read about what the median is, its benefits and downsides.

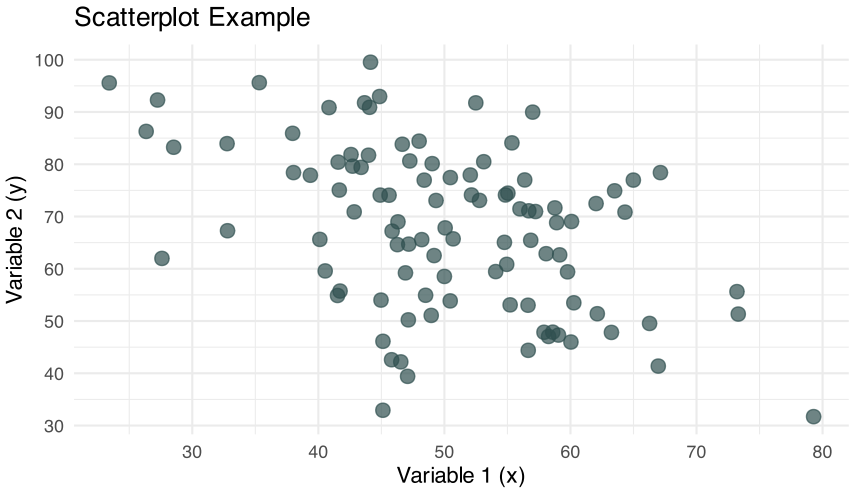
Third, the whiskers extend from the box to the minimum and maximum values within a certain range. The length of the whiskers can vary. There are several equations that may be used. It is common for any data outside of the whiskers to be considered extreme or as outliers.

Fourth, outliers are individual data points that fall significantly outside the typical range of the data. They are often plotted as individual points or dots.

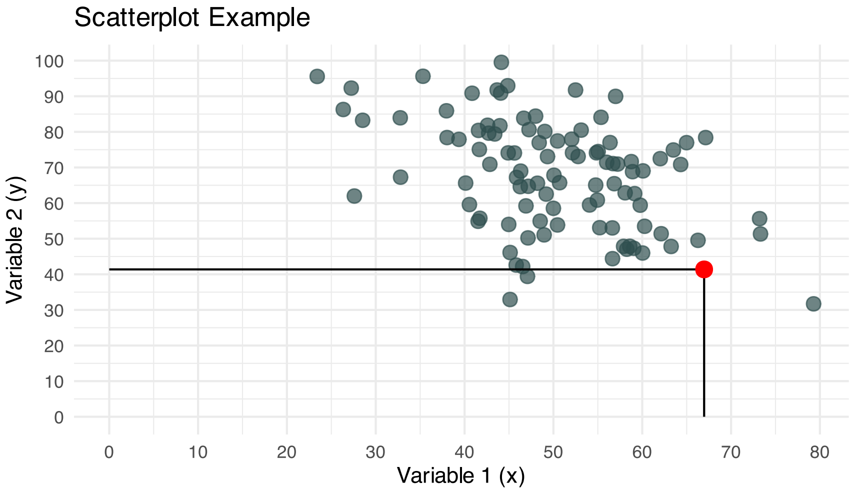
Box-plots are a great way to visualize a distribution and can inform viewers on the nuances of the data not understood through a histogram.

4.2.3 Scatter plots

Scatter-plot display the relationship between two variables in a two-dimensional space.



Each dot represents a participant. Each participant has a score on two variables. For example, let's consider one participant who scored around 67 on variables 1 (x) and 41 on variable 2 (y). Can you pinpoint their location on the above figure?



Scatter-plots serve as powerful tools for exploring the relationship between two variables in a data-set. The direction in which points trend across the two-dimensional plane, whether upward (i.e., dots get higher as you look from left to right) or downward (i.e., dots get lower as you look from left to right), offers insights into the association between the variables. Conversely, if the dots seems to form a circle, horizontal line, or vertical line, no meaningful relationship may exist.

As the relationship between two variables increases, the points will converge to a diagonal line. If the correlation between two variables is exactly −1.0 or 1.0, all dots will fall perfectly on a diagonal line. For example, here are sample scatter plots for various correlations.

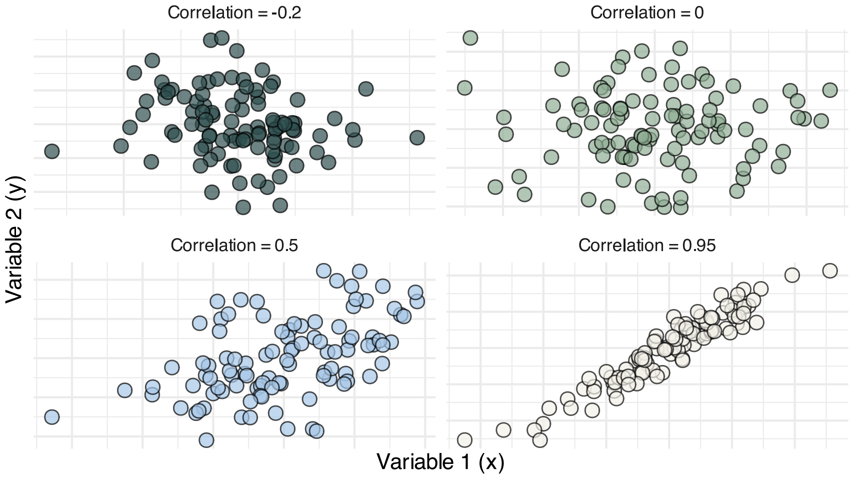
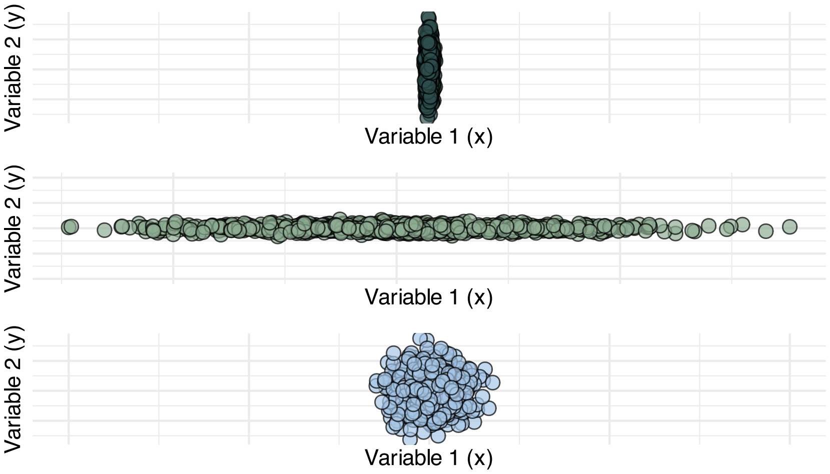


Figure 6: Visualizing different correlations.

The following correlations are all 0 (i.e., no relationship).



Descriptive statistics provide a concise summary of the main features of a data-set, aiding in the interpretation and communication of data patterns. These statistics are fundamental for understanding the characteristics of a data-set before applying more advanced statistical analyses or drawing conclusions based on the data.

4.3 Inferential Statistics

Our theories and hypotheses are typically assumed to apply to a specific population. Some examples of populations we may be interested in are men, those with depression, university students, or Grade 3 French Immersion Students. Sometimes our population of interest is the general public, everyone.

Sampling from a population is essential in research because studying an entire population is often impractical due to limitations in time, cost, and accessibility. For example, in large populations, such as all residents of a country or all patients with a particular medical condition, it is usually impossible to gather data from every individual. Thus, we must take a **sample–or a subset of the population**. Researching a sample that is representative of the population allows researchers to draw meaningful conclusions about the whole population while conserving resources. However, the challenge lies in selecting a sample that accurately represents the population to ensure the findings can be generalized.

A sample is a subset of the population that will be studied, which allows researchers to draw meaningful conclusions about the whole population.

4.3.1 Sampling Methods

To ensure that a sample is unbiased, researchers use several strategies. First, **random sampling** is one of the most reliable methods, where each individual in the population has an equal chance of being chosen. This reduces the likelihood of selection bias and ensures that the sample mirrors the diversity of the population. For instance, if studying a town's attitudes towards healthcare, random sampling would give each resident an equal chance of being selected, regardless of age, gender, or socioeconomic status. Because each individual is randomly selected, the sample as a whole should not be systematically different from the population of interest.

Second, researchers can use **stratified sampling**. Here, the population is divided into subgroups or strata based on specific characteristics (such as age, gender, or income). From each subgroup, random samples are then drawn. This method is particularly useful when researchers want to ensure that all key groups within the population are represented in the sample. For example, in a study of high school students' academic performance, stratified sampling might involve dividing students by grade level and then randomly selecting a proportional number from each grade. This ensures an equal number

A third method is **systematic sampling**. Systematic sampling offers a structured–yet efficient!–way of selecting a sample. In this method, every n^{th} person in the population is selected, with the starting point chosen randomly. This method can be easier to implement than random sampling while still providing a fair representation of the population, as long as there is no hidden pattern in the population that could influence the selection.

A fourth method is **cluster sampling**. Here, the researcher divides the population into groups or clusters, such as geographic regions or schools, and then randomly selecting entire clusters for study. This method is particularly helpful when the population is large or spread out over a wide area, as it simplifies the data collection process. However, it may introduce bias if the selected clusters systematically differ from the population. When this is the case, the sample will not be representative of the entire population.

One last method I will mention is **snowball sampling**. Snowball sampling can begin with another sampling strategy, such as random sample. However, when a participant completes the study, they are asked to recruit or refer additional participants from their networks. This method is particularly useful for niche populations of interest, or groups that are hard to identify and reach through other sampling methods.

Sampling from a population is vital for making research manageable and cost-effective. Using methods like random sampling, stratified sampling, systematic sampling, and cluster sampling helps ensure that the sample is unbiased and representative, allowing researchers to confidently generalize their findings to the broader population. Deciding which method is best for you will depend on your time, money, population of interest, and research question.

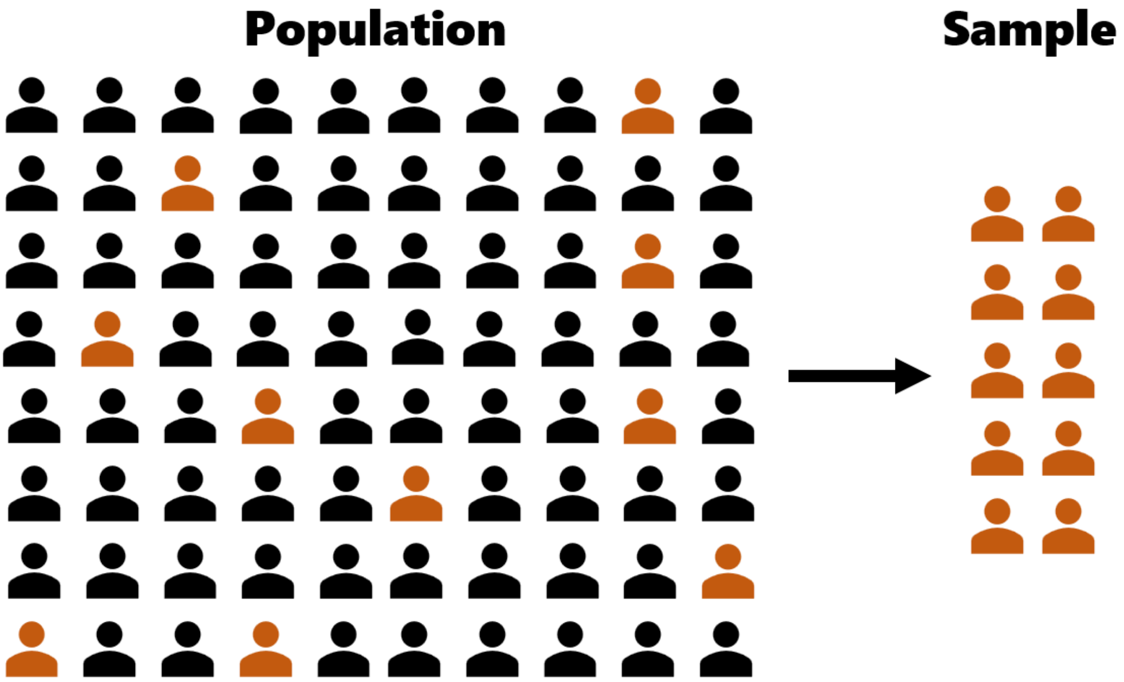
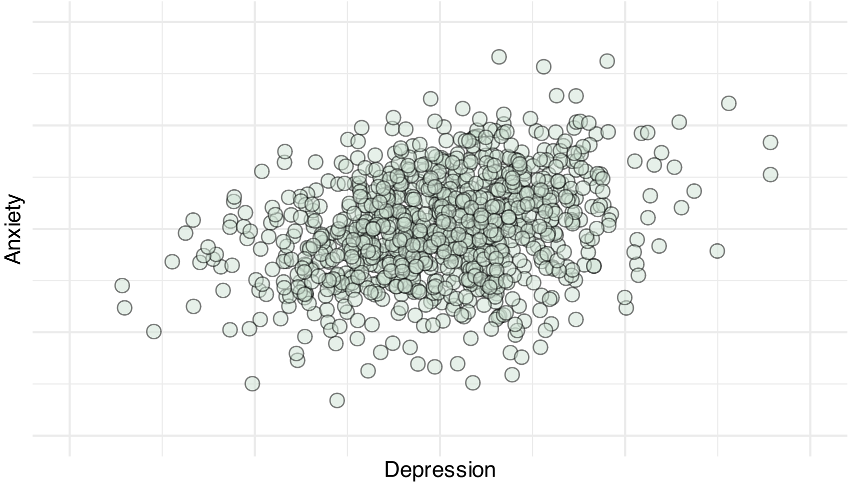


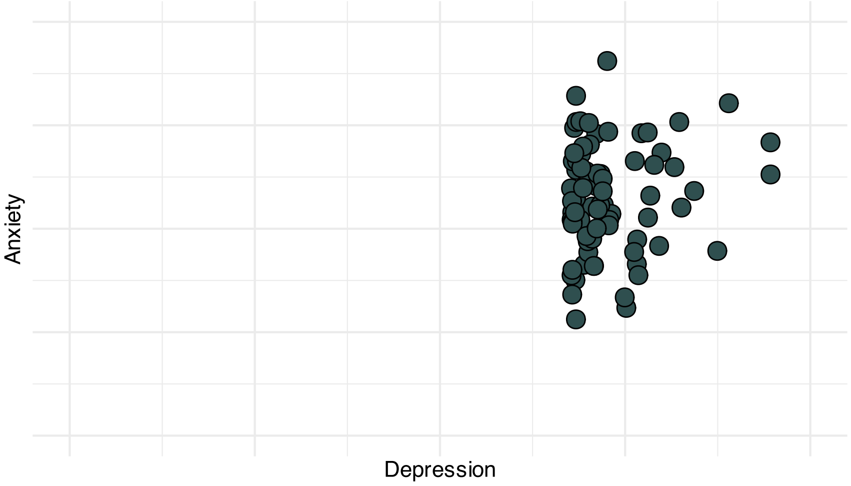
Figure 7: The left depicts the population. The orange individuals, who are randomly selected, are the sample.

After we identify a suitable sample, we collect data from that sample. As discussed, a major requirement is that *the sample is representative of the population of interest*. It is imperative to inferential statistics that this is the case. Specifically, the sample should share the same characteristics as the population of interest. Why does this matter? Importantly, once we analyze our data using the sample, we assume the results generalize to the entire population. *We infer about the population based on research using samples*: **inferential statistics**. The more discrepant our sample is from our population, the less likely the results of our results will generalize to the population.

To demonstrate, consider the following example. Imagine we are interested in understanding the link between depression and anxiety in Grenfell Campus Students (our population). While we typically will never know the true data of the population–if we did, we wouldn't need to sample in the first place–let's assume we do. Here is a figure representing **all** Grenfell Students.



We don't have the time or money to sample **all** students. Perhaps we recruit students outside of heath services–they conveniently let us set up a a recruitment table–and end up with data from 100 students. These 100 individuals are our sample. Their specific data can be viewed below (indicated by dark purple). Visually inspect the data we collect and compare it to the population values. Do you notice any discrepancies between the trend?



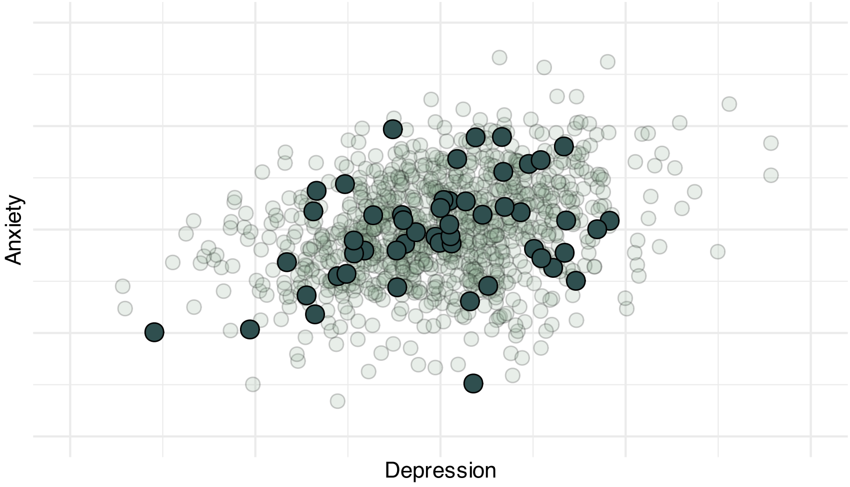
A formal analysis would reveal that this sample has a correlation between depression and anxiety of $.001$, which is an underestimate of the true population correlation of $.3$. Why might this sample be unrepresentative of the population? Sometimes it's simply random chance, but in this example it's likely something systematic.

Think about it.

We sample outside of the campus health center. Might these students be truly representative of the population of interest?

Perhaps students who frequent health services, which also includes mental health services, are more likely to have higher levels of depression and/or anxiety. Thus, this sample is unlikely to be truly representative of all Grenfell students.

Instead, imagine we take a truly random sample by randomly selecting 100 students numbers and having those students complete our study. We obtain the following data. Note that **light green** represents all students, while **dark green** represents our sample.



We would conduct analysis with the sample and infer that they generalize to the population! In the above example, the relationship between anxiety and depression is 0.34. While in the real world we don't know the true population parameter/data, in this hypothetical example, the TRUE correlation is 0.3. Our sample statistics aligns relatively well with the population parameter.

4.4 Conclusion

This chapter provided a introduction to some commonly used descriptive statistics, and their benefits, downsides, and uses. Furthermore, commonly used figures/graphs were introduced. Last, the ways in which samples are used to infer larger scale conclusions about populations of interest were detailed.

The following chapters outline the various analyses we can use to make population-based inferences from samples of data. Each analysis has situations where it can be best used, which will largely depend on the research question and hypotheses of interest. Remember, the research question determines the method, not the other way around.

4.5 Practice Questions

1. How can inferential statistics help psychologists draw meaningful conclusions from data, beyond just describing the sample at hand?
2. Can inferential statistics be misinterpreted or misused, and what steps can psychologists take to ensure the accuracy and validity of their statistical inferences?
3. What challenges and opportunities arise when applying inferential statistics to complex psychological phenomena, such as emotions, cognition, or interpersonal relationships?
4. How do cultural and contextual factors impact the appropriateness and interpretation of inferential statistics in cross-cultural psychological research?

The following are 10 salaries of randomly sampled Canadians.

Salaries
42000
42000
36000

Salaries
58000
17000
24000
67000
87000
41000
32000
525000

5. Calculate the mean with and without the outlier (the last observation).

4.6 Answers

5.

Mean with: $94,727.27

Mean without: $53,111.11

As you can see, the mean with is WAY higher than the mean without with one outlier. Imagine the answer to the question, ‘what is the average Canadian salary?’. How might you answer?

5 Probability

Imagine you are at a party. People slowly start coming to the party; the room you are in is starting to get full. Probability questions: How many people need to be at the party for there to be a 50/50 chance (p=.5) that two people have the same birthday? The answer is the end of the chapter.

Understanding probability is integral to understanding research and statistics in psychology. Specifically, probability is a branch of mathematics that deals with the likelihood of an event occurring. It quantifies uncertainty and helps us make informed decisions based on the likelihood of various outcomes. Probability values range from 0 to 1, where 0 indicates an impossible event (i.e., it will never happen), and 1 indicates a certain event (i.e., it's guaranteed to happen).

5.1 Basic Concepts

The following are some basic concepts we must understand prior to diving into probability. An **outcome** is a possible result of a situation. An **event** is a specific outcome or set of outcomes. Finally, the **sample space** is the set of all possible outcomes of a situation.

Definitions

Outcome is a possible result of a situation.

Event is a specific outcome or set of outcomes.

Sample space is the set of all possible outcomes of a situation.

As a practical example, consider flipping a coin. An outcome is a single possible result of the situation; in this case, the outcomes are Heads (H) and Tails (T) (we will ignore the fact that it is entirely possible for the coin to land on its side). The sample space is the set of all possible outcomes, which for a coin flip is represented as {H, T}. Additionally, an event is a specific set of outcomes. For example, we can define *Event A* as "getting Heads", which can be represented as the set {H}. Alternatively, we might define Event B as getting either Heads or Tails, represented as the set {H, T}. This framework helps us understand and calculate probabilities based on different scenarios.

5.2 Calculating Probability

The probability P of an event A is calculated using the formula (note the syntax of $P(A)$:

$$P(A) = \frac{\text{Number of favourable outcomes for } A}{\text{Total number of outcomes in sample space}}$$

💡 Think about it

Favorable does not mean we necessarily *want* the outcome. Favorable means that the event in question happens. It's favorable for the event.

For example, consider a simple experiment: rolling a six-sided die. We want to test to probability of rolling an even number. To this end, we can derive the following:

- **Sample Space (S)**: {1, 2, 3, 4, 5, 6}
- **Event (A)**: Rolling an even number {2, 4, 6}

To calculate probability, we can sum the number of favorable outcomes and divide it by the total sample space. There are three favorable outcomes: rolling a 2, 4, or 6. The total sample space is **all** the possible outcomes, of which there are six: rolling a 1, 2, 3, 4, 5, or 6. Thus:

1. **Number of favorable outcomes for A**: 3 (2, 4, 6)
2. **Total number of outcomes in sample space**: 6

Using the above formula, we can calculate the probability:

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

So, the probability of rolling an even number is $P = \frac{1}{2} = .5$, or 50%.

5.3 Compound Events

5.3.1 AND Probability

The **AND** probability (also known as joint probability) is used when we want to find the probability that two events A and B both occur. The formula for calculating the AND probability of two independent events is:

$$P(A \quad \text{AND} \quad B) = P(A) \times P(B)$$

So, the joint probability can be calculated by multiplying the independent probabilities together.

Consider two independent events. We have one die and want to determine the probability of:

- Event A: Rolling a 3 on the first roll and then
- Event B: Rolling an even number on the second roll

We first calculate the probability for each event independently.

1. **Probability of Event A**: $P(A) = \frac{1}{6}$ (only 1 favorable outcome)
2. **Probability of Event B**: $P(B) = \frac{3}{6} = \frac{1}{2}$ (3 favorable outcomes)

Using the formula:

$$P(A \quad \text{AND} \quad B) = P(A) \times P(B) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

Thus, the probability of rolling a 3 and then rolling an even number is $\frac{1}{12}$. Note that we cannot roll both a 3 and an even number on one die. Also, it's imperative to recognize that the events are *independent* and do not influence one another. Thus, in the example we are rolling a 3 and **then** rolling an even number. Or, having two dice and having one land on 3 and the other being an even number.

5.3.2 OR Probability

The **OR** probability is used when we want to find the probability that at least one of two events, A or B, occurs. The formula for calculating the OR probability is:

$$P(A \quad \text{OR} \quad B) = P(A) + P(B) - P(A \quad \text{AND} \quad B)$$

Let's calculate the OR probability for the same events from before. Imagine we have one die and our favorable events are:

- Event A: Rolling a 3 OR

- Event B: Rolling an even number

We can calculate the independent probabilities:

1. **Probability of Event A**: $P(A) = \frac{1}{6}$
2. **Probability of Event B**: $P(B) = \frac{1}{2}$
3. **Probability of Event A AND B**: $P(A \text{ AND } B) = 0$ (rolling a 3 and an even number is impossible)

We can substitute the values into our formula to derive:

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$$

$$= \frac{1}{6} + \frac{1}{2} - 0 = \frac{1}{6} + \frac{3}{6} = \frac{4}{6} = \frac{2}{3}$$

Thus, the probability of rolling a 3 or an even number is $\frac{2}{3}$. If we have a fair die, we will makes a bunch of rolls, we would expect to roll a 3 or an even number 75% of the time. This intuitively makes sense when we consider that their are four favorable outcomes $(2, 3, 4, 6)$ and six total outcomes $(1, 2, 3, 4, 5, 6)$.

5.4 Conclusion

Understanding probability is crucial in psychological research, particularly in the context of null hypothesis significance testing (NHST; our next chapter). Probability helps researchers determine the likelihood of observing results under the assumption that the null hypothesis is true. By calculating the probability of different outcomes, researchers can assess the strength of their evidence against the null hypothesis, making it easier to identify statistically significant findings. Furthermore, learning how to interpret and combine probabilities allows researchers to better evaluate the risks of Type I and Type II errors, ultimately leading to more informed conclusions and decisions in psychological studies.

Answer to the riddle

Let's start by having one person in the room. Let's think about the probability that their birthday is unique. That's easy! Of course it is! The probability that their birthday is unique is

$$\frac{365}{365} = 1.00$$

What about the second person. What's the probability that their birthday is unique?

$$\frac{364}{365} = .99726$$

What about the third person?

$$\frac{363}{365} = .99452$$

You know that when we require two or more events to occur (an AND probability), we can multiply their respective probabilities. Thus, the probability of all three people having unique birthday is:

$$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} = .9918$$

This equation can be simplified when we know the number of *unique pairs*. With two people, there is one unique pair. When we have 10 people there are $\frac{10*9}{2}$ unique pairs. This can be calculated as:

$$p = \frac{n \times (n-1)}{2}$$

Where p is the number of unique pairs. This can be plugged into:

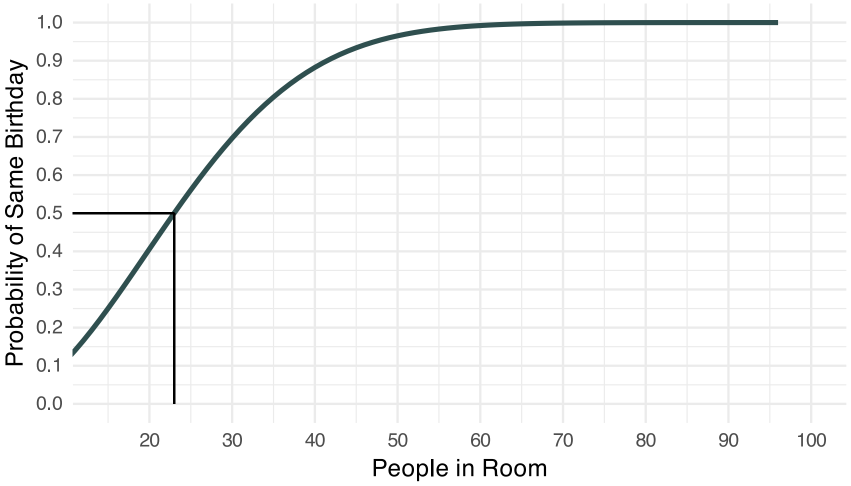
$$1 - \left(\frac{364}{365}\right)^{p}$$

Let's plug in a number....20. The results are:

$$1 - \left(\frac{364}{365}\right)^{\frac{20\times19}{2}} = 0.6474$$

This means there's a 64.74% chance that 20 people in a room all have different birthdays–or a ~35% chance they have at least one overlapping birthday.

We can plot the results as:



5.5 Practice Questions

Calculate the probability of the following:

1. A standard six-sided die is rolled. What is the probability of rolling a number greater than 4?
2. A standard deck of 52 playing cards is shuffled. What is the probability of drawing an Ace from the deck?
3. A fair coin is flipped once. What is the probability of getting Heads?
4. A bag contains 5 red marbles, 3 blue marbles, and 2 green marbles. If one marble is drawn at random, what is the probability of drawing a blue marble?
5. Two six-sided dice are rolled. What is the probability that both dice show a number greater than 3?
6. From a standard deck of 52 playing cards, what is the probability of drawing a King and then drawing a Queen without replacement?

7. A bag contains 4 red marbles and 6 blue marbles. If two marbles are drawn one after the other without replacement, what is the probability that both marbles are red?
8. Two fair coins are flipped. What is the probability that both coins show Heads?
9. A standard six-sided die is rolled. What is the probability of rolling a 2 or a 5?
10. From a standard deck of 52 playing cards, what is the probability of drawing a Heart or a Spade?
11. In a basket containing 3 apples, 2 oranges, and 5 bananas, what is the probability of randomly selecting an apple or a banana?
12. A fair coin is flipped twice. What is the probability of getting Heads at least once?

5.6 Answers

1. **Favorable outcomes:** Rolling a 5 or 6 (2 outcomes: {5, 6}). **Total outcomes:** 6 (numbers: {1, 2, 3, 4, 5, 6})

$$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}} = \frac{2}{6} = \frac{1}{3}$$

2. **Favorable outcomes:** There are 4 Aces in the deck. **Total outcomes:** 52 cards

$$P(A) = \frac{4}{52} = \frac{1}{13}$$

3. **Favorable outcomes:** Getting Heads (1 outcome: {H}). **Total outcomes:** 2 (Heads and Tails: {H, T})

$$P(A) = \frac{1}{2}$$

4. **Favorable outcomes:** There are 3 blue marbles. **Total outcomes:** 5 red + 3 blue + 2 green = 10 marbles.

$$P(A) = \frac{3}{10}$$

5. **Favorable outcomes:** Rolling a 4, 5, or 6 on each die. **Total outcomes:** Each die has 6 outcomes, so the total outcomes for two dice is $6 \times 6 = 36$.

The probability of one die showing greater than 3 is $\frac{3}{6} = \frac{1}{2}$.

Therefore, for both dice:

$$P(\text{both} > 3) = P(\text{die } 1 > 3) \times P(\text{die } 2 > 3) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

6. **Favorable outcomes:** 4 Kings, followed by 4 Queens. **Total outcomes:** 52 cards, then 51 remaining cards.

$$P(\text{King and Queen}) = P(\text{King}) \times P(\text{Queen} \mid \text{King})$$

$$= \frac{4}{52} \times \frac{4}{51} = \frac{16}{2652} = \frac{4}{663}$$

7. **Favorable outcomes:** 4 red marbles. **Total outcomes:** 10 marbles.

$$P(\text{both red}) = P(\text{first red}) \times P(\text{second red} \mid \text{first red})$$

$$= \frac{4}{10} \times \frac{3}{9} = \frac{12}{90} = \frac{2}{15}$$

8. **Favorable outcomes:** Both coins show Heads (1 outcome: {HH}). **Total outcomes:** 4 outcomes: {HH, HT, TH, TT}.

$$P(\text{both Heads}) = \frac{1}{4}.$$

9. **Favorable outcomes:** Rolling a 2 or a 5 (2 outcomes: {2, 5}). **Total outcomes:** 6 outcomes.

$$P(2 \quad \text{or} \quad 5) = P(2) + P(5) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}.$$

10. **Favorable outcomes:** 13 Hearts + 13 Spades = 26. **Total outcomes:** 52 cards.

$$P(\text{Heart or Spade}) = P(\text{Heart}) + P(\text{Spade}) = \frac{13}{52} + \frac{13}{52} = \frac{26}{52} = \frac{1}{2}.$$

11. **Favorable outcomes:** 3 apples + 5 bananas = 8. **Total outcomes:** 10 fruits.

$$P(\text{apple or banana}) = P(\text{apple}) + P(\text{banana}) = \frac{3}{10} + \frac{5}{10} = \frac{8}{10} = \frac{4}{5}.$$

12. **Favorable outcomes:** HH, HT, TH (3 outcomes). **Total outcomes:** {HH, HT, TH, TT} (4 outcomes).

$$P(\text{at least 1 Head}) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$$

6 Statistical Models

This chapter introduces statistical models. These models are of utmost importance in our analyses. Our models are often derived from our hypotheses and inform us the degree to which what we expected to happen aligns with the data. We are comparing our expectations with our observations. We expect a ball we throw up to come down. Does it? We expect that when we study, we will perform better. Do we? Or, we expect people with higher depressive symptoms to have an increase probability of making a suicide attempt compared to those with less depressed symptoms. Do they?

Definition

A statistical model is a mathematical framework used to represent, analyze, and make predictions about data.

It is often used to explain relationships between variables, such as the effect of one or more independent variables (predictors) on a dependent variable (outcome).

Statistical models help researchers. First, they help provide a simplified representation of complex data. This makes it easier to understand and interpret patterns or trends. Second, they help us directly test our hypotheses. Researchers use statistical models to evaluate whether there are significant relationships between variables as we expect, or whether observed patterns are random noise in the data. Last, they help make predictions. Based on the relationships identified in the data, models can predict future outcomes or behaviors.

When we use theory and generate hypotheses, we can translate our hypotheses into statistical models to test. These are derived from, and quite similar to our statistical hypotheses (see Section 2). We are attempting to create a model of real-world phenomenon. Perhaps the best way to conceptualize a model is to think of outcomes and their explanations. Given some outcome of interest, we can provide some model or explanation for individual differences in the outcome. But, our models may not *perfectly* explain the outcome and, thus, we must consider some degree of error. Formally:

$$outcome = model + e_i$$

For example, consider the following theory proposed by Schneidman (1998): suicide is caused by psychache (unbearable mental pain). Even if Shneidman was right, and this was **truly** how suicide operated in the real world, we simply cannot confirm this. As a result, others may disagree on the answer to 'what causes suicide?' There are many potential explanations and subsequent models. Consider the following three possible explanations:

1. Researcher 1: as pyschache increases, suicide risk increases
2. Researcher 2: as social connectedness decreases, suicide risk increases
3. Researcher 3: presence of specific genetic variant (SNP) Chr13:rs34399104 causes suicide

Each researcher could collect data to test how well their hypothesis fits the data that they collect. Each hypothesis can be represented as a model that is statistically testable. Researcher 1, who believes that knowing someone's score on psychache should allow us to predict their suicide risk, may model the data as:

$$y_{risk} = x_{psychache} + e_i$$

Above the outcome is suicide risk. The researchers believes that someone's degree of experiencing psychache can explain their degree of suicide risk. Because it's not a perfect explanation, there is some error incorporated. If the model explains a lot of the outcome, error will be low. If it doesn't, error will be high.

Researcher 2, who believes that people high on social connectedness should have less suicidal ideations than those with low social connectedness, may model the data as:

$$y_{ideations} = x_{connectedness} + e_i$$

Researcher 2 believes that if someone dies by suicide, then the Chr13:rs34399104 (Docherty et al. (2020)) gene should be present in a biopsy. They may model the data as:

$$y_{suicide} = x_{Chr13:rs34399104} + e_i$$

Also, the researcher would propose that the probability of someone dying by suicide, given the presence of Chr13:rs34399104 is 1.00, or:

$$p(death \mid Chr13 : rs34399104) = 1.00$$

So, we have three suicide researchers all proposing a different explanation for various aspects of suicidality. Thus, each researcher would collect slightly different data and analyse it differently. If their model fits the data well, it provides support for the hypothesis and theory. If it does not fit the data well, it likely does not accurately represent the real-world phenomenon of interest. For example, if researcher 3 collected genetic data and the presence of the hypothetical gene did not lead to suicide in some individuals, it would indicate a poor model fit. Thus, their hypothesis is not supported and the theory should be adjusted or scrapped.

Models can be simple or complex. Again, the research question and hypotheses precede the research design–and, subsequently, the model.

If you are still having difficulty conceptualizing a model, perhaps it's best to think about each individual and their data. The data will be the function of our model plus some error. Typically, we have their *observed* score and *predicted* score on some variable. Our predictions, however, are not always accurate. Instead, there is some degree of error. We may express it as:

$$y_i = model + error$$

Again, if our model–a proposed explanation on how the data should fit together–does a good job, errors will be relatively low. If our model does a bad job, the errors will be relatively low. A basic example may help further understand models.

6.1 A Basic Model

Let's try to model the mean height of psychology professors (in centimeters). You cannot measure all the psych professors in the world. Instead, you go to the Arts and Sciences Building at Grenfell Campus and measure the heights of four of your psychology professors. You get the following data.

Name	Height
Tyler	181
Steve	190
Jenny	173
Cindy	158

We can represent our model as:

$$y_i = X + e_i$$

Here: y_i presents the height of professor i, X represents some constant or model that we will use to try and predict a professor's height; and e_i, the error, represent the difference between the professor and the constant. Errors are also sometimes referred to as residuals. Note that each participants, or experimental unit, gets a residual. Here, professors–our experimental units–gets their heights–out variable–measured.

We can assess how well the model fits with the data we collected. For our model, we can try to calculate how large our residuals (e_i) are, as these represent the model error. Recall that a model that does a poor job will have more error compared to a model that does a good job.

Let's propose two models: 1) where $x = 150$ and 2)$x = 180$. Let's calculate the residuals for each professor for each model.

Model 1:

$$y_i = 150 + e_i$$

And the error for Tyler, who is 181cm, would be:

$$181 = 150 - e_{tyler}$$

Which means that:

$$e_{tyler} = 150 - 181 = -31$$

Here are the residuals for the other professors:

Name	Height	Error
Tyler	181	−31
Steve	190	−40
Jenny	173	−23
Cindy	158	−8

Let's do the same thing for Model 2:

$$y_i = 180 + e_i$$

An the errors are:

Name	Height	Error
Tyler	181	−1
Steve	190	−10
Jenny	173	7
Cindy	158	22

Simply inspecting the errors, we can see that the errors are on average higher for Model 1 compared to Model 2. Thus, Model 2 seems to be a better fit to the data. If you were asked to predict a professors height, you would be more accurate to guess 180cm versus 150cm.

Are these the best guess, though? While both 150 and 180 may seem like good guesses, in this simple scenario, the average height of the professors–the mean ($\bar{x}$)–will be the best fit for the data. The average height of these professors is 175.5cm. Thus, the best model that contains only one piece of information is:

$$y_i = \bar{x}_i + e_i = 175.5 + e_i$$

Where $\bar{x}_i$ is the mean of the group.

In the above model, I explicitly state that we only have one piece of information. However, some models will have more then one piece of information. For example, imagine we model using two pieces of information: the mean height of professors and the gender of the professor. We will now have two pieces of information to model: $\bar{x}_i$ and $gender$. We will also still have our error/residuals. Compare the errors of both of these models.

Model using mean only:

$$y_i = \bar{x}_i + e_i$$

Name	Height	Error
Tyler	181	−5.5
Steve	190	−14.5
Jenny	173	2.5
Cindy	158	17.5

Model using mean height and gender:

$$y_i = \bar{x}_i + gender(x_{2i}) + e_i$$

Where x_{2i} is someone's score on gender. We will revisit this later, but women will get a score of 0 on this variable, and men a score of 1. Let's assume men will be on average 20cm taller than women, meaning in the above model, $gender = 20$. Also, let's make the mean of the model the mean height for women, which is 166cm. Replacing the above equation with this new information, the following model is derived:

$$y_i = 166 + 20(x_{2i}) + e_i$$

Where x_2i is the gender variable for participants (0 for women and 1 for men). Thus, the equation will look differently for men versus women.

For women:

$$y_i = 166 + 20(0) + e_i = 166 + e_i$$

For men:

$$y_i = 166 + 20(1) + e_i = 166 + 20 + e_i$$

The following are the new errors:

Name	Height	Gender	Error
Tyler	181	Male	5
Steve	190	Male	−4
Jenny	173	Female	−7
Cindy	158	Female	8

As can be seen, the errors *seem* smaller. However, *seeming* smaller isn't quite scientific. Is there a better way?

6.2 Deviations

The astute reader may have noticed something peculiar. How can we tell how good a model fits the data? Perhaps we could add the residuals. Let's do this for our model that uses *only the mean of professors' heights (175.5cm)* to predict an individual professor's height and compare to the model that uses 150cm. As a refresher, here's the data:

Name	Height	Error175.5cm	Error150cm
Tyler	181	5.5	31
Steve	190	14.5	40
Jenny	173	−2.5	23
Cindy	158	−17.5	8

If we sum all the errors up across all our data when consider the mean, we get:

$$\sum e_i = 5.5 + 14.5 + (-2.5) + (-17.5) = 0$$

Let's compare that to the errors for the 150cm model:

$$\sum e_i = -31 + -40 + -23 + -8 = -103$$

Thus, the mean model seems a better fit than the 150 model; the residuals are closer to 0. As noted above, for this simple model with one piece of information, the mean will always provide the best estimate ($\bar{x}$) and have the residuals sum to 0:

$$\sum_{i=1}^{n} e_i = 0$$

Other more complex models will have this property. But some may not. Often, the squared residuals are used in place of the absolute residual. Variance and standard deviation are directly related to this.

6.2.1 Variance and Standard Deviation

We may effectively model the fit of our mean model with the variance and standard deviation. These are extremely important in statistics so it's imperative to become familiar with them.

Above we calculated the the deviation of each score. The variance is, in essence, the average squared difference between a score and its mean.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

But for a sample, our equation is (see Section 4):

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

This equation simply means we add up all the squared differences between a score and the mean (here this is the residual) and divide by the number of scores. So, the squared deviations are:

Name	Error	Squared
Tyler	5.5	30.2
Steve	14.5	210.2
Jenny	−2.5	6.2
Cindy	−17.5	306.2

We then add up the squared deviations, $30.2 + 210.2 + 6.2 + 306.2 = 552.8$. And divide by the number of scores (with sample adjustment to $N - 1$), $4 - 1 = 3$, to get:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1} = \frac{30.2 + 210.2 + 6.2 + 306.2}{4 - 1} = \frac{552.8}{3} = 184.27$$

Thus, the variance of the heights of psychology professors is 184.27. The standard deviation is simply the squared root of the variance:

$$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1}} = \sqrt{184.27} = 13.58$$

While you might think that the standard deviation (SD) is the average absolute difference between a score and the mean, it is not. For example, the SD of our heights is 13.58. But the average deviation is, in fact, $\frac{|5.5|+|14.5|+|-2.5|+|-17.5|}{4} = 13.33$. It is most likely helpful to think of the variance as the average squared deviation and the SD as the root of the variance.

6.3 Try This

Instead of using the mean in the above model, use a value of 190cm. Our new model would be:

$$x_i = 190 + e_i$$

Calculate the errors, variance and SD using this new model. Was the variance higher, the same or lower?

Which model seemed better? The one using the mean or 190cm?

Answers

Name	Height	NewDeviation	NewSquaredDeviation
Tyler	181	−9	81
Steve	190	0	0
Jenny	173	−17	289
Cindy	158	−32	1024

The sum of these new deviations is 1394.

The variance of these is 464.67.

The SD is 21.56.

When you compare the errors for one model, which uses the mean (175.5cm) as the best estimate of professors' heights versus the model that uses 190cm, we determine that the mean is the best model for the data. It has smaller errors/residuals. While the modeling the data with a mean is a simple model, there are more complex or advanced ways to model data.

6.4 Advanced Models

While above we have simply modeled a mean, later chapters will build up to more advanced models, such as:

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{2i}x_{3i}\beta_4 + e_i$$

Don't be intimidated, this is a whole lot like your classic high school's $y = mx + b$, with some intercepts and slopes. More to come. For now, a brief overview of some potential models will do. I will note that many common statistically models fall under the broader umbrella of general linear models (GLM). Some common types of statistical models in psychological research that will likely encounter in the literature include:

1. Linear regression models

These assess the relationship between one or more independent variables and a continuous dependent variable. For example, predicting levels of anxiety based on hours of sleep.

2. Analysis of Variance (ANOVA)

This model compares the means of different groups to determine if they are significantly different from one another, often used in experimental studies.

3. Structural equation modeling (SEM)

SEM is a more complex statistical model that can evaluate multiple relationships between variables simultaneously, including latent (unmeasured) variables.

6.5 Conclusion

Statistical models are essential for drawing conclusions about psychological phenomena, helping researchers identify patterns, test theoretical models, and inform practice. Our models are derived from our hypotheses. Each analysis in your hypothetical toolbox will allow you to model data in appropriate ways to test you hypotheses. We will revisit the idea of models throughout each chapter that follows.

6.6 Practice Questions

1. Calculate the mean, variance, and standard deviation for both the height (in cms) and weight (in kgs) of these NHL players.

Player	Height	Weight
Connor McDavid	185	99
Auston Matthews	190	93
Sidney Crosby	180	91
Alex Ovechkin	191	108

2. Write out the model for NHL height.
3. What are the e_i values for each player when modeling their height?

6.7 Answers

1.

```
                    V1
Mean_Height 186.725000
SD_Height     5.148058
var_Height   26.502500
Mean_Weight  97.725000
SD_Weight     7.587435
var_Weight   57.569167
```

2. Write out the model for NHL height.

$$height_i = \overline{x}_{height} + e_i$$

3. What are the e_i values for each player when modeling their height?

Player	Height	e_i
Connor McDavid	185	−1.3
Auston Matthews	190	3.8
Sidney Crosby	180	−6.7
Alex Ovechkin	191	4.3

7 NHST

Null hypothesis significance testing (NHST) is a controversial, yet widely used approach to testing hypotheses. It is likely the most commonly-used approach in psychological science. Simply, NHST begins with an assumption, a hypothesis, about some effect in a population. Interestingly, the hypothesis is that there is no effect or relationship. More to come on this. Regardless, data is collected from a sample that is believed to be representative of that population (see Section 4). Should the data not align with the null hypothesis, it is taken as evidence against it. There are several concepts related to NHST that need to be addressed to facilitate your understanding.

7.1 p-values

Prior to exploring p-values, let's ensure you have a basic understanding of probability notation. I recommend you Section 5 first. Back to the notation…:

$$p(x)$$

which would indicate the probability of x.

For example, the probability of flipping a fair coin and getting a heads is .5 ($p(heads) = .5$), indicating a probability of getting a heads. Probability ranges from 0 (no chance), to 1 (guarantee). For example, the probability that you, the person reading this, is a Grenfell student is high. Thus, my best guess right now that you are a 3950 student is that $p(3950student) = .95$. Hypothetically, if I could survey everyone who has read this sentence, I would expect 19 out of 20 of them (95%) to be a

Grenfell student. I would expect one of them not to be (e.g., someone randomly stumbled on this page…lucky you!).

Additionally, you must understand the notation for a conditional probability:

$$p(x \mid y)$$

which would indicates the probability of x, given we know y occurred/ is true.

For example, what if I informed you that the coin in the previous example is not a fair coin. The original probability will only be accurate if the coin is fair $p(heads \mid faircoin) = .5$. However, with the new information, the probability of getting heads is most certainly not .5, $p(heads \mid unfaircoin) \neq .5$. Perhaps our coin is biased to land on heads more often than tails. Thus, we would expect that $p(heads) > .5$.

The reverse of a conditional probability is not always equal to the original conditional probability. That is:

$$p(x \mid y) \neq p(y \mid x)$$

Consider the following: what is the probability that someone is Canadian, given that they are the prime minister of Canada: $p(Canadian \mid PrimeMinister)$? While not legally required, it is quite likely that if someone is the prime minister of Canada, they are Canadian. At the time of writing, Mark Carney is PM of Canada; he was born in NWT and grew up in Alberta. Thus, $p(Canadian \mid PrimeMinister) = 1.00$.

Do you think this is equal to the probability of someone being the Prime Minister of Canada, given that they are Canadian?: $p(PrimeMinister \mid Canadian)$? I would argue that the former is $p = 1.00$, while the later is not. In fact, the later can be calculated. Given there are about $41,000,000$ Canadians, and there is only one current Prime Minister, than the probability that someone is Prime Minister, given that they are Canadian $p(PM \mid Canadian) = \frac{1}{41,000,000} = .00000002$. So:

- $p(Canadian \mid PrimeMinister) = 1.00$
- $p(PrimeMinister \mid Canadian) = .00000002$

It is imperative understand conditional probability and that the inverse conditional probability is not necessarily different.

7.1.1 A Major "Given"

A key feature of NHST is that **the null hypothesis is assumed to be true**. Given this assumption, we can estimate how likely a set of data are. This is what a p-value tells you.

Definition

A **p-value** is the probability of obtaining data as or more extreme than you did, given a true null hypothesis. We can use our notation:

$$p(D \mid H_0)$$

Where D is our data (or more extreme) and H_0 is the null hypothesis.

When the p-value meets some predetermined threshold (i.e., *a priori criteria*), it is often referred to as statistical significance. This threshold has typically been $\alpha = .05$–there are debates over whether this is an arbitrary threshold or not (Cowles & Davis, 1982). Should data be so unlikely that is crosses the threshold (i.e., $p < .05$), we take it as evidence against the null hypothesis and reject it. If it is not below the threshold ($p \geq .05$), we fail to reject it.

Note that we do not *accept* the null hypothesis. Liken this to a courtroom verdict, which is that someone is guilty or not guilty. A not guilty verdict does not mean that someone is innocent, it means there is not enough evidence to convict. Likewise, failing to reject the null does not mean the null is true, rather, there's not enough evidence to conclude that it is false.

> …beginning with the assumption that the true effect is zero (i.e., the null hypothesis is true), a p-value indicates the proportion of test statistics, computed from hypothetical random samples, that are as extreme, or more extreme, then the test statistic observed in the current study.

> — Spence & Stanley (2018)

or, stated another way:

> The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.
>
> — Wasserstein & Lazar (2016)

Simply, a p-values indicates the probability of *your*, or more extreme, test statistic, assuming the null: $p(data \mid null)$. When we consider the results of our analyses, we can differentiate reality–how the world truly operates–with our research results–how the world should operate according to our results. Despite not knowing how the world truly operates, they should sometimes align. For ease, consider the following table:

		Reality	
		No Effect (H0 True)	Effect (H1 True)
Research Result	No Effect (H0 True)	Correctly Fail to Reject H0	Type I Error (α)
	Effect (H1 True)	Type II Error (β)	Correctly Reject H0

When we conduct NHST, we are assuming that the null hypothesis is indeed true, **despite never truly knowing this**. As noted, a p-value indicates the likelihood of your or more extreme data given the null.

If the null hypothesis is true, why would our data be extreme?

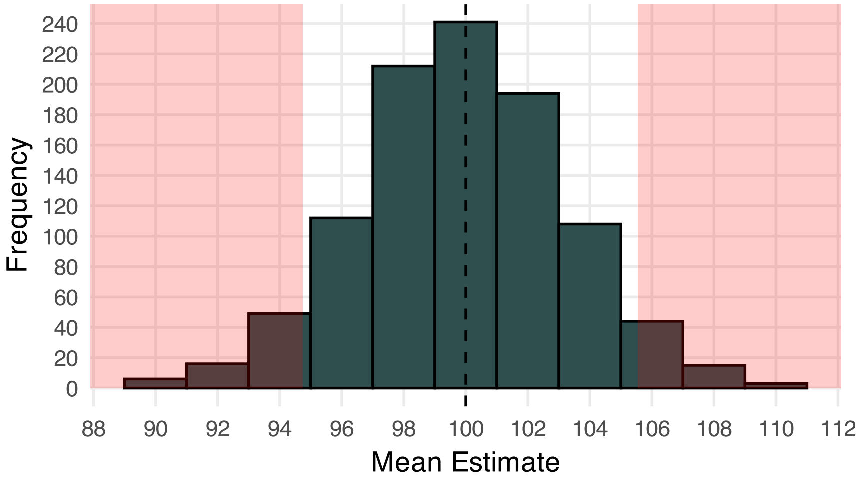
In inferential statistics we make inferences about population-level parameters based on sample-level statistics. For example, we infer that a sample mean is our best estimate of a population mean:

$$\hat{\mu} = \bar{x}$$

Where: - $\hat{\mu}$ is the estimate of the population mean and - $\bar{x}$ is the sample mean

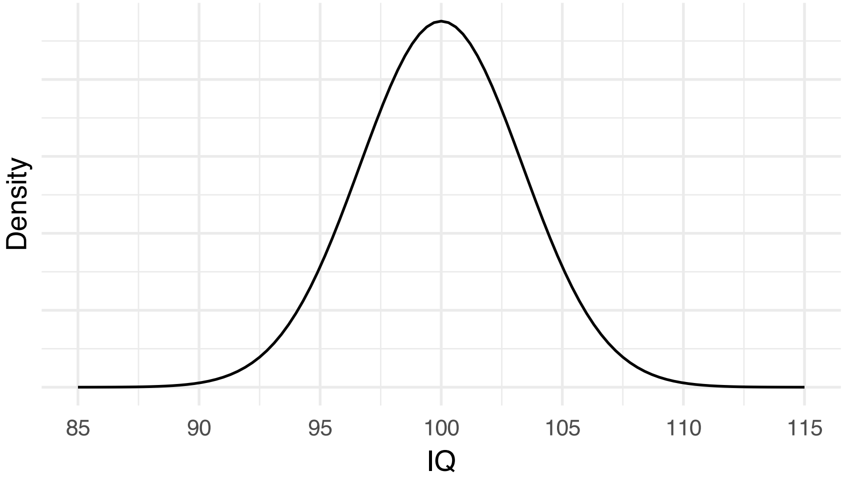
In, NHST, we assume that population-level effects or associations (i.e., correlations, mean differences, etc.) are zero, null, nothing. However, samples are only *estimates* of populations. Thus, if we sampled from a population whose true effect or association was 0 an infinite number of times, we would not always get a sample statistic of 0 (i.e., our true null effect) because of random variations in our sample. Instead, our *sample statistics* would form a distribution around the *population parameter*.

To help your understanding, let's first consider the IQ scores of a population–say residents of Corner Brook, NL. We will assume we know the true mean ($\mu = 100$) and standard deviation ($\sigma = 15$). If we took $1,000$ samples of size 20 from this population with $\mu = 100$ and $\sigma = 15$, we may get a distribution that looks something like:



The above shows the distribution of the means of 1,000 samples from the population where $\mu = 100$ and $\sigma = 15$. Although not quite an infinite number of samples, I hope you get the idea that the sample statistics form a normal distribution around the true population parameter. The red shaded region shows the two tails of this distribution that encompass the most extreme 5% of the samples (2.5% per tail). That is, 5% of the samples are in the tails (2.5% in each tail). These extreme 5% fall either below 94.75 or above 105.54. Thus, there is a low probability sampling from the population, measuring their IQ, and having the mean be in that red region. Only 5% of the samples are at or beyond this number. That data is quite unlikely given a true value of $\mu_{IQ} = 100$.

While we used simulated data from a hypothesized population, math people have derived formulas that represent the approximate shape of the curve for "infinite" samples. These are sometimes called probability density functions. We don't need to know the calculus behind it, but the PDF for $mu = 100$ and $\sigma = 15$ when sampling 20 people from the distribution is:



In the above, the extreme 5% of the PDF is associated with a sample mean lower than $\bar{x} = 92.98$ or higher than $\bar{x} = 107.02$; that is, the extreme 2.5% of samples have a mean IQ that is larger higher than $\bar{x} = 107.02$ or 2.5% have sample means lower than $\bar{x} = 92.98$.

To hammer home the concept, let's conduct a one-sample t-test on the following sample, assuming a population mean of 100. We obtain the following data:

Person	IQ
1	94.43
2	130.89
3	97.97
4	89.43
5	119.68

Person	IQ
6	106.44
7	129.16
8	115.69
9	81.65
10	93.58
11	106.19
12	124.06
13	117.38
14	108.2
15	127.77
16	100.86
17	91.02
18	113.98
19	105.75
20	86.26

This is a sample mean of 107.0199. The following are the results of a formal one-sample t-test.

```
Effect sizes were labelled following Cohen's (1988)
recommendations.

The One Sample t-test testing the difference between df_pop2$IQ
(mean = 107.02)
and mu = 100 suggests that the effect is positive,
statistically not
significant, and small (difference = 7.02, 95% CI [100.00,
114.04], t(19) =
2.09, p = 0.050; Cohen's d = 0.47, 95% CI [-1.90e-05, 0.92])
```

These results suggest the a sample mean of $\bar{x} = 107.00$ drawn from a population with mean $\mu = 100$ results in $p = .05$. What does this mean? How would you interpret this p-value? Try to link it to the distribution of samples means from above.

Think about it

The p-value indicates that given a true null hypothesis (here that the population mean is 100), we would expect a sample mean as big or bigger than ours only 2.5% of the time.

In the distribution of sample test statistics above, this aligns with the fact that 5% of the statistics (2.5% in each tail) were at least 7.02 away from the mean. 2.5% is at $100 + 7.02 = 107.02$ and beyond, the the other tail would be $100 - 7.02 = 92.98$ and below.

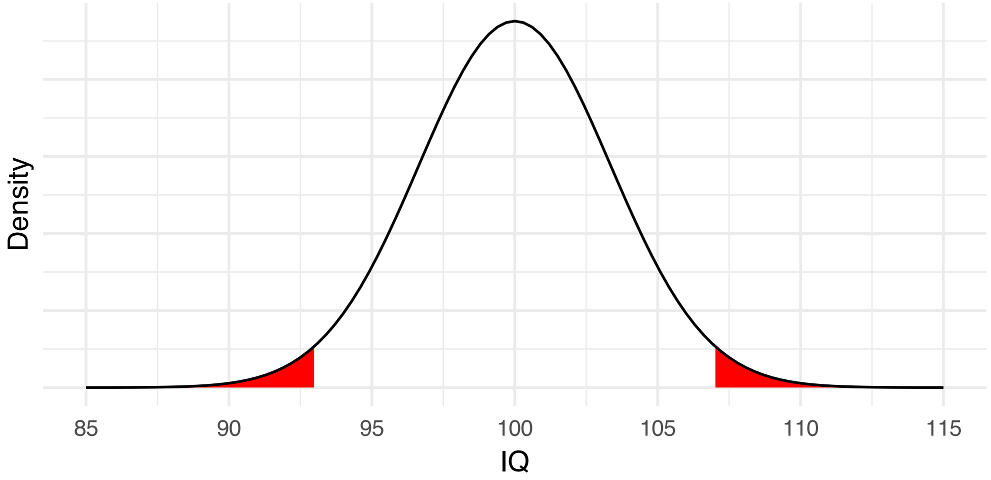
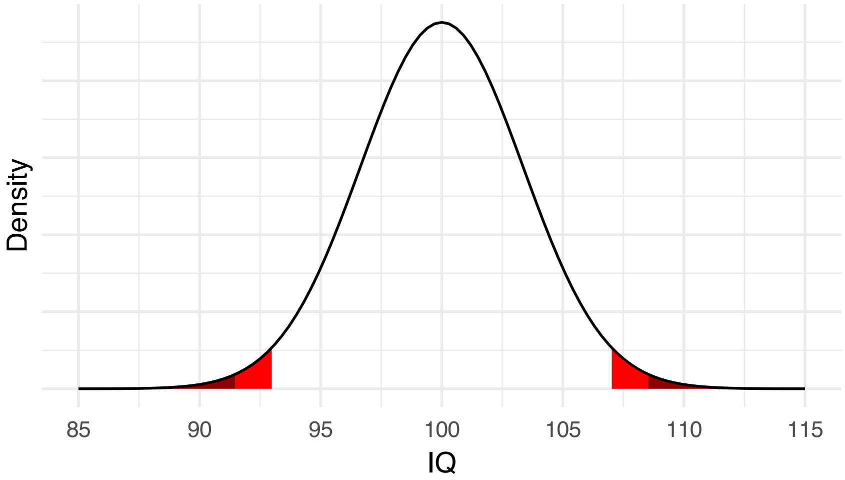


Figure 8: The red regions are those <92.98 and >107.02.

Because the probability of this data is quite low (i.e., we obtained an extreme statistic), given H_0, we often reject the null hypothesis. *Our data is very unlikely given a true null hypothesis, therefore we reject the null.*

7.1.2 Why is $\alpha = .05$?

You have often looked for "$p < .05$" to map onto your *a priori* alpha level ($\alpha = .05$). Why though? The magnitude of the tails is arguable arbitrary, but has been set to a standard of 5% (corresponding $\alpha = .05$) since the early 20th century (Dahiru, 2008), which is a further criticism of NHST. However, we set our α threshold/criteria, and as demonstrated in the previous paragraph, our α level influences the extremeness needed in our data to consider it *statistically significant*. Imagine we wanted to make our criteria more conservative or stringent, such that $\alpha = .01$. Here, our data would need to be *more* extreme to fit the criteria. What do you think would happen to the red shaded region in the above graph? The red regions would shift outward and our critical values would be larger in absolute magnitude. In this case, our graph would be:



The dark red region represents the extreme 1% of the distribution (0.5% per tail). The red region is where the original regions were for $\alpha = .05$.The proportion of samples in the red region is? You guessed it, 4%. Remember, these distributions are for a samples of $n = 20$. The critical regions and subsequent statistics required to be in those regions will differ based on sample size.

So, we set the criterion of α *a priori* and our resultant *p-value* lets us decide if our data are probable (or not) given H_0. If it's lower than our criterion we can reject H_0, suggesting **only** that the population effect is probably not zero. It is often called statistically significant (Spence & Stanley, 2018). Otherwise, if our *p-value* is larger than our criterion, we decide that our data is not that unlikely given a true null, and *fail to reject* that H_0. Here, it's plausible that the true population parameter is 0 (but not confirming that the parameter is 0).

> In short, **p-values** are the probability of getting a set of data or more extreme given the null. We compare this to a criterion cut-off, alpha.

> If our data is very improbable given the null, so much that is is less than our proposed cutoff, we say it is statistically significant.
>
> — Spence & Stanley (2018)

Importantly, rejecting the null can happen when the null is, in fact true. In this case we have committed a type 1 error.

7.1.3 p-value Misconceptions

Many of these misconceptions have been described in detail elsewhere (e.g., Nickerson (2000)). We will revisit only some misconceptions.

1. Odds against chance fallacy

The odds against chance fallacy is when a p-value is interpreted as *the probability that the null hypothesis is true*. For example, someone might conclude that if their $p = .04$, there is a 4% chance that the null hypothesis is true. You have learned that p-values indicate $p(D \mid H_0)$ and that you cannot simply flip the conditional probabilities: $p(D \mid H_0) \neq p(H_0 \mid D)$. The p-value tells you nothing about the probability of a null hypothesis other than it is assumed true. Under a NHST p-value, $p(H_0) = 1.00$.

Cohen (1994) outlines a nice example wherein he compares $p(D \mid H_0)$ to the probability of obtaining a false positive on a test of schizophrenia. Given his hypothetical example, the prevalence of schizophrenia and the sensitivity and specificity of assessment tests for schizophrenia, an unexpected result (a positive test) is more likely to be a false positive than a true negative.

If you want $p(H_0 \mid D)$, you may need another route such as Bayesian Statistics.

2. Odds the alternative is true

In NHST, no likelihoods are attributed to hypotheses. Instead, all p-values are predicated on $p(H_0) = 1.00$. It is assumed that the null is true. Thus, statements such as '$1 - p$ indicates the probability H_A is true' is false.

3. Small p-values indicate large effects

This is not the case. p-values depend on other things, such as sample size, that can lead to statistical significance for minuscule effect sizes. For example, consider the results of the following two statistical results, where both analyses results in the same statistic: $r = .1$:

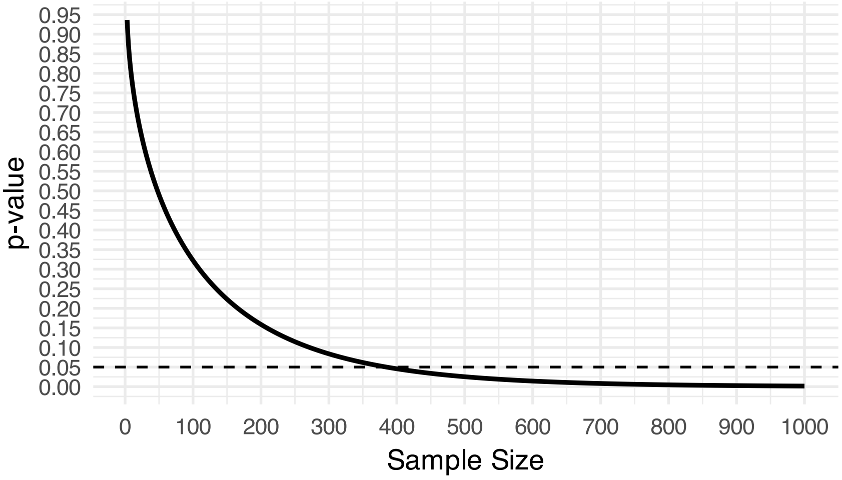
Result from test 1:

r	p
0.1	0.001544

Result from test 2:

r	p
0.1	0.3222

In the above, the two tests have the exact same effect size and correlation statistics, $r = .1$. However, the p-values vary substantially. Thus, effect size does not map directly onto p-value. One major other consideration is sample size. For extremely large sample sizes, a small effect size may have a small p-value (i.e., minor effect is statistically significant). For extremely small sample sizes, a large effect may have a large p-value (i.e., a major effect is not statistically significant). The following represents the relationship between p-value and sample sizes for a correlation of $r = .1$:



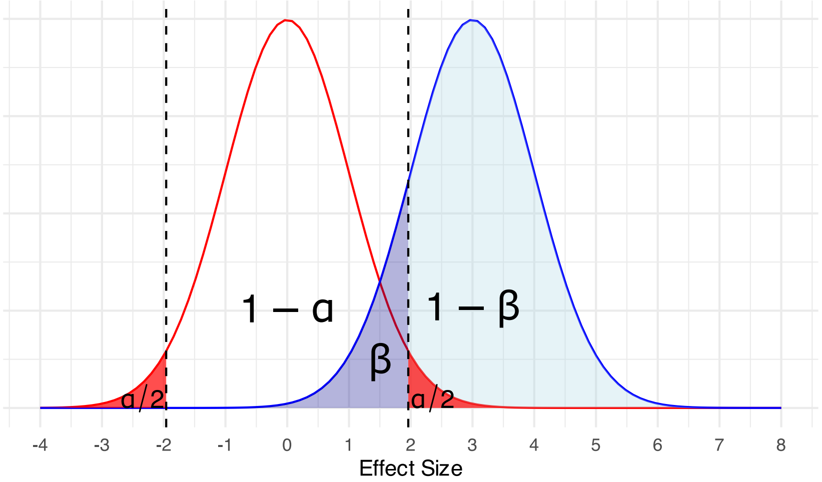
A p-value can be quite different for the exact same correlation coefficient, depending on sample size. More to come. Regardless, it is important to not interpret p-values alone. Effect sizes and confidence intervals are much more informative.

7.2 Power

Whereas p-values rest on the assumption that the null hypothesis is true, the contrary assumption, *that the null hypothesis is false (i.e., an effect exists)*, is important. Many of us would not be doing research if we didn't think a true effect existed (e.g., the drug reduced depression; being connected reduced suicide risk). Indeed, the assumption that an effect exists is also important for determining statistical power. Statistical power is defined as **the probability of correctly rejecting the null hypothesis given an true population effect size and sample size**, or more formally:

> Statistical power is the probability that a study will find *p* < alpha IF an effect of a stated size exists. It's the probability of rejecting H_0 when H_1 is true.
>
> — Cumming & Calin-Jageman (2024)

To help us understand statistical power, let's visualize two hypothetical distributions. One where the null is true (i.e., the null distribution of test statistics), and another where the alternative is true (i.e., the alternative distribution of test statistics).



This plot illustrates the concept of statistical power and has two distribution and two dotted vertical lines. The red curve represents the null hypothesis (H_0) distribution, which assumes no effect, while the blue curve represents the alternative hypothesis (H_1) distribution, which assumes an effect (here, with a mean shifted to 3). The two dotted vertical lines represent the critical values for statistical significance on the null distribution. The light red shaded area under the null distribution beyond the critical value represents the Type I error rate (α), or the probability of rejecting the null hypothesis when it is true. We have already seen how adjust our test criteria can shift those red regions (e.g., $\alpha = .05$ versus $\alpha = .01$).

In contrast, the dark blue shaded area under the alternative distribution to the left of the critical value of the null distribution represents the Type II error rate (β), or the probability of failing to reject the null hypothesis when it is false. The light blue shaded region under the alternative distribution the the left of the critical value reflects the statistical power ($1 - \beta$), which is the probability of correctly rejecting the null hypothesis when it is false, thus detecting a true effect. $1 - \alpha$ represents the probability of correctly failing to reject the null hypothesis when it is true, $1 - \beta$ represents statistical power, $\alpha/2$ marks the Type I error probability in each tail, and β indicates the probability of making a Type II error.

We can conclude from this definition that if your statistical power is low, you will not likely reject H_0 regardless of if there is a population-level effect (i.e., $H_1 \neq 0$). As will be demonstrated, under-powered studies are doomed from the start. Conversely, if a study is overpowered (i.e., extremely large sample size), you can get a *statistically significant* result for what might be a infinitesimal or meaningless effect size.

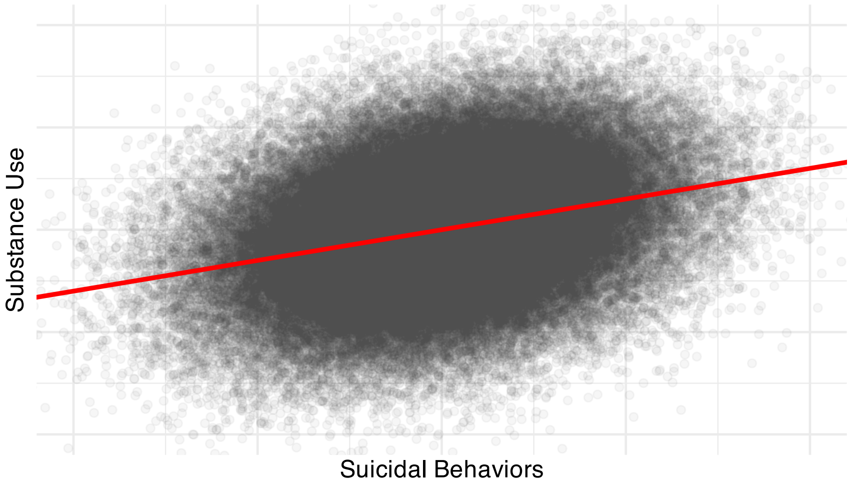
Note

You are likely familiar with the symbol r for a correlation. This is the *sample statistic*. Before proceeding to the next example, please note that you will often see the symbol 'rho', which is represented by the Greek symbol ρ (it's a blend of r and o). This is the population parameter of a correlation. This is the not the same as a p-value, which is represented by the letter p. It may get confusing it you mix up these symbols, so be sure you know the difference between the two.

Consider a researcher who is interested in the association substance use (SU) and suicidal behaviors (SB) in Canadian high school students. They design a study and use the NHST framework. Under this framework, they set the following hypotheses:

- $H_0 : \rho = 0$
- $H_1 : \rho \neq 0$

Let's assume that the *true* correlation between SU and SB is $\rho = .3$; the researchers do not know that this is the true value. The population data plotted on a scatterplot might look like this:



Every faint gray dot represents a student. Their score on SU is on the y-axis, while their score on SB is on the x-axis. The plot looks cloudy toward the center because most student score in the middle on both variables, with fewer scoring on the extremes. Remember, there are $100,000$ dots there! We can see a trend where students with higher SU also have higher, on average, SB. So, it appears that as substance use increases, so do suicidal behaviors. Although we aren't so omniscient in the real world, the population correlation here is $\rho = .3$.

However, before conducting the study, the researcher conducts a power analysis to determine an appropriate sample size required to adequately power their study. They want to have a good probability of rejecting the null, if it were false (here, we know this is the case). To calculate the appropriate sample size, they will require: 1) α criterion level, 2) desired power ($1 - \beta$), and 3) hypothesized effect size. The researcher uses the standard for α criterion and power. Thus, they have: 1) $\alpha = .05$, 2) 1-β = .8, and they estimate the true effect based on the literature to be 3) ρ = .300. Although **we** know the population correlation and that the researcher has accurately estimated the population effect, researchers may not accurately estimate the population effect size. In real-world research we would need to find a good estimate of the effect size, which is typically through reading the literature for effect sizes in similar populations. Power can be calculated many ways such as using the software R or GPower. Using R, the researcher uses the *pwr* package to conduct their power analysis. This package is very useful; you can insert any three of the required four pieces of information (alpha, power, estimated sample size, and sample size) to calculate the missing piece. More details on using `pwr` are below. For our analysis, we get:

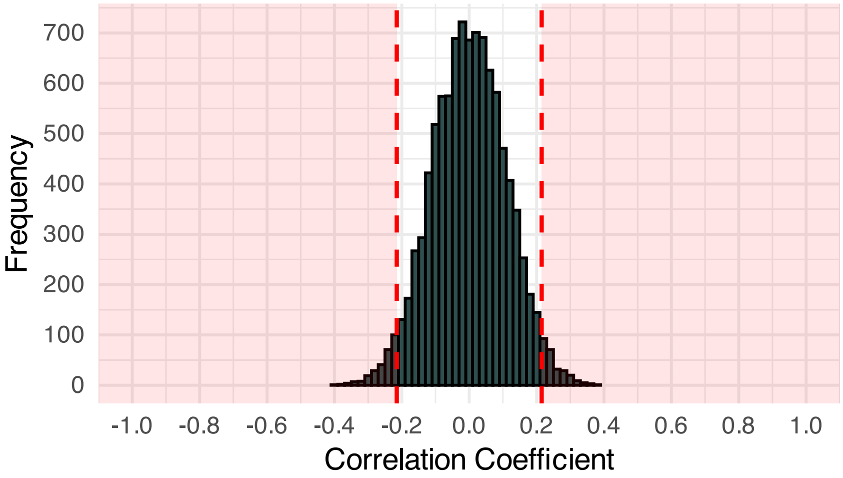
```
     approximate correlation power calculation (arctangh
transformation)

              n = 84.0736
              r = 0.3
      sig.level = 0.05
```

```
          power = 0.8
    alternative = two.sided
```

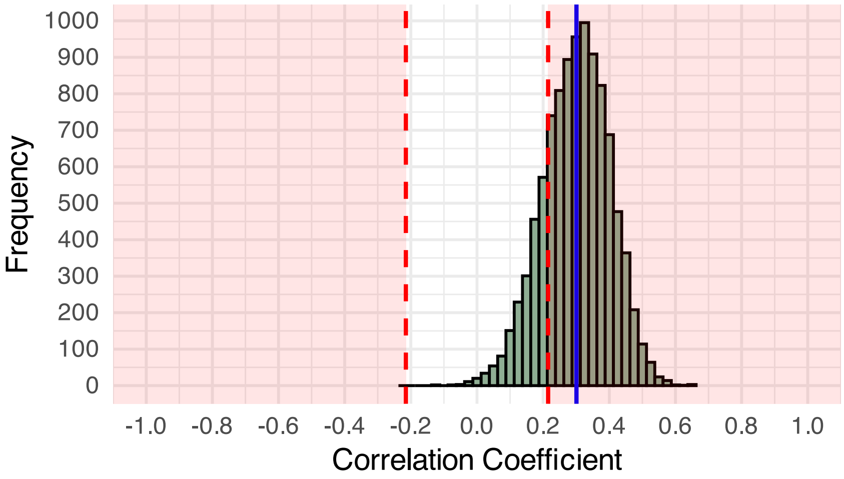
The results of this suggest that we need a sample of about 85 people (rounding up from $n = 84.07$; rounding down would mean power would drop below $.8$) to achieve our desired power ($1 - \beta = 0.8$), using our known population correlation ($\rho = .3$) and $\alpha = .05$.

What does this mean? It means that *if* the true population correlation was $\rho = .3$, then about 80% of all hypothetical studies using a sample size of $n = 85$ drawn from this population will yield $p < \alpha$.

We can create a histogram that plots the results of *many* random studies ($n = 85$)–10,000 to be exact–drawn from our population to determine which meet $p < \alpha$. But first, I will show you what the distribution of test statistics would look like for 10,000 samples of size 85 drawn from the null distribution.



So, let's now replace the null distribution histogram with the alternative distribution histogram. Note that the red regions **will not change**; these are critical values under the null distribution.



This graph represents the distribution of correlation coefficients for each of our random samples, which were drawn from our population. Notice how they form a seemingly normal distribution around our true population correlation coefficient, $\rho = .300$. Although it may seem normal, it is actually skewed because of the bounds of the correlation (i.e., −1 to 1).

The red lines and shaded regions represent correlation coefficients that are beyond the $r = .215$ cut-off, which represent the extreme 5% (2.5% per tail) of sample under a true null and, thus, would result in $p < .05$. The blue line represents the true population correlation coefficient ($\rho = .300$). Just how many samples scored at or beyond the critical correlation value of $r = .215$? Think about it before proceeding.

The results of our correlations suggest that 8016 correlation coefficients were at or beyond the critical value. Do you have any guess what proportion of the total samples that was? Recall that power is the probability that any study will have $p < \alpha$. Approximately eighty percent (0.8016%) of these studies met that criteria: this was our power! Why is it 0.8016 and not 80% exactly? Remember, power refers to a hypothetically infinite number of samples drawn from the population, for any given sample size, effect size, and α. Had we drawn ∞ samples, we would have 80% having $p < \alpha$. Also, recall that we rounded up to 85 people per sample, not 84. Having more people will result in higher power, holding all other things constant. Rounding down would have lower power below $1 - \beta =$

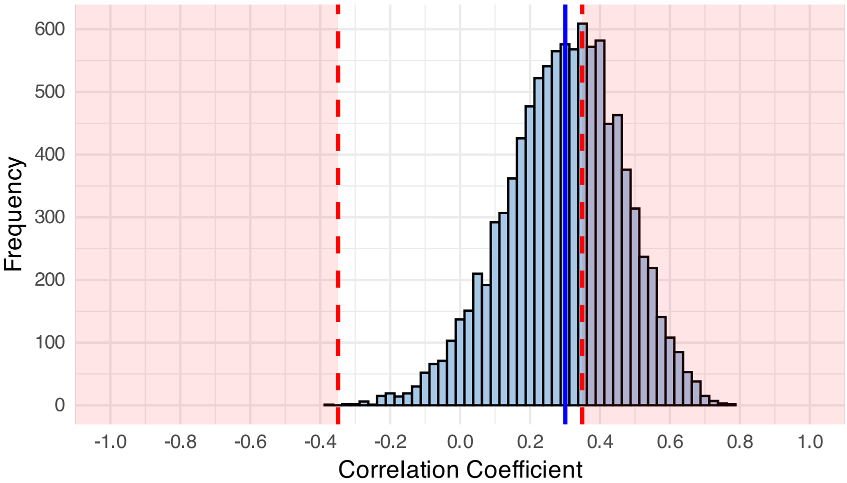
$.8$ and a sample size of $n = 84.07$is impossible. Thus, rounding up is the best course of action.

What if we couldn't recruit 85 participants?

Perhaps we sampled from one high school in a small Canadian city and could only recruit 32 participants. How do you think this would affect our power? First, smaller sample sizes give less precise estimations of population parameters under the alternative distribution (i.e., the histogram above of distribution of sample statistics may be more spread out).

Second, this also influences our critical region of the null distribution; the red regions shift outward. That is, a lower sample size also increase variability in the distribution of sample statistics for the null distribution, as well. This means that the extreme 5% will be further out. Overall, both of these reduce power.

Let's rerun our simulation with 2,000 random samples of $n = 32$ to see how it affects out power. Remember, the true population effect is $\rho = .3$. We can plot the resultant correlation coefficients. Note that the critical correlation coefficient value for $n = 32$ is more narrow for $n = 85$ compared to $n = 32$. The new critical value is $r = 349$ (compared to $r = 215$). The following figure represents the 2,000 samples:



Hopefully, you can see that despite the distribution still centering around the population correlation of $\rho = .3$, it has spread out more. We are less precise in our estimate. Furthermore, the red region (critical r region) is shifted outward due to a smaller n. This would mean that fewer of the samples result in correlation coefficients that fall in the red regions and be *statistically significant* (i.e., power is lower). Would a formal power calculation agree? First, let's find out how many studies resulted in $p < \alpha = .05$ and then do a formal power analysis to determine if they are equal.

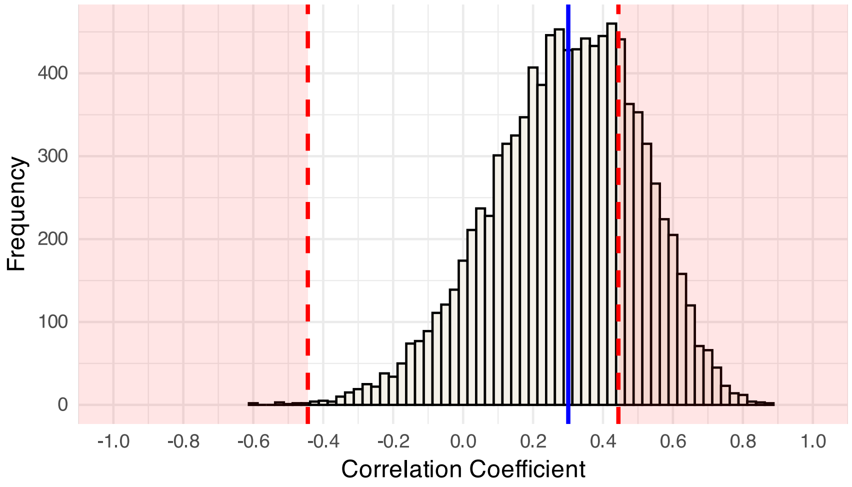
Our results suggest that of the 10,000 simulated studies, 3982studies yielded statistically significant results ($3982/10000 = 39.8\%$). Would our power align?

Formal Power Analysis

```

     approximate correlation power calculation (arctangh
transformation)

              n = 32
              r = 0.3
      sig.level = 0.05
          power = 0.393231
    alternative = two.sided
```

Whoa! Power was $1 - \beta = .3932$ (i.e., 39.32% of samples). Close enough!

Let's take it one step further and assume we could only get 20 participants. This is the plot of the resultant correlation coefficients. Note that the critical correlation coefficient value for $n = 20$ is larger than for $n = 32$ (remember that the null distribution or samples would be spread out more for smaller sample sizes). The new critical value is $r = 444$. The following is adjusted for 20 participants:



Hopefully, you can see that despite the distribution still centering around our population correlation of .3, the red regions are further outward than the previous *two* examples. Again, let's see how many studies resulted in $p < \alpha = .05$ and then do a formal power analysis to determine if they are equal.

A total of 2585 studies yielded statistically significant results ($2585/10000 = 25.8\%$). More, you may even notice that in the previous figure, a statistically significant result occurred in the *opposite direction* (look in the left red region). That is, some results would suggest a negative correlation despite the true parameter being positive. Would a formal power analysis align?

```

     approximate correlation power calculation (arctangh
 transformation)

              n = 20
              r = 0.3
      sig.level = 0.05
          power = 0.255924
    alternative = two.sided
```

So, out of 10,000 random samples from our population, 25.85% *had* $p < \alpha = .05$ and our power analysis suggests that 25.56% *would*. Again, given infinite samples, there would be 25.59237...% with $p < \alpha = .05$.

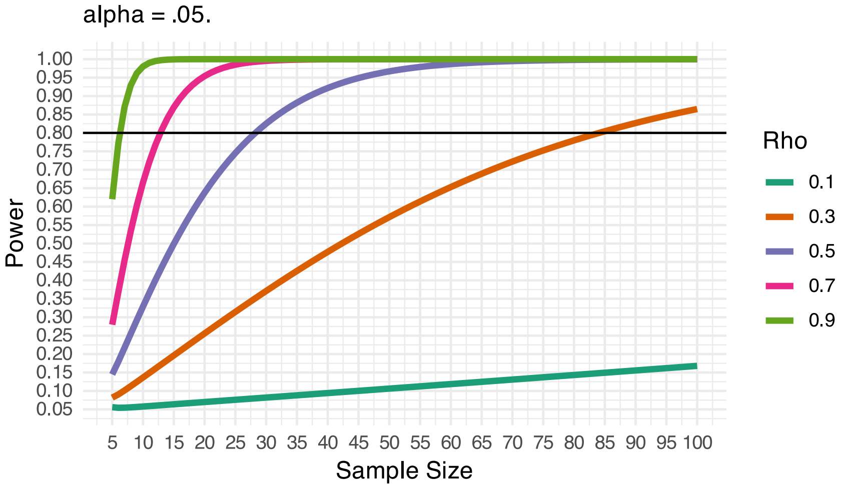
In sum, as sample size decreases and all other things are held constant, our power decreases. If your sample is too small, you will be unlikely to reject H_o even if a *true effect exists*. Thus, it is important to ensure you have an adequately powered study. **Plan ahead. Otherwise, even if a population effect exists, you may not conclude that through NHST'. Or, maybe a non-NHST approach is best.**

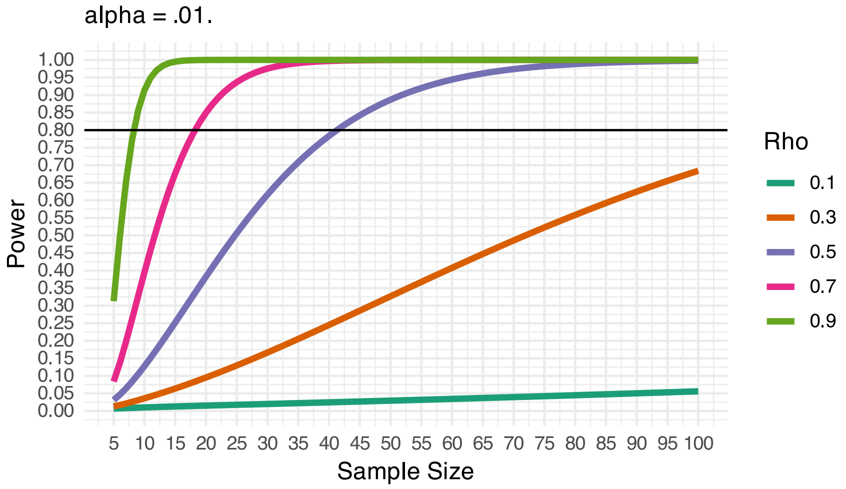
7.2.1 Increasing Power

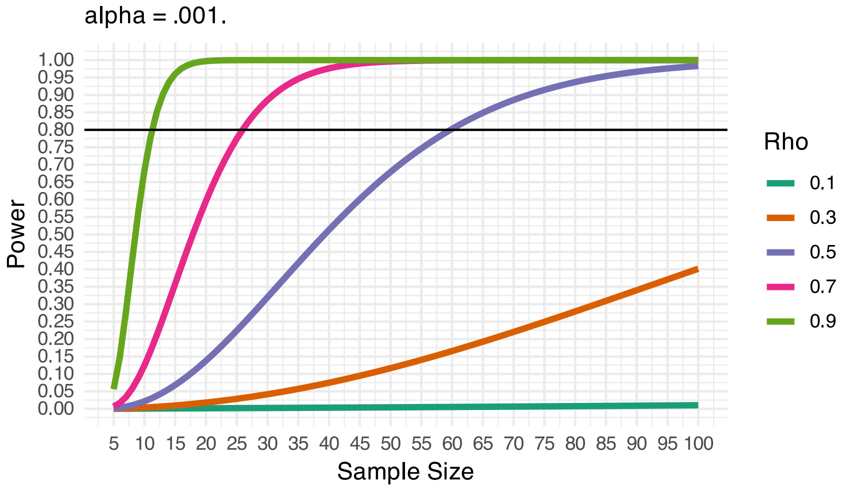
Power is the function of three components: sample size, hypothesized effect size, and $alpha$. Thus, power increases when:

1. The hypothesized effect is larger;
2. You increase your alpha level (i.e., making it less strict; $.1$ versus $.05$).
3. You can collect more data (increase n);

This relationship can be seen in the following graphs:







7.2.1.1 Effect Size

As the true population effect reduces in magnitude, your power is also reduced, given constant n and α. So, if the population correlation between substance abuse and suicidal behaviors was $\rho = .1$, we would require approximately $n = 782$ to achieve power of $1 - \beta = .8$. In other words, 80% of hypothetically infinite number of samples of $n = 782$ would give *statistically significant results*, $p < \alpha$, when $\rho = .1$.

If the population correlation was $\rho = .05$, we would require approximately $n = 3136$ to achieve the same power. This is the basis of the argument that large enough sample sizes result in statistically significant results, $p < \alpha$, that are meaningless (from a practical/clinical/real world perspective). If $\rho = .05$, which is a very small and potentially meaningless effect, large samples will likely detect this effect and result in *statistical significance*. Hypothetically, 80% of the random samples of $n = 3136$ will result in $p < \alpha = .05$ for a population correlation of $\rho = .05$. Despite the *statistical significance*, there isn't much practical or clinical significance. Statistical significance $\neq$ practical significance.

Ways to Estimate Population Effect Size

There are many ways to estimate the population effect size. Here are some common examples, ordered by recommendation:

1. **Existing Meta-Analysis Results**: Meta-analyses compile and combine results from multiple studies to give a more accurate estimate of the population effect size. By aggregating data across various studies, meta-analyses reduce bias and provide a more stable estimate. I **recommendation** using the lower-bound estimate of the confidence interval (CI) from the meta-analysis. This is a more conservative approach that accounts for uncertainty and publication bias, which helps avoid overestimating the true effect size.
2. **Existing Studies with Parameter Estimates**: When no meta-analysis is available, individual studies reporting effect size estimates can be used. These should come from studies with similar designs, populations, or theories. I **recommend** using the lower-bound estimate of the confidence interval presented in the study or halve the reported effect size from a single study to ensure a more conservative estimate. This is because single studies are more prone to sampling variability and biases like publication bias. By halving the effect size or using the lower bound, you reduce the risk of overestimating the true population effect size.
3. **Smallest Meaningful Effect Size Based on Theory**: Theoretical frameworks can suggest the smallest effect size that would be considered meaningful or practically important in your study. This might come from expert consensus or previous theoretical work that identifies thresholds for meaningful differences. I **recommend** that you choose the smallest effect size that is theoretically or practically

significant. This ensures your study has enough power to detect meaningful effects, even if they are small.

4. **General Effect Size Determinations (Small, Medium, Large)**: If no specific guidance is available, general benchmarks for effect sizes can be used:

- **Small**: Cohen's *d* = 0.2, Pearson's *r* = 0.1
- **Medium**: Cohen's *d* = 0.5, Pearson's *r* = 0.3
- **Large**: Cohen's *d* = 0.8, Pearson's *r* = 0.5

I **recommend** choosing the general effect size that aligns best with your theory. If a strong relationship is expected, use a larger effect size; if subtle, use a smaller one. These general benchmarks provide a starting point when no other data is available. They help design studies with realistic expectations for effect sizes, even without prior research.

7.2.1.2 α level

Recall that the alpha level (commonly set at $.05$) represents the threshold we choose to determine statistical significance. Specifically, it is the probability of committing a Type I error–rejecting the null hypothesis when it is actually true. In other words, it defines the "extreme" region of our null distribution, where we would reject the null hypothesis in favor of the alternative hypothesis.

Reducing Alpha and Its Consequences

When we reduce our alpha level (e.g., from $.05$ to $.01$), we are essentially making our criterion for rejecting the null hypothesis more strict. This means that fewer observed outcomes will fall into the "extreme" region, which is now smaller. Visually, this reduction in alpha shrinks the size of the red areas in a power curve (those areas representing where we reject the null hypothesis under the alternative distribution). These areas are associated with detecting true effects, so shrinking them decreases the likelihood of finding a significant result when an effect is present. In the case of correlations, which we have used as example, this results in a larger correlation coefficient threshold that we consider "extreme" enough to indicate significance. Essentially, we are raising the bar for what qualifies as a significant result.

Holding all other factors constant, *reducing the alpha level will decrease the power of the test*. Power refers to the probability of correctly rejecting the null hypothesis when it is false (i.e., detecting a true effect). As the threshold for significance becomes stricter, it becomes harder to detect effects because the criterion for rejection is less lenient. Consequently, more true effects may go undetected, increasing the risk of committing a Type II error (failing to reject a false null hypothesis). Visually, imagine the alternative distribution and the red regions, as shown above. Hold everything constant except those red regions. Slide those outward. Hopefully, you can intuitively understand this means less studies would achieve statistical significance.

Trade-offs in Reducing Alpha

Reducing alpha is often done to minimize the risk of a Type I error, but this comes with trade-offs. While a stricter alpha reduces the probability of falsely rejecting the null hypothesis, it also reduces the power of your test, making it harder to detect true effects. Therefore, when designing a study, it's important to carefully balance the chosen alpha level with the desired power, especially when small effects are being studied.

If alpha is too lenient (e.g., $.10$), the power may increase, but this comes at the expense of a higher likelihood of committing a Type I error, which could undermine the validity of your findings.

Thus, reducing alpha lowers the probability of Type I errors but also decreases the test's power, making it more difficult to detect real effects. This trade-off must be carefully considered during the study design phase, particularly in studies where detecting small but meaningful effects is crucial.

7.2.1.3 Increasing n

There is a direct relationship between n and power. Keeping both the hypothesized effect size and alpha constant, increasing n will increase power.

7.3 Conclusion

Although controversial, NHST is widely used to test hypotheses. We know that NHST begins with an assumption that there is no population-level effect or relationship. Then data is collected from a sample that is believed to be representative of that population. Should the data not align with the null hypothesis, it is taken as evidence against it.

Additionally, there are several concerns and misinterpretations were discussed and allow you to make your own decisions on the frequency and weight of using *p-values* in your research.

7.4 Power in R

This section will focus on conducting power analysis across various statistical platforms. For now I focus on R, but may include other software in future editions.

We will focus on two packages for conducting power analysis: `pwr` and `pwr2`.

7.4.0.1 Correlation

Correlation power analysis has four pieces of information. You need any three to calculate the other:

- `n` is the sample size
- `r` is the population effect, ρ
- `sig.level` is you alpha level
- `power` is power

So, if we wanted to know the required sample size to achieve a power of .8, with a alpha of .05 and hypothesized population correlation of .25:

```
## You simply leave out the piece you want to calculate
pwr.r.test(r = .25,
           power = .8,
           sig.level = .05)
```

```
     approximate correlation power calculation (arctangh
transformation)

              n = 122.447
              r = 0.25
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

7.4.0.2 t-test

With same sized groups we use `pwr.t.test`. We now need to specify the type as one of 'two.sample', 'one.sample', or 'paired' (repeated measures). You can also specify the alternative hypothesis as 'two.sided', 'less', or 'greater'. The function defaults to a two sampled t-test with a two-sided alternative hypothesis. It uses Cohen's d population effect size estimate (in the following example I estimate population effect to be $d = .3$:

```
pwr.t.test(d = .3,
           sig.level = .05,
           power = .8)
```

```
     Two-sample t test power calculation

              n = 175.385
              d = 0.3
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

7.4.0.3 One way ANOVA

One way requires Cohen's F effect size, which is kind of like the average Cohen's d across all conditions. Because it is more common for researchers to use η_2, you may have to convert something reported fro another study. You can convert η_2 to F with the following formula:

$Cohen's\ F = \sqrt{\frac{\eta^2}{1-\eta^2}}$

`pwr.anova.test()` requires the following:

- `k` = number of groups
- `f` = Cohen's F
- `sig.level` is alpha, defaults to .05
- `power` is your desired power

```
pwr::pwr.anova.test(k = 3,
               f = .4,
               power = .8,
               sig.level = .05)
```

```

     Balanced one-way analysis of variance power calculation

              k = 3
              n = 21.1036
              f = 0.4
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

7.4.1 Alternatives for Power Calculation

7.4.1.1 G*Power

You can download here.

7.4.1.2 Simulations

Simulations can be run for typical designs, which you have seen above through our own simulations to demonstrate the general idea of power. For example, we can repeatedly run a t-test on two groups with a specific effect size at the population level. Knowing that Cohen's d is:

$d = \frac{\overline{x}_1 - \overline{x}_2}{s_{pooled}}$

We can use `rnorm()` to specify two groups where the difference in means is equal to Cohen's d and when we keep the SD of both groups to 1.

```r
# One time
sample_size <- 20
cohens_d <- .4 ## our hypothesized effect is .4
t.test(rnorm(sample_size),
       rnorm(sample_size, mean=cohens_d), var.equal = T)
```

```

	Two Sample t-test

data:  rnorm(sample_size) and rnorm(sample_size, mean =
cohens_d)
t = -1.771, df = 38, p-value = 0.0845
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -1.2910722  0.0860217
sample estimates:
mean of x mean of y
 0.116143  0.718668
```

We can use various R capabilities to simulate 10,000 simulations and determine the proportion of studies that conclude that $p < \alpha$.

This returned a data.frame will 10,000 p values from simulations. The results suggest that 2350 samples were statistically significant, indicating 23.5% were statistically significant.

You may be thinking, why do this when I have `pwr.t.test`? Well, the rationale for more complex designs is the same. For more complicated designs, it can be difficult to determine the best power calculation to use (e.g., imagine a 4x4x4x3 ANOVA or a SEM). Sometimes it makes sense to run a simulation.

Simulation of SEM in R, which can help with power analysis.

Companion shiny app regarding statistical power can be found here.

8 Confidence Intervals

If you read the last chapter on NHST, you now recognize that a p-value provides a limited amount of information. Essentially, a p-value can indicate whether the data is probable given a true null hypothesis. Often, we want more than that. Consider Miller and his famous research on short-term memory. Imagine if we said that the results of his famous studies on memory suggest that his data are unlikely if people could not hold anything in short-term memory, $p = .03$. Here, the null would be that people could not hold anything in memory. Wow. Not that informative when compared to the typically-communicated point and precision estimate that short-term memory capacity is around 7 ± 2 pieces of information.

Confidence intervals are intervals that present the most plausible values for a parameter based on a given sample. For example, we might conduct a correlation study and determine that $r = .3, 95\% CI[.20, .34]$, indicating that the best guess for the population parameter is .3, while anywhere from $.20$ (the lower limit) to $.34$ (the upper limit) is plausible. Typically, when working with CIs, we have:

1. Point estimate: a single number that indicates the resulting test statistic. For any given sample, it is the most plausible value of the population parameter.

2. Confidence interval: "an interval or range of plausible values for the population parameter of interest. A CI is a set of parameter values that are reasonably consistent with the sample data we have observed." (Cumming & Finch, 2001). The range has two numbers: the lower and upper limit. The lower limit represents the smallest plausible value or the population parameter and the upper limit represents the largest plausible value.

3. Confidence level: the degree of certainty we wish to put around a group of CIs. Typically set to 95% to correspond with $\alpha = .05$ in NHST. That is, typically, the CI percentage and α sum to 1.

8.1 Benefits of Using CIs

There are numerous potential benefits of using CIs, as Cumming and Finch (2001) explain. Let's explore these in detail.

First:

> They give point and interval information that is accessible and comprehensible and so, as the examples above illustrate, they support substantive understanding and interpretation.
>
> — Cumming & Finch (2001)

Specifically, CIs give point *and* interval information that is accessible and comprehensible, which supports understanding and interpretation (Cumming & Finch, 2001). Where p-values tell us little, CIs tell a lot more. While p-values are generally misinterpreted or difficult to interpret, particularly for novice researchers and the general public, CIs are more easily interpretable.

For example, consider this statement. "**The relationship between depression and anxiety is unlikely given a true null hypothesis**, $p = .021$". That's much more difficult to tease apart and understand than "**The relationship between depression and anxiety is plausibly range from a small medium positive relationship**, $r = .22$, $95\%CI[.15, .33]$".

Or consider another example. I want to know how tall you are. You say: given the 10 measurements I've taken over the last week, I'm probably not 0cm tall, $p < .001$. Compare that to: I'm probably 162cm tall, but am plausibly between 160-165cm tall. The former is what a p-value can tell us, the latter what a CIs seeks to tell us.

Second:

> There is a direct link between CIs and familiar null hypothesis significance testing (NHST): Noting that an interval excludes a value is equivalent to rejecting a hypothesis that asserts that value as true —at a significance level related to C. A CI may be regarded as the set of hypothetical population values consistent, in this sense, with the data.
>
> — Cumming & Finch (2001)

What is meant here, is that if a CI excludes a null hypothesis value at whatever confidence level (e.g., 95%), it would also reject that value through NHST and the same α level. For example, if you concluded that the mean difference between two groups is $\overline{x}_{diff} = .3$, $95\% CI[.10, .38]$, then a standard NHST would results in $p < .05$ for the same test. If the CIs excludes 0, you would get a statistically significant result. The CI tells you just as much, and much more, than a p-value.

Third:

> CIs are useful in the cumulation of evidence over experiments: They support meta-analysis and meta-analytic thinking focused on estimation. This feature of CIs has been little explored or exploited in the social sciences but is in our view crucial and deserving of much thought and development.
>
> — Cumming & Finch (2001)

Specifically, CIs propose that their values will inform us of population parameters over the long run of many studies of similarly conducted tests. Meta-analysis can facilitate this by pooling multiple studies into one strong evidence base. Because a 95%CI indicates that 95% of CIs over the long run will contain the true population parameter, meta-analysis can inform just where that parameter may be. Forrest plots are helpful in this regard.

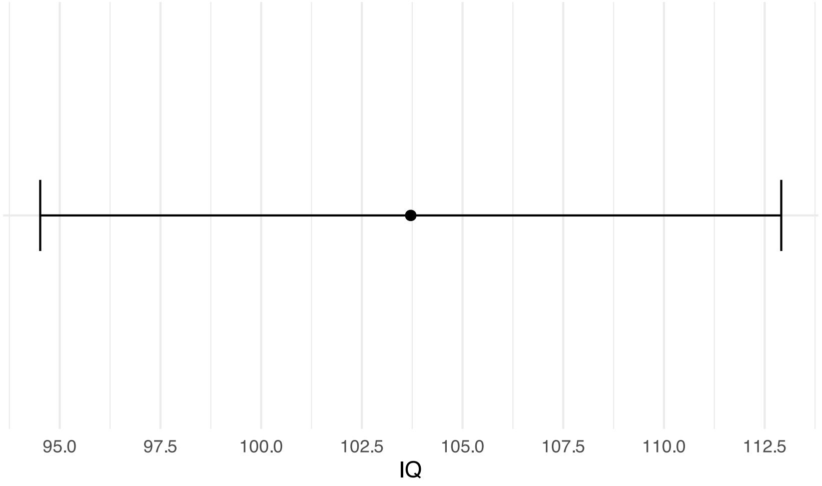
Last:

> CIs give information about precision. They can be estimated before conducting an experiment and the width used to guide the choice of design and sample size. After the experiment, they give information about precision that may be more useful and accessible than a statistical power value.
>
> — Cumming & Finch (2001)

8.2 CI Basics

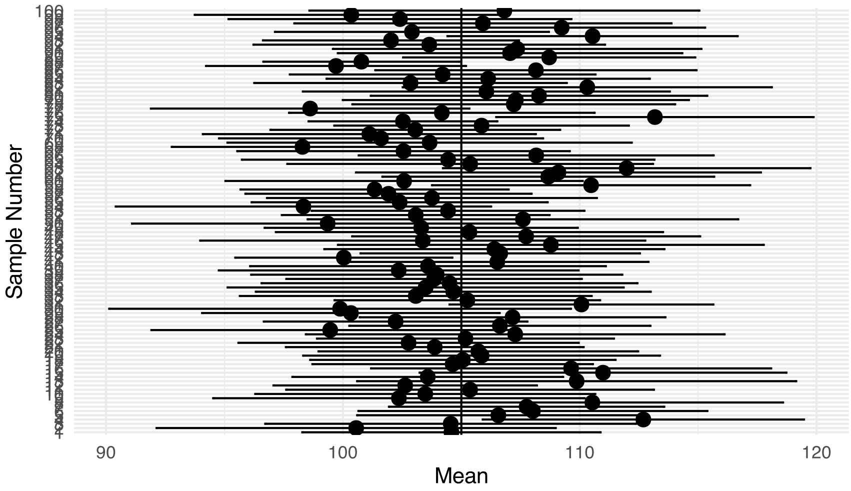
Given we know that CIs are made up of point estimates and intervals, we can visualize them to help us understand what they are. Imagine a population (e.g., all Grenfell students). We want to sample from the population and infer from the sample statistics about the population parameter. Our hypothetical construct of interest is IQ. Imagine that we know the population parameters: $\overline{X}_{IQ} = 105$ and $\sigma_{IQ} = 15$.

If we sample from the population we can calculate a CI (let's calculate the 95% CI, but we could for any number). In our sample we get a mean of 103.71, with a 95% CI of $[94.51, 112.91]$. Let's visualize it:

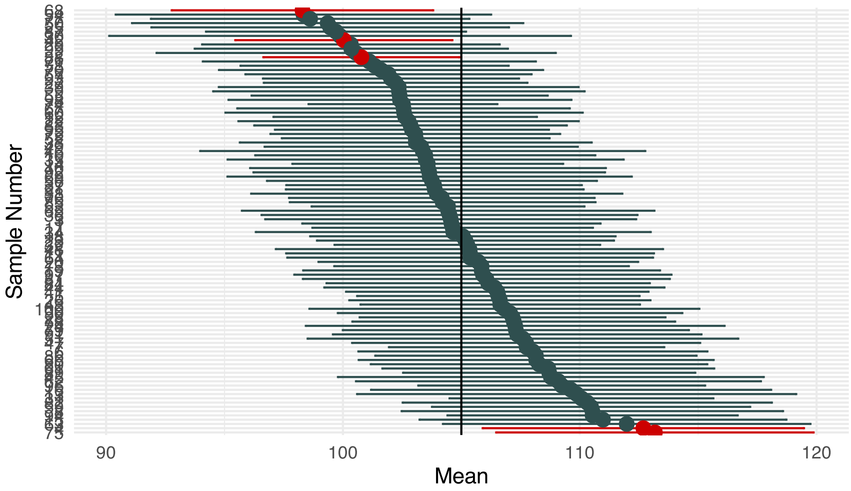


In the above we have the mean, which is represented as a dot, and the confidence interval, which is the space between the two vertical lines.

Imagine that instead of taking one sample, we re-ran the study over, and over, and over.... 100 times–each time with a slightly different sample! If we plot the CIs for each of these studies, we get the following. For convenience, I'll put a solid vertical line to represent the known population mean:



You may notice that some of the CIs contain the true population parameter and some do not. Given what you know about CIs, you know that that those CIs that do not contain the true parameter would reject it in a NHST test; those tests would say that the population parameter is probably not 105. To help visualize, let's sort them by their mean:



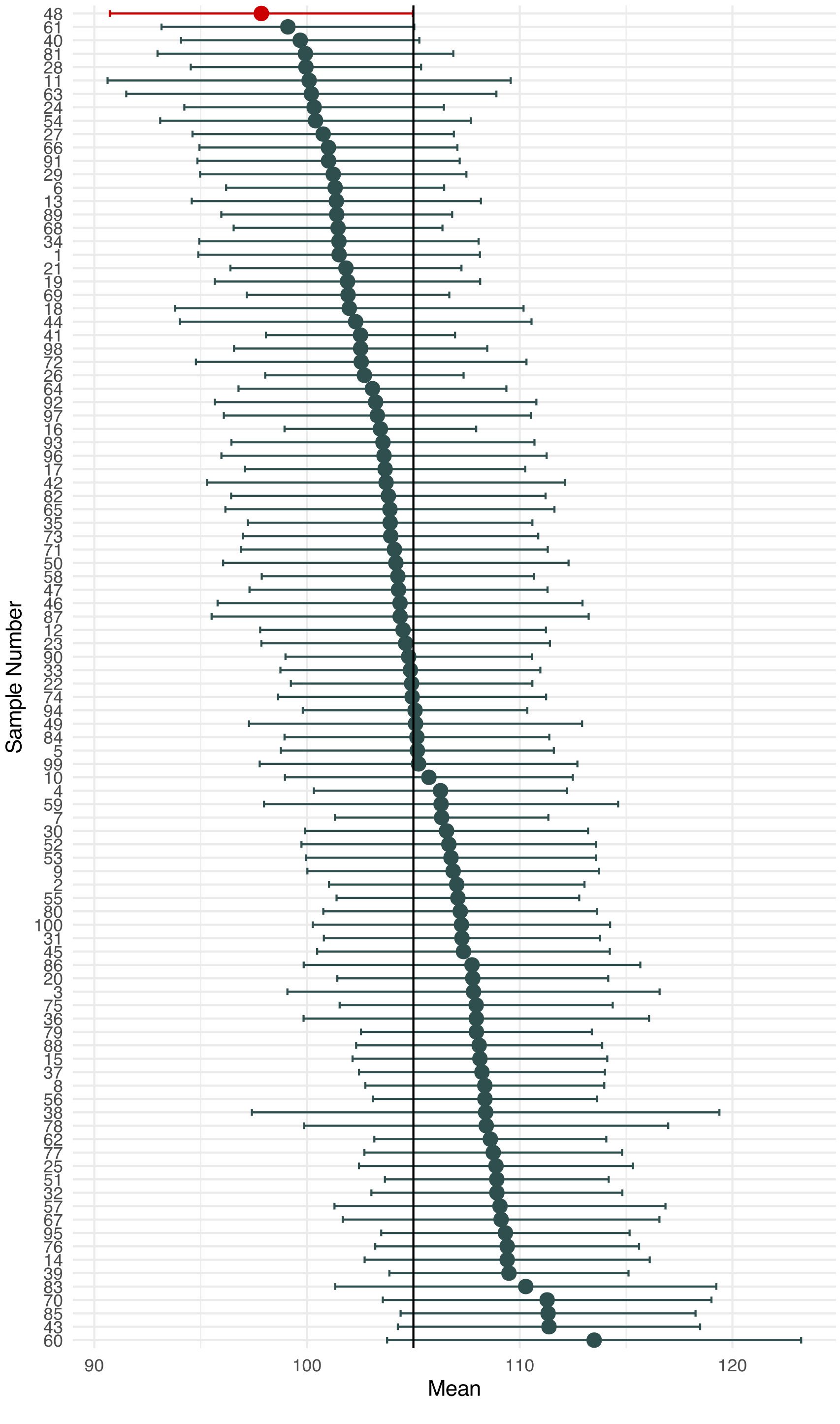
Think about it

Stop and think before moving on. How many of the CIs contain the true population parameter?

$95/100 = 95\%$ of the confidence intervals contain the true parameter. This is an easy way to interpret and understand the meaning of a CI.

Importantly, a CI tells you *nothing about the probability of any single CI* containing the true population parameter. It tells you the most plausible values of the parameter, given the obtained sample. Also, the % of CIs containing the parameter will equal the confidence level over a hypothetically infinite number of samples. In our example, 95% of the CIs contained the true population parameter.

We can use the exact same data to calculate a 99% CI instead of a 95% CI for each sample. Notice what happens when we use the new CIs:



By adjusting the CIs on the data, 99/100 (i.e., 99%) contain the parameter.

Think about it

If we change our CIs from 99% versus 95% (using the same data), what should happen to the width of CIs? Use the graphs above to help you visualize.

Picture 100 samples. If we calculate 95% CIs around the mean, then approximately 95% will overlap the population mean. If we calculate 99% CIs, then approximately 99% will overlap the population mean. This means that the 99% CIs *must* be longer than the 95% CIs. More of them will over lap the mean.

Conversely, let's imagine we calculate the 50% CIs. Now approximately 50% will overlap the population mean. Thus, they will be shorter/narrower.

Let's work out how to calculate CIs for a mean. While we will not learn how to calculate them for more advanced statistics, the rationale is typically the same: statistic $\pm$ margin of error.

8.3 CI of a Mean

The CI of a mean can be calculated using the following formula.

$$\overline{x} \pm t_{(n-1,\frac{\alpha}{2})}\left(\frac{s}{\sqrt{n}}\right)$$

where:

- $t_{(n-1,\frac{\alpha}{2})}$ is the critical t value for $n-1$ df and
- α is your test criteria.

Let's do a concrete example. Imagine the following data: $8, 5, 9, 5, 4$. Let's calculate a CI for the population mean, given the data we have collected. We need to calculate a few things. We need the mean ($\bar{x}$), n, SD, standard error (SE), and critical t.

In this example, we get: $\overline{x} = 6.2$, $SD = 2.1679$, and $SE = \frac{SD}{\sqrt{n}} = \frac{2.1679}{\sqrt{5}} = 0.9694$.

Next we need critical t. You can look this up in any t-distribution table for $\frac{\alpha}{2}$ and $n-1$ degrees of freedom. For us, t_{crit} = 2.776445.

Our resulting CI is:

$$\overline{x} \pm t_{(n-1,\frac{\alpha}{2})}\left(\frac{s}{\sqrt{n}}\right)$$

$$= 6.2 \pm 2.776(0.9694)$$

$$= 6.2 \pm 2.69105$$

$$= [3.51, 8.89]$$

Thus, the population parameter is plausibly anywhere between 3.51 and 8.89. Our best guess is 6.2.

8.4 Other CIs

Usually your statistical program will calculate CIs for your various sample statistics. For the purposes of courses at Grenfell, I do not require you to know the formal calculations for various confidence intervals. **You should, however, know how to interpret them.**

8.5 Conclusion

CIs represent the most plausible values of a population parameter. Due to the misinterpretation of the work 'confidence', some propose that they should instead be termed plausibility intervals. For our purposes, they provider a best guess point estimate and set of plausible values for the true population parameter. Indeed, we could write our results as 'the plausible values for population parameter are $95\%CI = [LL, UL]$. However, given we never know the true population parameter and whether our CI contains it, there is the likelihood of committing an error. Regardless, CIs provide much more information than p-values and should be the focal point of a results section rather than simple statistical significance.

8.6 Practice Questions

1. Calculate the mean and 95% CI for the following list of numbers:

10, 3, 4, 3, 7

2. Interpret the following confidence intervals:

a. $r = .3$, $90\% CI[.13, .42]$
b. mean difference between group 1 and group 2 $= 3.2$, $95\% CI[-1.2, 6.3]$

8.7 Answers

Mean	SD	N	SE	LowerCI	UpperCI
5.4	3.05	5	1.364	1.613	9.187

a. This means that the correlation between two variables is best estimated to be 0.3. However, anywhere between 0.13 and 0.42 are plausible values. Since this CI does not include 0, the correlation is likely statistically significant.

b. This indicates that the estimated mean difference is 3.2. However, the mean difference is plausibly between −1.2 to 6.3. Since the CI includes 0, this suggests that the difference between the two groups may not be statistically significant. The data is not conclusive about whether there is a true difference between the groups.

9 z-test

This chapter will cover the z-test. Although more details follow, in short, a z-test is a statistical method used to determine if there is a significant difference between a sample mean and a known population mean assuming the population variance is known. It calculates a z-score, which measures how many standard deviations the sample mean is from the population mean. By comparing this z-score to a critical value from the standard normal distribution, researchers can determine whether the observed difference is statistically significant. These tests are commonly used in large-sample studies where population parameters are available.

9.1 Betcha' can't eat just one!

Imagine we wanted to model the average weight of a bag of Lay's Potato Chips. Let's use our scientific method, as discussed in an earlier chapter, to conduct some science.

Specifically, you have a theory that, *despite being listed as 200g, the large bag of chips actually weighs less because the company cuts corners to save money and is, thus, dishonest about the weight*. So you hypothesize that **Lays chips that are listed as 200g do not weigh 200g**. Let's work through the steps.

9.1.1 1. Generating hypotheses:

Our conceptual hypothesis that "Lays chips that are listed as 200g do not weigh 200g" can be translated into a statistical hypothesis, which is represented as the null and alternative hypotheses:

$$H_0 : \mu_{lays} = 200g$$

$$H_1 : \mu_{lays} < 200g$$

You email Lays and they respond, indicating that their chips, on average, weigh 200g, but have some variability. Specifically, they say the standard deviation of weight of the chips is 6g. You aren't satisfied with that response and continue with your research.

It is impossible for you to weigh every single produced 200g bag of Lays chips, so you decide to take a sample. You decide to use NHST to test the weights of bags and use a $\alpha = .05$ criterion.

9.1.2 2. Designing a study

Before conducting the study, you write up your proposed design and submit it to the Grenfell research ethics board. You plan a study wherein you will drive around Corner Brook to three popular stores: Sobey's, Myles's, and Coleman's. You will buy two 200g bags of Lays at each location, resulting in a total sample size of six (n = 6). Once you have the chips, you will bring them to your home and weigh them on a professionally calibrated weight scale. You decide to pour the chips out of the bag, as Lay's communicated that the weight indicates the chips put in a bag and does not include the bag.

The Grenfell Campus research ethics board foresee no risks and allow you to complete the study.

9.1.3 3. Collecting data

You follow through with your research plan. You get the following data:

Bag	Store	Weight
1	Wilsons	201.1
2	Wilsons	191.7
3	Myles	194.6
4	Myles	191.3
5	Colemans	189.9

Bag	Store	Weight
6	Colemans	195

9.1.4 4. Analyzing data

We know the population mean ($\mu = 200$) and standard deviation ($\sigma = 6$). Let's calculate the mean and standard deviation of our sample.

Mean	SD
193.9	4.017

9.1.4.1 Mean, Standard Deviation, and Variance

Remember, the mean is a measure of central tendency and is (here, population):

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

and population standard deviation is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N}}$$

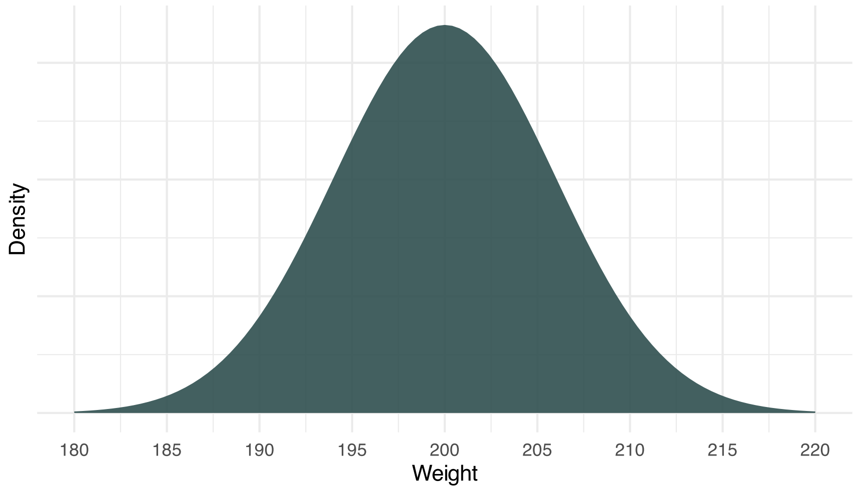
and population variance is:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N}$$

You may remember that the population and sample standard deviation differ in the denominator. The sample SD and variance have $n - 1$ as a denominator versus the population's n. Revisit Bessel's Correction for a refresher as to why.

9.1.4.2 Visualizing Chips

Given that Lays has communicated the population parameters, we can calculate how likely our sample is. First, let's visualize the distribution of chips with a $\mu = 200$ and $\sigma = 6$:



When we repeatedly take samples from a population and calculate the mean of each sample, the distribution of these sample means forms what's known as the **sampling distribution of the sample mean**.

Importantly, according to the Central Limit Theorem (CLT), no matter what the original distribution of the data looks like (as long as it has a finite mean and variance), the sampling distribution of the sample mean will tend to be normally distributed as the number of samples increases. This is true even if the population distribution is not normal.

The **mean** of the sampling distribution of the sample mean (often denoted $\mu_{\bar{x}}$) will be the same as the mean of the original population. So if the average weight of chip bags in the population is, say, 200 grams, then the mean of all those sample means will also be 200 grams.

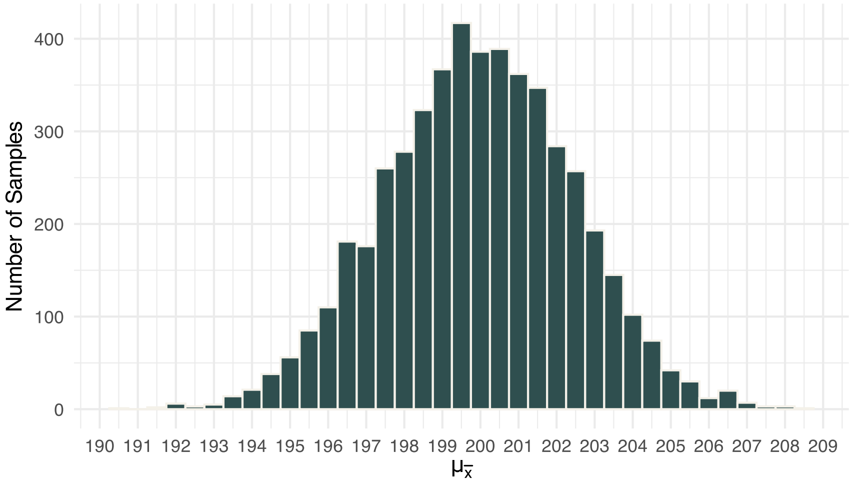
The **standard deviation** of the sampling distribution (called the **standard error**) is smaller than the population standard deviation. It is calculated as:

$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}}$$

Where σ is the population standard deviation and n is the sample size. This means that the variability (spread) of the sample means is less than the variability of individual data points in the population. Additionally, as the sample size increases, the standard error approaches 0. That is, a larger sample size provides a more precise estimate of the population mean.

If you plot the means of a lot of samples of 6 bags of chips, you would get a **normal-shaped bell curve** (assuming you've taken a large number of samples), with the peak centered at the population mean. The spread (or width) of this bell curve depends on the **sample size** you drew and the **population standard deviation**. The more samples you take, the smoother and more normal the distribution of sample means will look.

In our example, if we take a sample of six bags, calculate the average weight of those six bags, and repeat this process many times, we would end up with a collection of sample means.



As you can see from the figure, when you collect the mean of six random bags of chips many times, they form another normal distribution.

Definition

The **standard error** is the standard deviation of sample means. A large standard error indicates high variability between the means of different samples. Therefore, your sample may not be a good representation of the true population mean. This is not good.

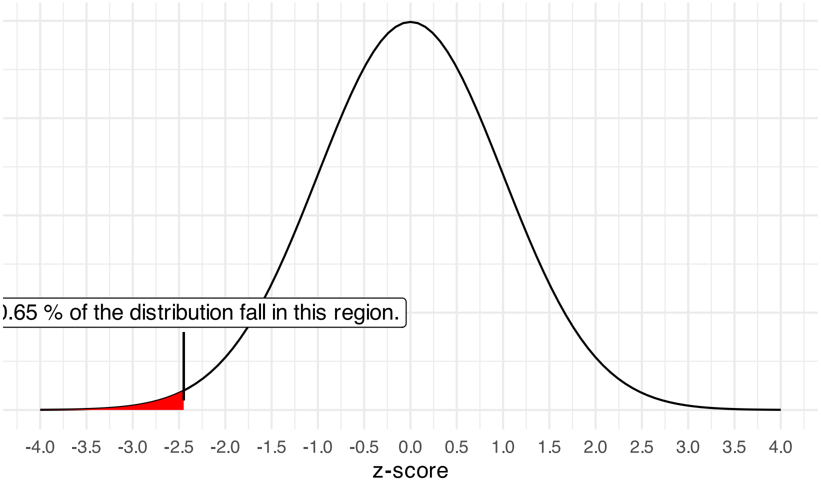
A small standard error indicates low variability between the means of different samples. Therefore, our sample mean is likely reflective of the population mean. This is good.

Should the above distribution of sample means truly follow a normal distribution, then we should be able to calculate how likely our sample of six bag of chips is! We can fill in the what we know, according to Lays: $\mu_{\bar{x}} = \mu = 200$ and; $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{6}} = 2.4495$.

People have used calculus and other math to help us identify the proportions of sample means in various tails, or the proportions associated with various scores. Using this math, the probability of getting our sample mean of 193.92 can be converted into a z-score:

$$z = \frac{x - \bar{x}}{\sigma_{\bar{x}}} = \frac{193.9185 - 200}{2.4495} = -2.482752$$

Additionally, the probability of getting a z-score that low is 0.0065. Recall that the normal distribution has some unique properties and we can find out the proportion of scores that fall in the tails. While you may have used table like the previous link in past courses, computers can easily determine the exact quantity (e.g., `pnorm()` in R).



Hypothetically, out of 1000 samples of six bags of chips drawn from the distribution of $\mu = 200$ and $\sigma = 6$, we would get a score as low as our sample mean or lower less than 7 times (7/1000 is close to 0.65%). Is this unprobably enough? This is determined by our apriori criteria; for us it was $\alpha = .05$. Thus, if the data are at extreme 5% of the distribution, we would conclude that it is unlikely given a true null, which is $p < .05$. Our data was much less likely: $p = .0065$. Our sample is very unlikely if Lays is telling the truth!

You just did a z-test. Let's run it in R to ensure we get the same numbers! Recall our data:

Bag	Store	Weight
1	Wilsons	201.1
2	Wilsons	191.7
3	Myles	194.6
4	Myles	191.3
5	Colemans	189.9
6	Colemans	195

The assumption is that we have randomly sampled. That is:

$$H_0 : \mu = 200$$

and

$$H_1 : \mu < 200$$

What would you conclude from the following output?

```

	One-sample z-Test

data:  chip_data$Weight
z = -2.483, p-value = 0.00652
alternative hypothesis: true mean is less than 200
95 percent confidence interval:
      NA 197.948
sample estimates:
mean of x 
  193.918 
```

9.1.5 5. Write your results/conclusions

When interpreting this, we can say that:

The sample of six bags of chips had a mean weight of $\bar{x} = 193.91$ ($SD = 4.02$). A z-test indicated that the sample mean was unlikely given a true null hypothesis that $\mu = 200$ ($\sigma = 6$), $z = -2.48, p = .0065$.

9.2 Conclusion

A z-test is a statistical method used to determine if there is a significant difference between a sample mean and a known population mean assuming the population variance is known. It calculates a z-score, which measures how many standard deviations the sample mean is from the population mean. By comparing this z-score to a critical value from the standard normal distribution, researchers can determine whether the observed difference is statistically significant. These tests are commonly used in large-sample studies where population parameters are available.

9.3 Practice Questions

1. Calculate the mean, sample SD, and sample variance of the following two variables, *x* and *y*:

x	y
6	10
3	6
12	8
3	9
6	4

2. What's the difference is the probability of sampling a single bag of chips weighting 190g in a sample versus getting a mean weight of 190g for 10 bags of chips? Why are they different? How to the distributions differ?
3. What happens to the SD of the distribution of sample means as the sample size increases? Imagine drawing 100,000 bags of chips.
4. Suppose a university claims that the average score on a standardized test for psychology students is 75 with a standard deviation of 10. We collect a sample of 15 students' test scores to see if the average score differs from 75. The following are the scores:

Person	Score
1	78
2	74
3	61
4	75
5	57
6	80
7	82
8	66
9	74
10	90

Person	Score
11	81
12	64
13	52
14	45
15	87

a. What is the sample mean?
b. Perform a one-sample z-test: Is the sample mean significantly different from the claimed population mean of 75? Use a population standard deviation of 10 and a significance level of 0.05.
c. What is the z-score for this test?
d. What is the p-value for the z-test? Does it allow us to reject the null hypothesis?

9.4 Answers

1.

Mean_x	SD_x	Mean_y	SD_y
6	3.674	7.4	2.408

2.

- **Single Bag**: The probability of randomly selecting one bag weighing 190g is based on the overall distribution of weights. For example, if most bags weigh around 200g, the chance of getting exactly 190g might be moderate.
- **Mean of 10 Bags**: The probability of getting an average weight of 190g from 10 bags is much lower because averaging reduces variability. If the average weight of a sample is far from the population mean, it's less likely to occur.

3. As the sample size increases, the **standard deviation of the sample means** (called the standard error) gets smaller. This means the aver-

age weight of the sample gets closer to the true mean weight of the population.

Example with 100,000 Bags: If we take 100,000 bags, the standard error will be very small. This indicates that the average weight of these bags will be very close to the true population mean, making it a precise estimate.

4. z-test results

```

	One-sample z-Test

data:  test_scores$Score
z = -1.523, p-value = 0.128
alternative hypothesis: true mean is not equal to 75
95 percent confidence interval:
 66.0061 76.1273
sample estimates:
mean of x 
  71.0667 
```

APA Write-Up A z-test was conducted to determine if the average score of psychology students on a standardized test significantly differed from the university's claimed average score of 75. Our results are not unlikely given a true null hypothesis (mean difference = 0), $z = -0.41, p = .681$.

Explanation Since the p-value (.141) was greater than the significance level of .05, the null hypothesis was not rejected. These results suggest that there is insufficient evidence to conclude that the average score of psychology students significantly differs from the claimed mean score of 75.

10 Independent t-test

This chapter will cover the independent t-test. Although more details follow, in short, an independent t-test is a statistical method used to determine if there is a significant difference between the means of two independent groups. Unlike the z-test, which has one group's mean, the independent t-test is used to compare the means of two groups. Additionally, the z-test is used when the population variance is known, where the independent samples t-test is used when the population variances are unknown and estimated from the sample data. It calculates a t-statistic, which measures how many standard deviations the difference between the two sample means is from the expected difference (usually zero). By comparing this t-statistic to a critical value from the t-distribution, researchers can determine whether the observed difference is statistically significant (i.e., unlikely given a true null of *no difference*). Independent t-tests are commonly used in psychological research.

10.1 Some Additional Details

The t-test is used to compare two groups, which is considered one categorical (independent) variable. As examples:

- The variable gender could have two potential groups (male versus female);
- The variable time could have two potential groups (Time 1 versus time 2; Age 8 versus age 12)
- The variables treatment could have two potential groups (medication versus therapy).

Additionally, we measure some continuous outcome (dependent) variable. In most psychological research, and experiments in particular, we aim to both *randomly sample* participants from the population and *randomly assign* them to different groups. This process helps ensure that the groups are comparable, so any differences in groups means on the outcome variable can be attributed to the grouping variable. In other words, any changes in the dependent variable (DV) are assumed to result from the independent variable (IV).

Definition Refresher

In a **random sample**, a subset of individuals chosen from a larger population in such a way that every member of the population has an equal chance of being selected. This ensures that the sample accurately represents the population, which reduces bias and ensures results are generalizable.

Random assignment is the process of assigning participants in an experiment to different groups (e.g., treatment vs. control) using a random method. Random assignment ensures that each participant has an equal chance of being placed in any group, which should balance out any other differences between group members. This helps control for confounding variables, which allows researchers to make causal inferences.

If you recall, the null hypothesis typically purports that there is no difference/association. Thus, imagine we are comparing the means two groups: μ_1 and μ_2. The null hypothesis states:

$$H_0 : \mu_1 = \mu_2$$

or, in another form:

$$H_0 : \mu_1 - \mu_2 = 0$$

and the alternative hypothesis states (for a two-sided test)

$$H_1 : \mu_1 \neq \mu_2$$

or, in another form that aligns with above:

$$H_1 : \mu_1 - \mu 2 \neq 0$$

In the last chapter we learned about the z-test, which carries a somewhat unrealistic assumptions: that we know some population's variance. Most times we do not know the population standard deviation or variance, so we must estimate it using our sample data. Furthermore, in last chapter we simulated the distribution of a sample's means over and over to demonstrate central limit theorem and show how standard error is a function of a sample's standard deviation and sample size ($SE = \frac{SD}{\sqrt{n}}$). In a t-test, we are not interested in one mean, but two. We also have two potential standard deviations. Thus, our analysis is slightly different.

10.2 Betcha' can't eat just two...?

Imagine we wanted to model the average weight of bags of two different flavors of Lay's Potato Chips. Let's use our scientific method, as discussed in an earlier chapter, to conduct some science.

Specifically, you have a new theory: **Lays purposely puts fewer chips in a bag of Ketchup Chips than Regular Chips because the seasoning in Ketchup Chips costs more to produce**. Based on this, you hypothesize that **200g bags of Ketchup Chips weigh less than 200g bags of Regular Chips**.

10.2.1 1. Generating hypotheses:

Your hypothesis can be translated into a statistical hypothesis, represented as the null and alternative hypotheses:

$$H_0 : \mu_{ketchup} = \mu_{regular}$$

This states that the average weight of Ketchup Chips is equal to that of Regular Chips. The alternative hypothesis is that

$$H_1 : \mu_{ketchup} < \mu_{regular}$$

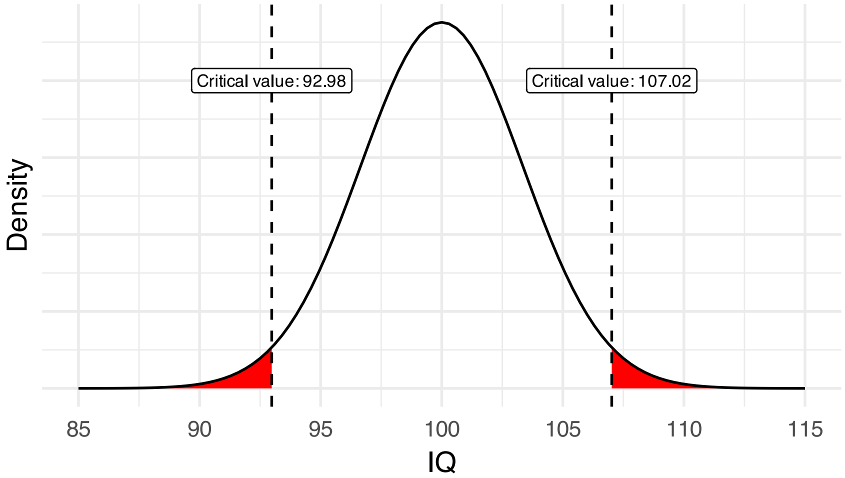
This states that the average weight of Ketchup Chips is less than that of Regular Chips.

You email Lays, and they respond similarly to before: “All our bags weigh, on average, 200g regardless of flavor! Also, we don’t know the standard deviations of ALL the flavors…measure them yourself! And, oh…stop emailing us!”

One and Two-sided Tests

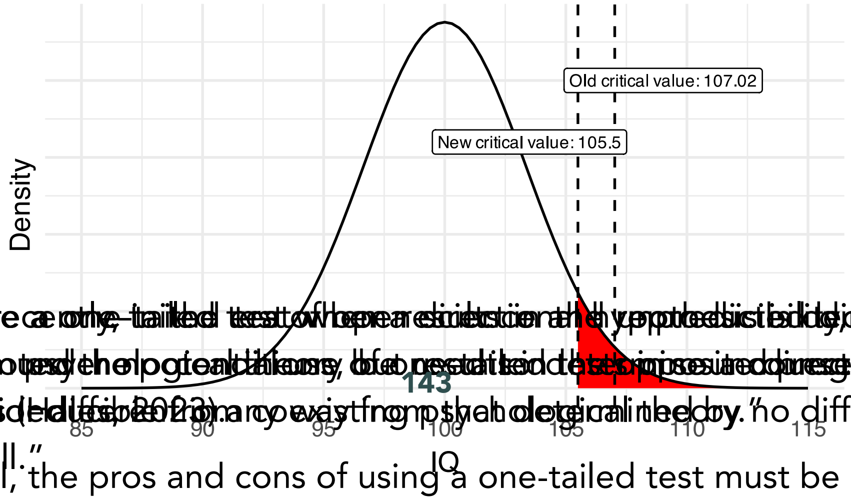
So far we have looked at two-sided tests. Recall in a previous chapter we looked at distributions of tests statistics based on certain sample sizes and population parameters.

For example:



In the above, each of the two tails has 2.5%, which total 5%. The horizontal lines represents the critical values. Some researchers may have hypothesis so specific that they are confident their tests statistics will be in one specific fail of that distribution. Well, those research may wish to keep a total 5% of extreme values to represent *statistical significance*, but put them exclusively in one tail. Compare the following to the above, which both have red regions accounting for 5% of the total distribution:

```
[1] 105.517
```



Why would a researcher do this? Well, what do you notice about the

[illegible]

[illegible] at all.”

Overall, the pros and cons of using a one-tailed test must be considered and determine **prior** to conducting any formal analyses.

10.2.2 2. Designing a study

Back to our Lay's research. Determined to investigate further, you plan to purchase two cases of chips (15 bags each) directly from Lays: one case of Ketchup Chips and one case of Regular Chips. Your total sample size is 30 bags–15 Ketchup, 15 Regular.

Once you have the chips, you will weigh each bag using a professionally calibrated scale, ensuring the weights reflect only the chips and seasoning themselves, excluding the bags. You decide to use null hypothesis significance testing (NHST) to analyze your data with a significance level of $\alpha = .05$.

Prior to conducting your research, you submit your research plan to the Grenfell Campus research ethics board, which approves your study and classified it as low-risk.

10.2.3 3. Collecting data

You follow through with your research plan. You get the following data:

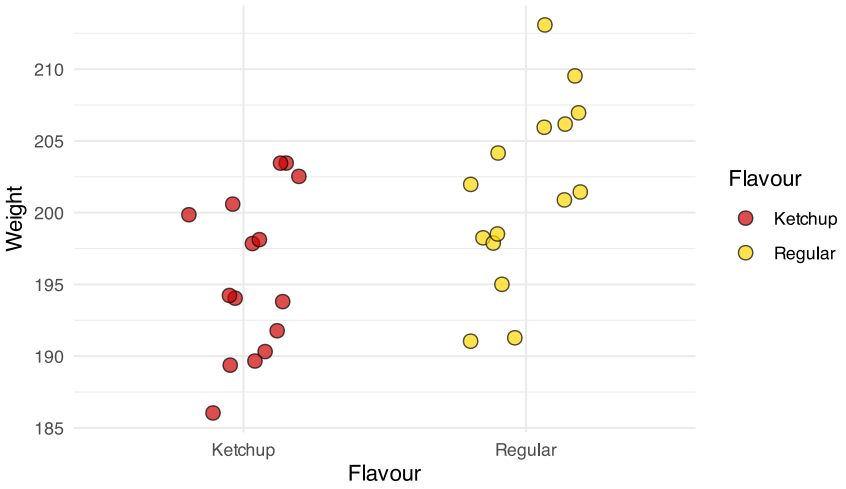
Bag_ID	Flavour	Weight
1	Ketchup	203.5
2	Ketchup	202.5
3	Ketchup	189.4
4	Ketchup	190.3
5	Ketchup	203.4
6	Ketchup	199.9
7	Ketchup	200.6
8	Ketchup	197.8
9	Ketchup	191.8
10	Ketchup	186
11	Ketchup	193.8
12	Ketchup	194
13	Ketchup	194.2
14	Ketchup	189.7

Bag_ID	Flavour	Weight
15	Ketchup	198.1
16	Regular	207
17	Regular	201.4
18	Regular	209.5
19	Regular	195
20	Regular	204.2
21	Regular	206.2
22	Regular	198.2
23	Regular	200.9
24	Regular	197.9
25	Regular	198.5
26	Regular	191.3
27	Regular	205.9
28	Regular	191
29	Regular	202
30	Regular	213.1

And we can summarize the data:

Flavour	Mean	Min	Max	SD
Ketchup	195.673	186	203.5	5.616
Regular	201.475	191	213.1	6.359

It is also helpful to visualize the data using a graph/plot. There are several options for a t-test (see an earlier chapter regarding ways to visualize data). For now, we will create a dot plot that has a dot for each bag of chips. The y-axis represent the weight, and the x-axis represent the flavor. I have added some slight x-axis movement within each flavor to prevent dots from overlapping.



Think about it

How do you make sense of this figure? Are there trends you see? If you have to guess, without having access to a formal analysis, would the flavors weight the same?

10.2.4 4. Analyzing data

If you recall the concept of the distribution of sample means from the z-test chapter, you know that sample means vary due to random sampling. For example, even if a population mean is 200g with a standard deviation of 6g, taking a sample from that population might not give us exactly 200g every time due to this variability.

We can apply this same logic to the differences between two groups' means. Even if the true means of both groups are identical, sampling variability will cause the observed difference between the sample means to deviate from 0. In other words, we may see some differences between the two groups' sample means, even if both are drawn from the same population or populations with identical means.

A t-test helps us assess whether the observed difference between these sample means is large enough to be statistically significant. By using a pre-specified significance level (like $\alpha = 0.05$), we can determine whether the difference is so large that it's unlikely to have occurred if the groups indeed came from the same population—suggesting that the two groups' means are different.

We will need several pieces of information prior to analyzing our results.

10.2.4.1 Deriving the Distribution of Differences in Sample Means

The distribution of differences in sample means, known as the *sampling distribution of the difference between two means*, is derived from the individual sampling distributions of each group's mean.

10.2.4.1.1 1. Sampling Distribution of Each Group's Mean

For each group in an independent t-test, the sampling distribution of the sample mean is normally distributed (or approximately normal for large enough sample sizes, due to the Central Limit Theorem). The mean of each group's sampling distribution is the population mean (μ), and the variability is represented by the standard error of the mean. For Group 1, the standard error of the mean (SE1) is:

$$SE_1 = \frac{s_1}{\sqrt{n_1}}$$

where s_1 is the sample standard deviation of Group 1, and n_1 is the sample size. For Group 2, the standard error of the mean (SE2) is:

$$SE_2 = \frac{s_2}{\sqrt{n_2}}$$

10.2.4.1.2 2. Sampling Distribution of the Difference Between Means

To obtain the sampling distribution of the *difference* between the two sample means ($\bar{X}_1 - \bar{X}_2$), we combine the two individual sampling distributions. The mean of the difference between the two sample means is the difference between the population means:

$\mu_{(\bar{X}_1 - \bar{X}_2)} = \mu_1 - \mu_2$

If the null hypothesis (H_0) is true (i.e., the two population means are equal), then the mean difference will be 0 ($\mu_1 - \mu_2 = 0$).

10.2.4.1.3 3. Standard Error of the Difference Between Means

The variability (spread) of the distribution of the difference between means is captured by the *standard error of the difference*. This is calculated by combining the standard errors of both groups. If we assume that the variances of the two groups are equal, we use the *pooled standard deviation* to compute the standard error of the difference:

$$SE_{\bar{X}_1-\bar{X}_2} = \sqrt{\left(\frac{s_p^2}{n_1}\right) + \left(\frac{s_p^2}{n_2}\right)}$$

Where s_p is the pooled standard deviation, and n_1 and n_2 are the sample sizes of the two groups. The pooled standard deviation (s_p) is calculated as follows. It's often called the summation form:

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

or, an alternative formula for pooled variance follows. You can take the square root of the following to obtain the former. It is often called the weighted variance form:

$$s_p^2 = \frac{\Sigma(x_{i1}-\bar{x_1})^2 + \Sigma(x_{i2}-\bar{x_2})^2}{n_1+n_2-2}$$

10.2.4.1.4 4. t-Distribution

The sampling distribution of the difference between means follows a t-distribution when the population variances are unknown. The degrees of freedom (*df*) for this t-distribution are:

$$df = n_1 + n_2 - 2$$

This t-distribution is used because we are estimating the population variances from the sample data, and it accounts for the uncertainty associated with that estimation.

10.2.4.1.5 5. Calculating the t-statistic

The t-statistic is calculated by comparing the observed difference between the two sample means to the expected difference under the null hypothesis. The formula is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE_{\bar{X}_1 - \bar{X}_2}}$$

Where $\bar{X}_1$ and $\bar{X}_2$ are the sample means for Group 1 and Group 2 and $SE_{\bar{X}_1 - \bar{X}_2}$ is the standard error of the difference between the two means, which is calculated using the pooled standard deviation.

The t-statistic tells us how many standard errors the observed difference between means is away from 0 (the expected difference under the null hypothesis). A large t-value suggests that the difference between the means is unlikely to have occurred by chance, leading to the potential rejection of the null hypothesis.

10.2.4.2 Assumptions

There are several assumptions we must have when testing from the t distribution.

1. The data are continuous. For our purposes, this will be interval or ratio data.
2. The data are randomly sampled.
3. The variance of each group is similar.

10.2.4.3 Welch's t-test

So far we have used formulas from Student's t-test. Specifically, Welch's t-test is an alternative test that is more robust to unequal group variances and smaller sample sizes. Welch's t-test alters the denominator for the t-test in the equation to:

$$\sqrt{s_{x1}^2 + s_{x2}^2}$$

where

$$s_{xi}^2 = \frac{s_i}{\sqrt{n_i}}$$

Thus, the overall equation for Welch's t-test is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_{\overline{x1}}^2 + s_{\overline{x2}}^2}}$$

Furthermore, Welch's t-test alters the degrees of freedom (*v*) to:

$$v \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 v_1} + \frac{s_2^4}{n_2^2 v_2}}$$

Importantly, there are no major disadvantages to using Welch's versus Student's and you should probably use it in your own research. R's function `t.test()` automatically uses Welch's t-test.

For the purposes of this course, we will use Student's t-test for our hand calculations. However, you can use Welch's t-test for any analyses conducted using statistical software.

10.2.4.4 Effect Size

Cohen's d is the standard effect size estimate for a t-test. It provides us with an estimate of the standardized mean difference. It is:

$$d = \frac{\overline{X}_1 - \overline{X}_2}{s_{pooled}}$$

This is a standardized effect size that be compared across groups of metrics. Please review the chapter that discussed how to best determine meaningful effect sizes. However, Cohen suggested the following cut-offs:

- Small - $d = .2$
- Medium - $d = .5$
- Large - $d = .8$

10.2.4.5 Ketchup a rip-off?

Let's apply this to our chips example. We have all the data to calculate our t-statistic.

Flavour	Mean	Min	Max	SD
Ketchup	195.7	186	203.5	5.616
Regular	201.5	191	213.1	6.359

We calculate our squared differences between each bag and the mean of that group, which will be needed later. Not that in the following table, the last column is the squared difference between the weight of a bag of chips (X) and the mean of that bag's GROUP ($\bar{X}$; ketchup or regular):

Bag_ID	Flavour	Weight	x_minus_mean_square
1	Ketchup	203.5	60.632178
2	Ketchup	202.5	47.013878
3	Ketchup	189.4	39.732011
4	Ketchup	190.3	28.658178
5	Ketchup	203.4	60.476544
6	Ketchup	199.9	17.528178
7	Ketchup	200.6	24.272044
8	Ketchup	197.8	4.737878
9	Ketchup	191.8	15.236011
10	Ketchup	186	92.801111
11	Ketchup	193.8	3.509378
12	Ketchup	194	2.667778
13	Ketchup	194.2	2.083211
14	Ketchup	189.7	36.160178
15	Ketchup	198.1	5.986178
16	Regular	207	30.081568
17	Regular	201.4	0.001248
18	Regular	209.5	64.877655
19	Regular	195	41.800535
20	Regular	204.2	7.207435
21	Regular	206.2	22.039895
22	Regular	198.2	10.402775
23	Regular	200.9	0.342615

Bag_ID	Flavour	Weight	x_minus_mean_square
24	Regular	197.9	12.926422
25	Regular	198.5	8.733995
26	Regular	191.3	103.944822
27	Regular	205.9	20.022642
28	Regular	191	108.896182
29	Regular	202	0.244695
30	Regular	213.1	134.668288

So, for bag 1:

$$(X-\bar{X})^2 = (203.46-195.6733)^2 = 60.63$$

Let's fill in the missing data to compute our t-statistic. We have:

- $\bar{X}_1 = 195.67$;
- $\bar{X}_2 = 201.48$;
- $s_p^2 = \frac{\Sigma(X_{i1}-\bar{X}_1)^2+\Sigma(X_{i2}-\bar{X}_2)^2}{n_1+n_2-2} = \frac{441.49+556.19}{15+15-2} = \frac{1007.67}{28} = 35.99$

and, therefore:

$$t = \frac{195.6733-201.4753}{\sqrt{\frac{35.99}{15}+\frac{35.99}{15}}} = \frac{-5.802}{2.190586} = -2.6486$$

You may wish to look the p-value of the resulting test up in a critical value table. However, most likely you will use statistical software to provide you with an exact p-value.

Formal Results

The following is the formal results of our t-test:

```

    Two Sample t-test

data:  Weight by Flavour
t = -2.649, df = 28, p-value = 0.0131
alternative hypothesis: true difference in means between group
```

```
Ketchup and group Regular is not equal to 0
95 percent confidence interval:
 -10.28914  -1.31486
sample estimates:
mean in group Ketchup mean in group Regular
              195.673               201.475
```

10.2.4.6 Cohen's D

Recall that:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_{pooled}}$$

From our above means and pooled variance, we have:

$$d = \frac{195.67 - 201.48}{\sqrt{35.99}} = -0.968$$

10.2.5 5. Write your results/conclusions**

A two Sample t-test testing the difference of weight of bags of chips by Flavor suggest that Ketchup chips $(\bar{X} = 195.67$ weigh less than Regular chips $(\bar{X}) = 201.48)$. The results suggests that the effect is statistically significant, and large $\bar{X}_{diff} = -5.80, 95\% CI[-10.29, -1.31]$, $t(28) = -2.65, p = .013$, Cohen's d = −1.00, 95% CI [−1.78, −0.21]$.

10.3 Conclusion

an independent t-test is a statistical method used to determine if there is a significant difference between the means of two independent groups. Unlike the z-test, the independent t-test is used when the population variances are unknown and typically estimated from the sample data. It calculates a t-statistic, which measures how many standard deviations the difference between the two sample means is from the expected difference (usually zero). By comparing this t-statistic to a critical value from the t-distribution, researchers can determine whether the observed difference is statistically significant.

10.4 Practice Questions

1. Practice Question: Calculate the degree of freedom that would be used in Welch's t-test on the chip data:

Bag_ID	Flavour	Weight
1	Ketchup	203.5
2	Ketchup	202.5
3	Ketchup	189.4
4	Ketchup	190.3
5	Ketchup	203.4
6	Ketchup	199.9
7	Ketchup	200.6
8	Ketchup	197.8
9	Ketchup	191.8
10	Ketchup	186
11	Ketchup	193.8
12	Ketchup	194
13	Ketchup	194.2
14	Ketchup	189.7
15	Ketchup	198.1
16	Regular	207
17	Regular	201.4
18	Regular	209.5
19	Regular	195
20	Regular	204.2
21	Regular	206.2
22	Regular	198.2
23	Regular	200.9

Bag_ID	Flavour	Weight
24	Regular	197.9
25	Regular	198.5
26	Regular	191.3
27	Regular	205.9
28	Regular	191
29	Regular	202
30	Regular	213.1

2. Practice Question: Calculate Student's t statistics for the following data comparing Sour Cream and Onion (SCO) chips to Salt and Vinegar (SV).

Bag_ID	Flavour	Weight
1	SCO	198.3
2	SCO	192.1
3	SCO	204.8
4	SCO	201.6
5	SCO	198.3
6	SCO	196.6
7	SV	193.7
8	SV	197.4
9	SV	199.2
10	SV	198.3
11	SV	213
12	SV	205.8

10.5 Answers

1.

```
     df
27.5777
```

2.

```
Effect sizes were labelled following Cohen's (1988)
recommendations.

The Two Sample t-test testing the difference of Weight by
Flavour (mean in
group SCO = 198.62, mean in group SV = 201.23) suggests that
the effect is
negative, statistically not significant, and small (difference
= -2.62, 95% CI
[-10.09, 4.86], t(10) = -0.78, p = 0.453; Cohen's d = -0.49,
95% CI [-1.74,
0.78])
```

11 Paired t-test

This chapter will cover the repeated measures t-test, a statistical method used to determine if there is a significant difference between the means of two related, repeated, or dependent groups. Unlike the independent t-test, which compares two separate groups, the repeated measures t-test is used when the same participants are measured under different conditions or at multiple points in time. Because any participant is more similar to his or herself than others, their scores at two time points will be more similar. As a result, we can use this to our advantage. The paired t-test calculates a t-statistic based on the *differences* between paired scores, which allows researchers to determine whether any observed changes are statistically significant.

11.1 Some Additional Details

The repeated measures t-test is appropriate for situations in which there is a natural pairing of the data such as when measuring a group of participants who are measured on some outcome both before and after an intervention. Or, as another example, when participants undergo two different experimental conditions; an outcome is measured after receiving each condition.

Importantly the null hypothesis posits that there is no change in the mean difference scores between the two time points or conditions. Specifically, the null hypothesis states:

$$H_0 : \Delta\mu_D = 0$$

where $\Delta\mu$ (*delta mu*) is the mean of the *differences* of participants across two time points or two conditions. The alternative hypothesis states (for a two-sided test)

$$H_1 : \Delta\mu_D \neq 0$$

In this context, we're testing whether there is a statistically significant difference in the mean scores of the difference between participants at two time points or under two conditions, rather than comparing two independent groups. So instead of calculating a mean for group 1 and another for group 2, we calculate the difference scores for each participant and get the mean of those differences.

11.2 Key Assumptions

A repeated measures t-test can be conducted under certain assumptions. We will explore these in more detail later, but in short:

First, the data are continuous. The dependent variable should be at the interval or ratio level. **Second**, the paired scores are normally distributed. Stated another way, the differences between paired scores should follow a normal distribution. **Last**, there is independence of observations within each pair; each observation in one condition should correspond to a *single* observation in the other condition. With these assumption in mind, let's work through an example.

11.3 Have you tried… just not being anxious?

Imagine a researcher that wants to test the effectiveness of a new anxiety-reduction therapy. The researcher plans on recruiting individuals who are diagnosed with generalized anxiety disorder (GAD) and measures their anxiety levels before and after completing the therapy program. The researcher believes that **the therapy will reduce participants' anxiety levels**. However, they decide that a two-tailed test would be best, in case the new therapy program *worsens* anxiety.

11.3.1 1. Generating hypotheses

We can translate this conceptual hypothesis into a statistical one. For a repeated measures t-test, we are interested in whether the mean of the differences between the two sets of measurements (pre-therapy and post-therapy anxiety levels) is significantly different from zero. Thus, our hypotheses are as follows:

$$H_0 : \Delta\mu = 0$$

$$H_1 : \Delta\mu \neq 0$$

11.3.2 2. Designing a study

In brief, the method for this study is:

Participants: Participants will be recruited by placing recruitment posters at a local hospital. Interested participants will complete an anxiety questionnaire and those with a score of 50 or above on the anxiety measurement will meet criteria for participation. Those currently receiving psychological services outside of the study will be excluded. A power analysis revealed that 19.34 participants (20) are required for adequate power.

Measures: Anxiety was measured using the Anxiety Questionnaire for Adults (AQA). The questionnaire consists of 20 items, designed to evaluate the frequency and intensity of anxiety symptoms experienced over the past week. Each item is rated on a 5-point Likert scale, ranging from 1 (not at all) to 5 (very often). Higher total scores indicate greater levels of anxiety. The AQA has demonstrated strong psychometric properties in community and clinical samples.

Procedure: Participants for the study were recruited through community centers and online platforms, where they were provided with information about the study's purpose, procedures, and potential risks and benefits. Interested individuals who met the inclusion criteria (ages 18 and older) were screened via a brief eligibility questionnaire.

Upon obtaining informed consent, participants were administered the Anxiety Questionnaire for Adults (AQA) in a controlled environment. The questionnaire was presented either individually or in small groups, depending on the setting. The administration of the AQA took approximately 10-15 minutes.

Included participants were enrolled in a six-week therapy program. Participants completed individual therapy once per week for the six weeks with a doctoral-level clinical psychologist. Participants who missed more than two sessions were excluded from analysis.

The ethics review board at Grenfell Campus reviewed the project and ethics submission and approved the study.

11.3.3 3. Collecting data

The study was completed as described; a final sample size of 20 was used. The following data were obtained:

Participant_ID	Pre_Therapy	Post_Therapy
1	57.1	49.1
2	66.9	54.7
3	72.4	60
4	55.8	38.7
5	59.7	41.6
6	71.1	63.5
7	64.8	56.8
8	60.2	39.4
9	74.9	44.9
10	66.5	53.5
11	60.4	49
12	62.9	43
13	57.6	40.3
14	74.3	63.2
15	74.6	62
16	60.3	45.3

Participant_ID	Pre_Therapy	Post_Therapy
17	59.5	41.6
18	64.3	53
19	59	43.1
20	63.2	39.1

11.3.4 4. Analyzing data

To assess the significance of the mean of the differences in anxiety scores, we calculate the t-statistic for the paired differences. There is some information we need to calculate the statistics. We require:

11.3.4.1 Mean Difference

First, we need the mean of the differences between the pre- and post-therapy scores. We can simply calculate the difference between each group's mean. Here, the mean of the pre-therapy condition is 64.275 and the mean of post-therapy is 49.09. Thus:

$$\Delta\mu = \mu_1 - \mu_2$$

And for this study:

$$\Delta\mu = 64.275 - 49.09 = 15.185$$

11.3.4.2 Standard Deviation of Differences

Next, we need to calculate the standard deviation of these differences. For us, we will first need difference scores (D_i)for each person. For example, the difference score (representing by Δ) for person 1 (x_1) is:

$$\Delta x_1 = 57.1 - 49.1 = 8.0$$

We would do this for each individual. These are:

Participant_ID	Pre_Therapy	Post_Therapy	Difference
1	57.1	49.1	8
2	66.9	54.7	12.2
3	72.4	60	12.4
4	55.8	38.7	17.1

Participant_ID	Pre_Therapy	Post_Therapy	Difference
5	59.7	41.6	18.1
6	71.1	63.5	7.6
7	64.8	56.8	8
8	60.2	39.4	20.8
9	74.9	44.9	30
10	66.5	53.5	13
11	60.4	49	11.4
12	62.9	43	19.9
13	57.6	40.3	17.3
14	74.3	63.2	11.1
15	74.6	62	12.6
16	60.3	45.3	15
17	59.5	41.6	17.9
18	64.3	53	11.3
19	59	43.1	15.9
20	63.2	39.1	24.1

Next we would calculate the standard deviation for these difference scores. Note that the mean of the difference scores is the same as the mean of pre-therapy subtract the mean of post_therapy (15.185).

$$s_D = \sqrt{\frac{\sum_{i=1}^{n} \left(D_i - \bar{D}\right)^2}{n-1}}$$

where:

- D_i represents each individual difference,
- $\bar{D}$ is the mean of the differences,
- n is the number of paired observations.

For us, this works out to be:

- $\sum \left(D_i - \bar{D}\right)^2 = 611.365$
- $n - 1 = 20 - 1 = 19$

Thus:

$$s_D = \sqrt{\frac{611.365}{19}} = \sqrt{32.177} = 5.677$$

11.3.4.3 Standard Error of the Mean Difference

The standard error of the mean difference is calculated as follows:

$$SE_D = \frac{\sum\left(d_i - \overline{D}\right)^2 \frac{1}{N}}{\sqrt{n}} = \frac{s_D}{\sqrt{n}}$$

where:

- s_D is the standard deviation of the differences
- n is the number of paired observations.

Thus, our SE can be calculated as follows:

$$SE_D = \frac{5.677}{\sqrt{20}} = 1.27$$

11.3.4.4 Calculate the t-Statistic

The t-statistic is calculated by dividing the mean difference by the standard error:

$$t = \frac{\bar{X}_{difference}}{SE} = \frac{15.185}{1.27} = 11.96$$

Importantly, the paired t-test has one extra degree of freedom compared to the independent samples t-test. For paired t-test:

$$df = n - 1$$

A formal analysis would result in:

```
    Paired t-test

data:  df_anxiety$Pre_Therapy and df_anxiety$Post_Therapy
t = 11.96, df = 19, p-value = 2.73e-10
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 12.5282 17.8418
```

```
sample estimates:
mean difference
         15.185
```

The t-test output provides the t-statistic, degrees of freedom, and p-value. If the p-value is below the significance level ($\alpha = .05$), we can conclude that difference between pre- and post-therapy scores is unlikely if the null were true (i.e., that the difference between before versus after therapy was 0; no change).

11.3.4.5 Effect Size: Cohen's d

For a paired-samples t-test, Cohen's d provides an estimate of the standardized mean difference. Cohen's d for repeated measures is calculated as follows:

$$d = \frac{\bar{X}_D}{s_D}$$

For us:

$$d = \frac{15.185}{5.677} = 2.67$$

Note that due to a small sample size, some statistical software may apply a correction to test statistics. Specifically, you should use Hedge's g when dealing with a small sample size. In fact, there is no downside to using Hedge's g. For small sample sizes, it applies a correct. For larger sample sizes, it will approximate Cohen's. In r, our results are:

```
Cohen's d |       95% CI
------------------------
2.67      | [1.72, 3.62]
```

This standardized effect size allows us to determine the practical significance of the results. Cohen suggested interpreting d values as follows:

- Small: ($d = 0.2$)
- Medium: ($d = 0.5$)
- Large: ($d = 0.8$)

11.3.5 5. Write your results/conclusions

A paired t-test was used to determine the efficacy of the therapy by testing the difference between the Pre-Therapy Post-Therapy scores. The results suggests that the results are unlikely given a true null hypothesis, $\bar{D} = 15.19$, $95\% CI[12.53, 17.84]$, $t(19) = 11.96$, $p < .001$. Additionally, the effect is considered large, $d = 2.67$, $95\% CI[1.72, 3.62]$.

11.4 Conclusion

The paired t-test is a more powerful alternative to between subjects t-tests that uses the same participants across multiple conditions. It is good for testing changes in a variable over time such as in pre-post designs. The repeated measures t-test takes advantage of the dependence of observations, helping researchers draw conclusions.

Tip

Building Your Toolbox

It may be good practice to take note of the use cases for each analyses you learn about. For example, for a repeated measure/paired t-test, what are the main uses, number and type (e.g., NOIR) of IVs, number and type of DVs, major assumptions, statistical hypotheses, and effect sizes?

In future chapters, add to this table:

Test Names	Main Uses	Number of IVs	Number of DVs	IV Type	DV Type	Assumptions	Hypotheses	Effect Size
z-test	Compare one group's mean to a population mean.	0 (No IVs)	1	None or Categorical (e.g., Group)	Continuous	Normality, known population variance	Null: Mean of group equals population mean, Alternative: Mean of group differs from population mean	Cohen's d
Independent t-test	Compare means between two independent groups.	1 (Categorical, e.g., Group)	1	Categorical (2 groups)	Continuous	Normality, equal variances (for Student's t-test), independence	Null: Means of the two groups are equal, Alternative: Means of the two groups differ	Cohen's d (or Hedges' g)
Repeated Measures t-test	Compare means within the same group at different time points.	1 (Categorical, e.g., Time Point)	1	Categorical (1 group)	Continuous	Normality of differences, sphericity (if applicable)	Null: Means at different time points are equal, Alternative: Means at different time points differ	Cohen's d (for paired samples)

12 One way ANOVA

This chapter will cover the one-way ANOVA, a statistical method used to determine if there is a significant difference among the means of three or more independent groups. Unlike a t-test, which compares the means of only two groups, the one-way ANOVA extends this comparison to multiple groups to test whether at least one group mean is significantly different from the others. It calculates an F-statistic by comparing the variance between group means to the variance within groups, allowing researchers to determine whether any observed differences are statistically significant. This should make intuitive sense; if the groups do not differ, their mean differences shouldn't be that different from the random variation within each group.

12.1 Some Additional Details

The one-way ANOVA is appropriate for situations where there is one independent variable (IV) with three or more levels (e.g., conditions or groups) and one continuous dependent variable (DV). For example, researchers might use a one-way ANOVA to compare three different teaching methods' (IV with three levels) impact on test scores (DV). Or, as another example, to assess the effects of three different diets (IV with three levels) on weight loss (DV). Importantly, in a one-way ANOVA, the participants in each group are different.

The null hypothesis for the one-way ANOVA (hereafter, I will simply write 'ANOVA') posits that all group means are equal:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

where k represents the number of groups. The alternative hypothesis states that at least one group mean is different:

$$H_1 : \text{At least one } \mu_i \text{ differs from the others.}$$

💡 ANOVA Hypotheses

The above is the general structure of H_0 and H_1 in ANOVA designs.

By rejecting the null hypothesis, researchers conclude that a statistically significant difference exist between *at least* two group means: at least one group differs from at least one other group. However, post hoc tests are typically required to identify *where* these differences occur.

12.2 Key Assumptions

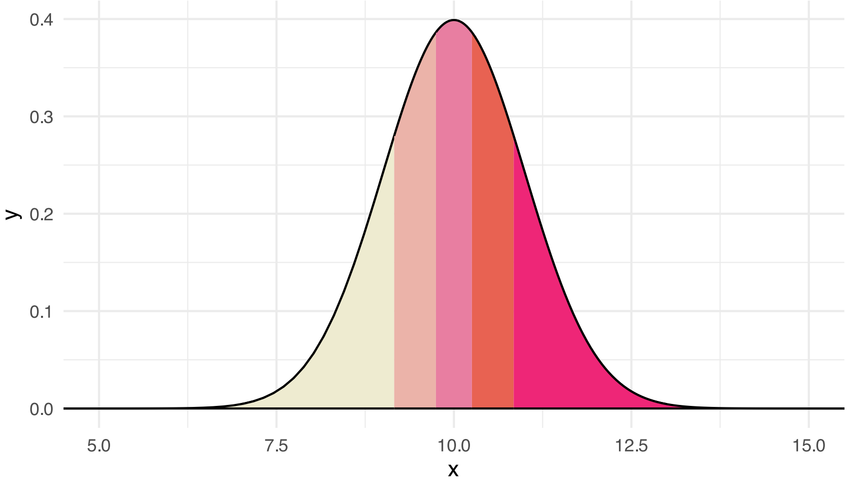
A one-way ANOVA can be conducted under certain assumptions. **First**, the data are continuous. That is, the dependent variable should be at the interval or ratio level. **Second**, there needs to be independence of observations. Each group should consist of independent individuals; this means that no participant is in more than one group. This assumption is sometimes called independence of residuals. **Third** is the homogeneity of variances. Here, the variance within each group should be approximately equal. This can and will be tested using Levene's test. We want a non-statistically significant result for Levene's test. **Fourth**, and last, the residuals should be normally distributed. Recall that the residuals are deviations from an observed and predicted score. It's a common misconception that the DV must be normally distributed. In reality, the residuals should be normally distributed. We can visualize normally distributed residuals using a Q-Q plot or formally test this using the Shapiro-Wilks test. Note that the SW test is less informative for large sample sizes.

Q-Q Plot

The Q-Q plot show what a variable would like like if it were normally distributed compared with what it actually is. In essence, we order our variable and compare to what the quantiles should look like under normality. Consider the following data:

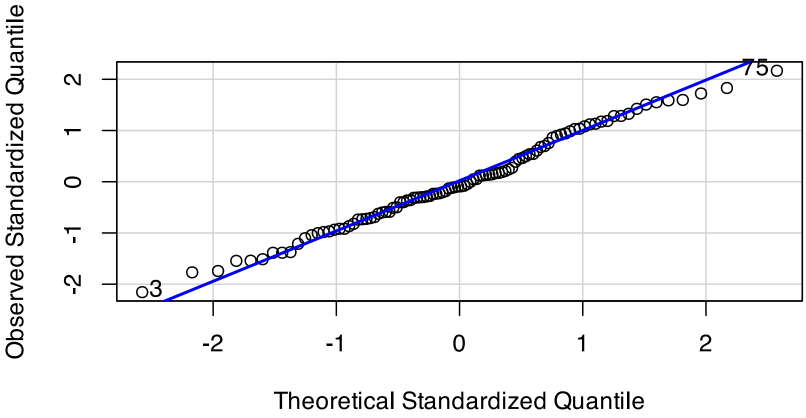
x
10.77
9.86
9.8
9.92
8.82

We could order these variables as: 8.82, 9.8, 9.86, 9.92, 10.77. Under a normal distribution, we would create five quantiles (because we have five score). Quantiles are cut-off points that each contain the same amount of the distribution. Here our quantiles have 20% ($\frac{100\%}{5} = 20\%$)



Believe it or not, each section in the above contains the same proportion. Thus, if data were normally distributed and we drew five numbers, like above, we would expect one to fall in each quantile.

The QQ plot compares where we would expect values to fall on a normal curve versus where they actually are (e.g., z-score). Quantiles are calculated for your data and for the theoretical distribution (e.g., Normal). Each data point's quantile is plotted against the corresponding theoretical quantile on the plot. The following is what a Q-Q plot looks like for data with 100 observations (there would be 100 quantiles, each with 1% of a normal distribution):



```
[1] 75  3
```

Normally distributed data will fall close to the line (vague, I know). The above graph does let us know that observation 75 appears to be particularly problematic. Typically, we would be concerned if the points systematically varied from the line.The tails typically stray farther from the line.

Shapiro-Wilks Test

The Shapiro-Wilks test assesses how far the data deviates from normality. A statistically significant result ($p < .05$) is typically interpreted as the data deviating from normality.

12.3 Treatments for OCD: What Works?

O'Connor et al. (2006) compared the efficacy of treatments for obsessive compulsive disorder (OCD). They measured the severity of OCD symptoms, with lower scores indicating better outcomes (i.e., fewer symptoms). Specifically, they compared differences in OCD severity after individuals received one of:

1. Cognitive Behavioral Therapy (CBT)
2. Medication
3. CBT + Medication

And they believed that the CBT + Medication would show lower OCD symptoms than the other groups.

12.3.1 1. Generating hypotheses

The main null and alternative hypotheses for the ANOVA can be converted into a statistical hypothesis stated as (for the null):

$$H_0 : \mu_{CBT} = \mu_{Rx} = \mu_{CBT+Rx}$$

And (for the alternative):

$$H_1 : \text{At least one } \mu_i \text{ differs from the others.}$$

12.3.2 2. Designing a study

We will use a research design to conduct a similar analysis as O'Connor et al. (2006) using fake/hypothetical replication data. Although there are some slight adjustments, our method follows:

Participants: Participants were recruited from the Montreal community, meeting criteria for severe OCD (Y-BOCS >16). Exclusions included major psychiatric cormorbidities like substance abuse or psychosis. A power analysis determined that a sample size of 10 individuals per group were needed for adequate power.

Materials: The Yale-Brown Obsessive Compulsive Scale (Y-BOCS) was used to assess OCD severity.

Procedure: Participants were randomly assigned to either:

1. Fluvoxamine (hereafter, **Rx**): for 5 months
2. CBT only: 20 sessions focusing on exposure and cognitive restructuring
3. Fluvoxamine + CBT: 20 sessions while also taking medication.

Symptoms were measured at the end of the treatment conditions. Unfortunately, we lost any pre-treatment data so we can only assess OCD severity at the end of treatment. Thus, we have three major groups (Rx, CBT, Rx + CBT) and one outcome (OCD severity).

The ethics review board at Grenfell Campus reviewed the project and ethics submission and approved the study.

12.3.3 3. Collecting data

The study was completed as described; a final sample size of 30 (10 per group) was used. The following data were obtained:

ID	Treatment	OCD_Severity
1	CBT	12
2	CBT	8
3	CBT	12
4	CBT	10
5	CBT	7
6	CBT	10
7	CBT	10
8	CBT	11
9	CBT	11
10	CBT	14
11	Rx	14
12	Rx	12
13	Rx	8
14	Rx	11
15	Rx	9
16	Rx	10
17	Rx	8
18	Rx	13
19	Rx	9
20	Rx	7
21	CBT_Rx	9

ID	Treatment	OCD_Severity
22	CBT_Rx	6
23	CBT_Rx	8
24	CBT_Rx	9
25	CBT_Rx	8
26	CBT_Rx	7
27	CBT_Rx	6
28	CBT_Rx	7
29	CBT_Rx	10
30	CBT_Rx	10

12.3.4 4. Analyzing data

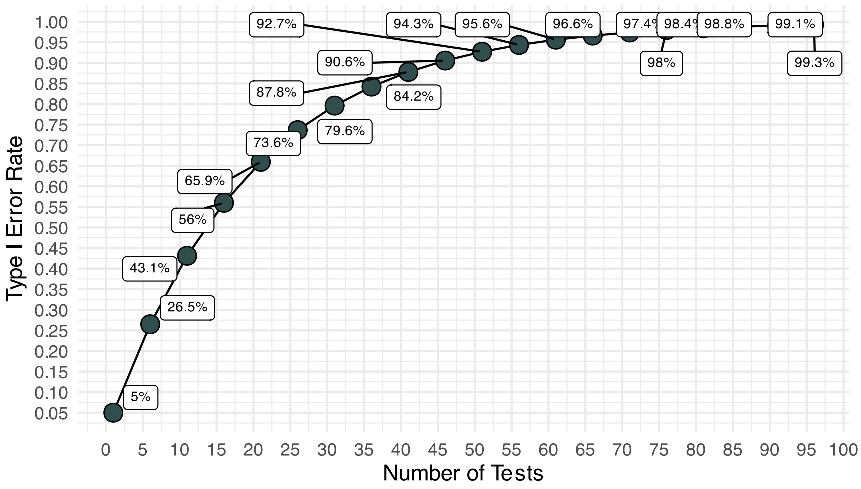
Theoretically, we could do three t-tests:

1. CBT versus Rx
2. CBT versus CBT + Rx
3. Rx versus CBT + Rx

However, this would result in an inflated Type I error rate. Recall that we typically set our $\alpha = .05$ $(1 - \alpha = .95)$. With three independent comparisons our alpha rate actually becomes:

$1 - (1 - \alpha)^{n_{comparisons}} = 1 - (1 - .05)^3 = .142625$

Our Type I error rate would go from 5% to 14.26%! If we had four groups, we would have six possible comparisons and our Type I error rate would increase to 26.49%. The following figure represents the relationship between number of comparisons and Type I error.



Understanding this figure is important. Yes, you could do a bunch of t-tests, but you trade off your alpha rate. Instead, you could use an ANOVA.

An ANOVA allows us to compare 2+ groups within a single test – and as you will see later, multiple IVs at once. Specifically, we can test an independent variable's association with a dependent variable and answer, "do different groups/levels differ on the DV?" Prior to diving into ANOVAs, we should understand the F-distribution.

Understanding the F-Distribution

To understand the *F*-distribution, we need to explore the chi-square distribution. Imagine we measure IQ scores in a random sample of four 20-year-olds. Let's assume IQ is normally distributed with a mean of 100 and a standard deviation of 15. For example:

name	iq
Alicia	85
Ramona	108
Marie	91
Julianna	108

We can standardize these scores by converting them to *z*-scores using the formula:

$$z = \frac{x_i - \mu}{\mathrm{SD}}$$

Here's how those z-scores look:

name	iq	z
Alicia	85	−1
Ramona	108	0.53
Marie	91	−0.6
Julianna	108	0.53

These z-scores are also called *standard deviates*: they represent how far each score deviates from the mean, measured in standard deviations. Since the IQ scores are independently and identically distributed (IID), we can calculate the **sum of squares** of these standard deviates:

$$SS = \sum_{i=1}^{n} z_i^2$$

For our sample:

$$SS = (-1)^2 + (0.53)^2 + (-0.6)^2 + (0.53)^2 = 1.92$$

The **chi-square value** is the sum of squares of standard deviates. The **chi-square distribution** describes the distribution of these summed squares when sampling repeatedly (theoretically, infinitely) from a normally distributed population. For example, if we randomly sampled four IQ scores repeatedly, the sum of squared standard deviates

Connecting Chi-Square to Variance

Recall that variance can be expressed as:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Interestingly, the chi-square value can be scaled to estimate variance. By dividing the sum of the standard deviates by the degrees of freedom, we approximate a variance. For our four people from the IQ example, recall that:

$$SS = (-1)^2 + (0.53)^2 + (-0.6)^2 + (0.53)^2 = 1.92$$

And the approximation of a variance being:

$$\text{Scaled Chi-square} = \frac{\chi^2}{df} = \frac{1.92}{4-1} = 0.64$$

The **F-distribution** is the ratio of two independent, scaled chi-square distributions. Imagine we have two chi-square distributions: one with df_1 and the other with df_2. The F-statistic is calculated as:

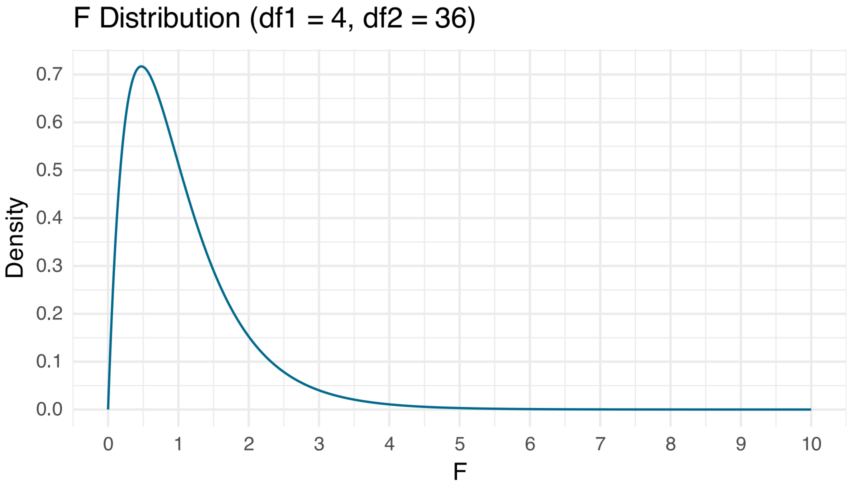
$$F = \frac{\chi_1^2/\mathrm{df}_1}{\chi_2^2/\mathrm{df}_2}$$

This ratio is central to **ANOVA**. In ANOVA, we compare two variances:

1. The variance of between the groups' means (*MSB*: mean square between).
2. The variance of individual scores within the groups (*MSE*: mean square error).

If the groups are sampled from the same population (i.e., the null hypothesis is true), the ratio of these variances should follow the *F*-distribution. If group means vary substantially more than the variance within the group, than the F value will be larger than expected. If the value is at or beyond a pre-specified value (i.e., beyond our critical alpha value), we assume that our data are unlikely given a true null hypothesis and we reject the null hypothesis.

To help you visualize the distribution, here's an example of the F-distribution with $df_1 = 4$ and $df_2 = 36$:



You can use critical values for F-distributions the same way you would for z or t tests. We can choose the most extreme percentage of the distribution that aligns with our critical value and compare our results to that.

We can test our hypotheses through an ANOVA and post-hoc tests.

12.3.4.1 Our Model

Using the above hypothesis, we can frame our model as:

$$outcome_i = group + error_i$$

or more specifically, and expanded on below:

$$y_i = \beta_0 + x_i group_i + error_i$$

So, we think that any participant or individual's (person i) score on the outcome will be a function of some coefficient/intercept (β_0), the person's group membership ($x_i group_i$), and some error ($error_i$).

Our Analysis

If you read the section about the F-statistic earlier, you'll remember that the standard deviates (or z-scores) were squared and summed to calculate a total. We can use a similar idea when analyzing data with multiple groups.

In ANOVA, we compare two sources of variability:

1. Variability *between* group means: This tells us whether the group means are different from each other.

2. Variability *within* groups: This reflects the differences between individual scores and their group's mean.

We calculate these as "sum of squares" and, subsequently, "mean squared", values:

1. **SST (Sum of Squares Total)**: The total variability in all the scores.
2. **SSB (Sum of Squares Between Groups)**: The variability explained by group differences.
3. **SSE (Sum of Squares Error)**: The variability within each group that isn't explained by group differences.

These three components are related in a simple way:

$$SST = SSB + SSE$$

In short, the one way ANOVA breaks the total variability into two parts: what's explained by the group differences (SSB) and what's left over as error (SSE). By comparing these two sources of variability, we can see if the group means are significantly different from each other. Note: later we will explore more complex types of ANOVAs that can break the total variability into more than two parts.

Descriptions of each follow along with visualizations.

12.3.4.2 SST

The sum of squares total represents all the deviations of the model. It is each score compared to the overall mean (sometimes called the *grand mean*). Is it represented by:

$$SST = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

We can calculate SST by subtracting each score from the mean, squaring it, and adding up across all participants. However, there is another formula that may be easier to use:

$$SST = \sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}$$

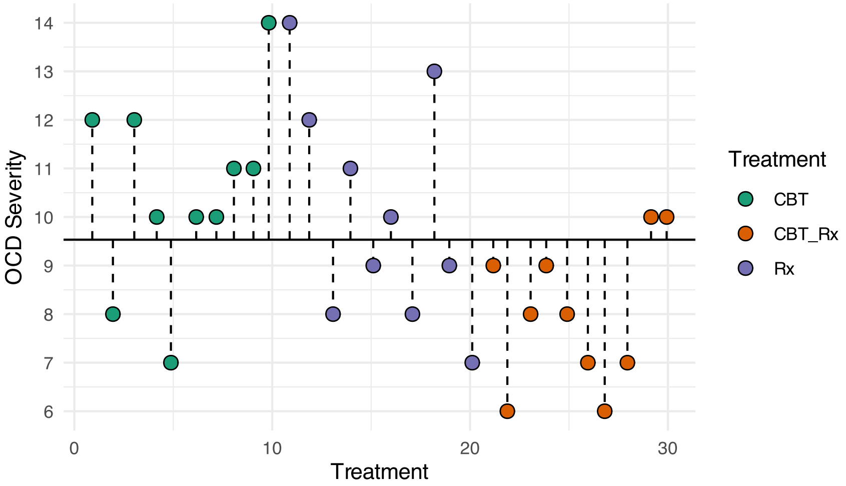
Here you only need three pieces of information:

1. Sum of each squared scores/value: $\sum_{i=1}^{N} x_i^2$
2. Sum of scores/values, then square the total: $\left(\sum_{i=1}^{N} x_i\right)^2$
3. Total sample size: N

For our OCD data, the sum of the squared scores is 2868, the sum of the scores, then squared, is 8.1796^{4}, and N is 30. Thus, our SST is:

$$SST = \sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N} = 2868 - \frac{81796}{30} = 141.467$$

We can also represent SST visually. This figure represents the total deviations used to calculate SST. Each dotted lined represents the difference between each individual and the grand mean (the solid black line). Remember, the grand mean is the mean of *all* individuals.



12.3.4.3 SSB

The sum of squares between groups (SSB) represents the deviation of each group's mean from the grand mean. In the following i represents individual i and j represents group j.

$$SSB = \sum_{j=1}^{n_j} n_j \left(\overline{x}_j - \overline{x}_{overall} \right)^2$$

where n_j is the sample size for group j and $\overline{x}_{overall}$ is the overall/grand mean.

Alternatively, the following formula may be used:

$$SSB = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} x_{ij} \right)^2}{n_i} - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} \right)^2}{N}$$

Here we need the following:

1. Sum of x_i, then squared for each group (j).
2. n_i for each group.
3. Sum of x_i for ALL individuals.
4. N, total sample size.

For us, the sum of all scores, then squared for each group are:

Treatment	Sum_then_squared
CBT	11025
CBT_Rx	6400
Rx	10201

The sum of all scores (regardless of group), then squared it 81796. Each group has $n = 10$ and, thus, $N = 30$.

We will calculate the first part of the formula:

$$\sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} x_{ij} \right)^2}{n_i}$$

for each group and then add them.

For group 1 (CBT):

$$\frac{\left(\sum_{j=1}^{n_i} x_{ij} \right)^2}{n_i} = \frac{11025}{10} = 1102.5$$

For Rx:

$$\frac{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2}{n_i} = \frac{10201}{10} = 1020.1$$

For CBT + Rx:

$$\frac{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2}{n_i} = \frac{6400}{10} = 640$$

Adding these three together give us the first part of the SSB equation:

$$\sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2}{n_i} = 1102.5 + 1020.1 + 640 = 2762.6$$

The second part of the SSB formula is:

$$\frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}\right)^2}{N} = \frac{81796}{30} = 2726.533$$

And our resulting SSB:

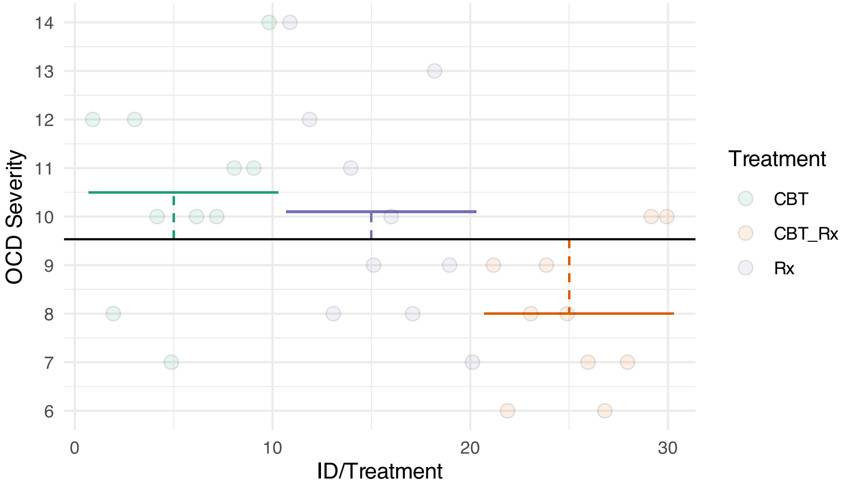
$$SSB = 2762.6 - 2726.533 = 36.067$$

A real head scratcher

It can get confusing when different places use different names for these. For example, we called the total sum of squares SST, but some places call the sum of squares between SST (*treatments*). Other may call the SSB, SSN (numerator). We will stick to the following for a one way ANOVA:

- Sum of squares total (SST)
- Sum of squares between groups (SSB)
- Sum of squares error (SSE)
- SST = SSB + SSE

This second figure represents SSB. Each dotted lined represents the difference between a group's mean and the grand mean. The colored lines represent each groups mean. The black line represents the grand mean.



12.3.4.4 SSE

The sum of squares error SSE represents the deviation of each individual from their group mean.

$$SSE = \sum \left(x_{ij} - \overline{x}_j\right)^2$$

where x_{ij} is individual i in group j and $\overline{x}_j$ is the mean of group j. For the above data, for example the CBT group, we can calculate SSE.

An easier way to calculate this is to subtract SSB from SST. Likewise, if you have SST and SSE, you can quickly calculate SSB. Recall the relationship between the three:

$$SST = SSB + SSE$$

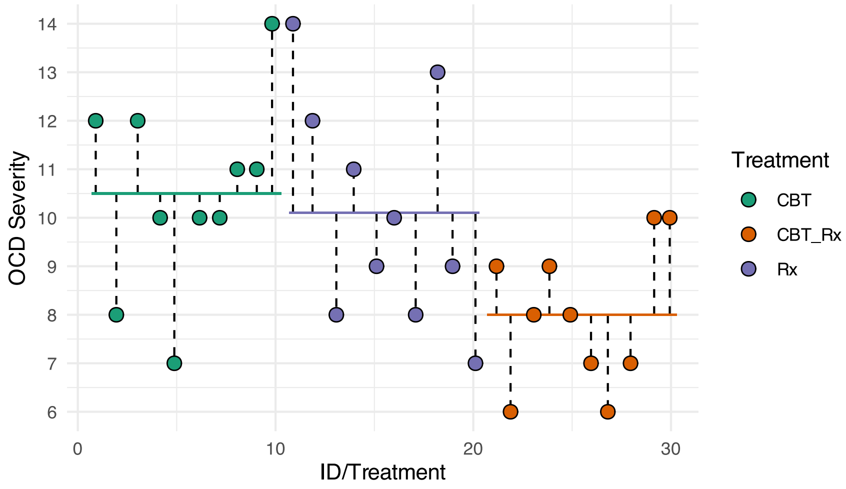
and, thus:

$$SSE = SST - SSB$$

For us:

$$SSE = 141.467 - 36.067 = 105.4$$

This third figure represents SSE. Each dotted lined represents the difference between an individual and their group's mean. The colored lines represent each groups mean.



12.3.4.5 Mean Squares

We must also scale the sum of squares to become variances by dividing by the degrees of freedom. Our df are $df_b = k - 1$ (where k is the number of groups) and $df_e = N - k$. This results in two **Mean Squares**. These are, essentially, variances.

$$MSB = \frac{SSB}{df_b}$$

and

$$MSE = \frac{SSE}{df_e}$$

If you remember the above regarding the F distribution, you may intuitively relate it to the OCD treatment example, wherein:

$$F = \frac{MSB}{MSE}$$

We can determine how likely or unlikely our data are using the F-distribution of df_b and df_w degrees of freedom.

Thus, the above example would results in:

$$MSB = \frac{36.06667}{3-1} = 18.033$$

and

$$MSE = \frac{105.4}{30-3} = 3.904$$

and, thus:

$$F = \frac{18.033}{3.904} = 4.619$$

We will compare this statistic to the distribution to determine how probable our data are compared to the null hypothesis. If our F is large enough to be unlikely given the null, we would reject it. The F statistic is an **omnibus test**.

Omnibus Tests

The F statistic provides the results of an *omnibus test*. These global tests determine the ratio of variance explained by the model versus error. In our F-test, it tells us that at least two of the means differ, but does not tell us which one. Thus, we need to conduct some form of post-hoc (*after the event*) tests.

If the omnibus test is not statistically significant, you can stop there. The groups in the IV do not differ.

We can use an F-distribution table to find out our approximate p-value. The table suggests that critical F for our degrees of freedom is $F(2,27) = 2.51$ for an $\alpha = .05$. Our obtained F is higher ($F_{obtained} > F_{critical}$; $4.62 > 2.51$) and, thus, $p < .05$. This site can calculate exact p-values. This site returns a p-value of $p = .019$.

Our total results can be summarized in what is commonly known as an ANOVA summary table.

Source	Sum of Squares	df	F Value	p-value
Between Groups	36.07	2	4.62	0.019

Source	Sum of Squares	df	F Value	p-value
Within Groups	105.4	27	-	-
Total	141.47	29	-	-

Any statistics program you use for your analyses will provide the appropriate p-value. For example, here is the output from R:

Predictor	SS	df	MS	F	p	partial_eta2	CI_90_partial_eta2
(Intercept)	1102.50	1	1102.50	282.42	.000		
Treatment	36.07	2	18.04	4.62	.019	.25	[.03, .42]
Error	105.40	27	3.90				

12.3.4.6 Effect Size

We can calculate η^2 (eta squared) as an effect size for our ANOVAs. This is simply the ratio of SSB and SST. That is, deviations explained by the model over all deviations. It is an indicator for fit of our model. It ranges from 0 (nothing explained by the model) to 1 (everything explained by the model). It is calculated as:

$$\eta^2 = \frac{SSB}{SST}$$

Typically, the following cut-offs are used:

- η^2= .01 (small effect size)
- η^2= .06 (medium effect size)
- η^2= .14 (large effect size)

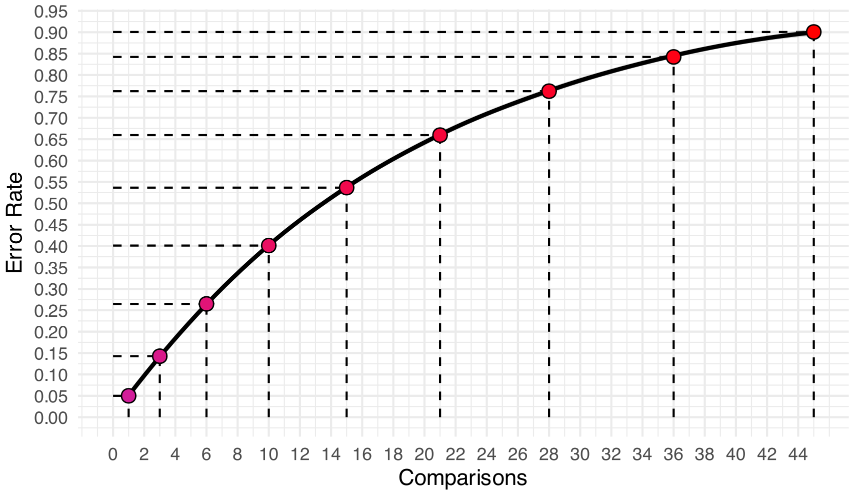
Let's calculate it for our OCD example above. We had $SSB = 36.067$ and $SST = 141.477$. Therefore:

$$\eta^2 = \frac{36.067}{141.477} = .255$$

12.3.4.7 Post-hoc Tests

So our ANOVA revealed a statistically significant results, but it was an omnibus test. Now what? Unfortunately, the results of the omnibus ANOVA test does not inform us *which* groups differ. It simply tells us at least one group differ from at least one other group.

Recall that our family wise error rate increases as we do more statistical tests. So, while we may have set a criterion of $\alpha = .5$, it increases as we do more tests.



which reflect the following error rates:

Groups	Number of Possible Comparisons	Error Rate
2	1	0.05
3	3	0.1426
4	6	0.2649
5	10	0.4013
6	15	0.5367
7	21	0.6594
8	28	0.7622
9	36	0.8422
10	45	0.9006

Initially, we should test only those comparisons with which we hypothesize to be different; doing more increases the likelihood that we commit a Type 1 error. However, sometimes it makes sense to also do exploratory analyses. If we conduct exploratory analyses, we should adjust our error rate to reflect the multiple comparisons.

We will cover one major method of comparing group means: Tukey's Honestly Significant Difference (HSD).

12.3.4.8 Tukey's HSD

Tukey's HSD provides an efficient ways to compare multiple groups simultaneously to determine if their difference is statistically significant. Basically, Tukey's HSD provides a number with which to compare differences in group means. If two groups means' differ by more then Tukey's HSD, then they are statistically different.

Think about it

Tukey's HSD is, essentially, all t-test comparisons using a correction for family-wise error rates.

The formula for Tukey's HSD is:

$$T = q \times \sqrt{\frac{MSE}{n}}$$

where q is a critical q value, MSE is the mean squared error from our ANOVA, n is the sample size per group. Should the difference between the means of two groups exceed T, they are considered statistically different.

From our ANOVA above, we got $MSE = 3.904$ and 10 individuals per group. When we look up the critical q value (for $k = 3$ and $df_e = 27$), we get $q = 3.508$. Thus, Tukey's HSD would be:

$$T = q \times \sqrt{\frac{MSE}{n}} = 3.508 \times \sqrt{\frac{3.904}{10}} = 2.19$$

Thus, we can classify any mean difference beyond 2.19 as statistically significant. That is, if the absolute value of one group's mean subtracted from another group's mean is greater than 2.19, they are statistically significantly different. Our groups are as follows:

```
           diff        lwr       upr     p.adj
CBT_Rx-CBT -2.5 -4.6908019 -0.309198 0.0228923
Rx-CBT     -0.4 -2.5908019  1.790802 0.8936294
Rx-CBT_Rx   2.1 -0.0908019  4.290802 0.0621952
```

💡 Think about it

Look at the confidence intervals in the data above (lwr = LL and upr = UL). Notice which ones contain zero and how that is linked to the respective p-values.

Most statistical software/programs will give you specific p-values for each comparison. For example, in R:

```
           diff        lwr       upr     p.adj
CBT_Rx-CBT -2.5 -4.6908019 -0.309198 0.0228923
Rx-CBT     -0.4 -2.5908019  1.790802 0.8936294
Rx-CBT_Rx   2.1 -0.0908019  4.290802 0.0621952
```

12.3.4.9 Bonferonni Correction

Sometimes you may wish to do some post-hoc comparisons, but without doing Tukey's HSD. You may also apply a Bonferonni correction to the α level to attempt to control for the family wise error. To do so, simply use the following:

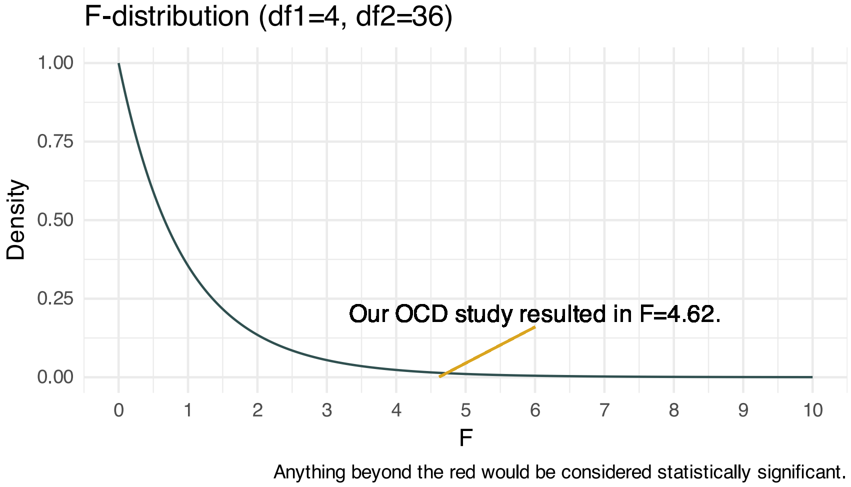
$$\alpha_{new} = \frac{\alpha}{n_{comparisons}}$$

So, imagine we decide post-hoc to conduct two exploratory analyses. If our original $\alpha = .05$, then our new critical value would be:

$$\alpha_{new} = \frac{\alpha}{n_{comparisons}} = .\frac{05}{2} = .025$$

More about the F-distribution

The following figure represents the F-distribution for $df_1 = 2$ and $df_2 = 27$.The red region represent the most extreme 5% of the distribution, which aligns with our pre-determine criteria of $\alpha = .05$.



12.3.5 5. Write your results/conclusions

We conducted a one way ANOVA to determine the association between OCD treatments and OCD severity post-treatment. The results suggest that the group means for OCD severity are unlikely given a true null hypothesis, $F(2, 27) = 4.62$, $p = .019$.

Post-hoc comparisons using Tukey's HSD were conducted to evaluate pairwise differences between the treatment groups. The results indicated a statistically significant difference between the CBT group $(\bar{x} = 10.5)$ and the CBT_Rx group $(\bar{x} = 8.0)$, $\text{mean difference} = -2.5$, $95\%CI[-4.69, -0.31]$, $p_{adj} = .023$. This suggests that participants in the CBT group had significantly higher OCD severity scores compared to those in the CBT_Rx group.

No significant differences were observed between the Rx group $(\bar{x} = 10.1)$ and the CBT group, $\text{mean difference} = -0.4$, $95\%CI[-2.59, 1.79]$, $p_{adj} = .894$, or between the Rx group and the CBT_Rx group $\text{mean difference} = 2.1$, $95\%CI[-0.09, 4.29]$, $p_{adj} = .062$.

12.3.6 ANOVA Model - Part Deux

I recommend returning to this section after you read the chapter on regression.

Surprisingly, you may learn that the ANOVA is the same as regression (both are **general linear models**). In our specific example above with three treatment groups:

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + e_i$$

where y_i is individual i's score on the DV, e_i is the residual/error for individual i, β_n are regression coefficients, and x_{ni} are individual i's score on the n^{th} variable.

What are these variables? For an ANOVA, we would use what is known as **dummy coding**. In dummy coding, there will be k-1 variables, where k is the number of groups. So, in the above example, we have three groups. Thus, we will have $k - 1 = 3 - 1 = 2$ dummy variables. In dummy coding, we choose a reference category, here we can pick CBT, and the other groups are represented by the dummy variables. Thus, the CBT group will score a '0' on all other β variables. The other groups will score '1' on one of the dummy variables. Importantly, each group will be the only one to score '1' on the a specific dummy variable (e.g., only the Rx group scores '1' on β_1. As a result, an individual's score on dummy coded variables will change depending on what group that individual is in.

Additionally, most general linear models require an intercept, which is typically referred to as β_0. We will include an intercept that is present for ALL groups. Thus, each group would score a '1' on the intercept, indicating that it is part of the linear equation for that group. The resulting table could be:

Group	CBT	Rx	CBT+Rx
β_0	1	1	1
β_1	0	1	0
β_2	0	0	1

Dummy Coding

As you can see, each group has a unique pattern on the dummy variables. So how does this fit within the GLM and regression? Well, let's focus on the placebo group. First, recall our means:

Treatment	Mean
CBT	10.5
CBT_Rx	8
Rx	10.1

Let's use our brains to understand how these numbers connect to dummy coding. First, let's consider the CBT group. Our resulting general linear model will be:

$$CBT = \beta_0 + (0)\beta_1 + (0)\beta_2$$

therefore:

$$CBT = \beta_0$$

because anything multiplied by 0 is 0! So this is our simplified equation for the CBT group. Our *best estimate* of the score of any individual in the CBT group is the mean of the CBT group. That is, if knew an individual was in CBT group and noting else about them, our best guess at their OCD severity score would be the mean of the CD severity scores of the CBT group. As a result, β_0 is the mean of the CBT group (here, 10.5)!

What about the Rx group?

$$Rx = \beta_0 + (1)\beta_1 + 0\beta_2$$

therefore

$$Rx = \beta_0 + (1)\beta_1$$

Well, we know that β_0 is the mean of the CBT group (10.5), and we know the mean of the Rx group (10.1). Therefore:

$$Rx = \beta_0 + \beta_1$$

$$10.1 = 10.5 + \beta_1$$

$$\beta_1 = 10.1 - 10.5 = -0.4$$

Interesting. β_1 is simply the difference in means in CBT and Rx group, $\beta_1 = \overline{x}_{Rx} - \overline{x}_{CBT}$. I hope you can intuitively figure out what β_2 represents.

One important consideration here is what group we choose as the reference group (the one with "no" coefficient). This determines which means are being compared through each regression parameter, β.

Think about it

Solve for β_2 in the OCD example.

Answers

It is the mean difference between our reference group (CBT) and the group that is coded 1 for that variable (CBT+Rx): 2.5.

The difference between Rx and CBT+Rx is not represented in dummy coding. We could calculate it by making either of those the reference group.

We can also use the model to solve for an individual's residual (error) score. Assuming we are interested in person 7 (from the CBT group; y_7). From the above equation, we get:

$$y_7 = \beta_0 + (0)\beta_1 + (0)\beta_2 + e_7$$

Which simplifies to:

$$y_7 = \beta_0 + e_7$$

because their scores are 0 on each dummy variable. Therefore, individual 7's score is simply $\beta_0 + e_7$. Interestingly, β_0 is simply the mean of the CBT group and e_7 is that person's deviation from the mean of the CBT group. Here:

$$10 = 10.5 + e_7$$

so

$$e_7 = 10 - 10.5 = -0.5$$

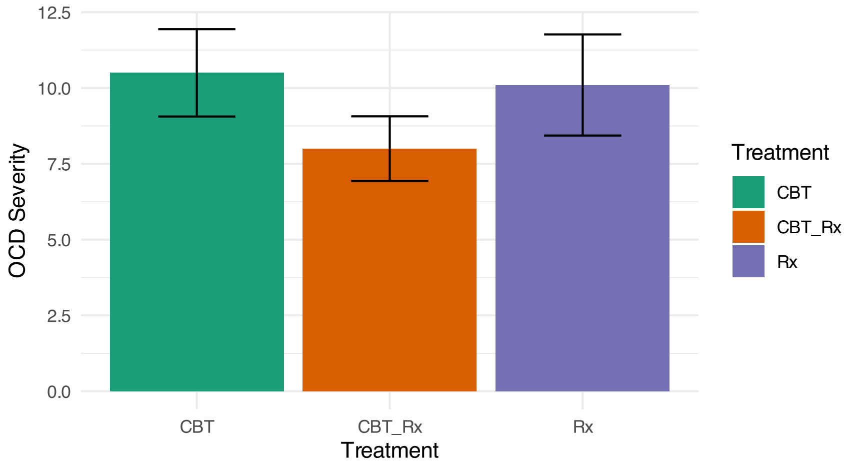
The person's residual score is −0.5, which is the difference between their actual score and their predicted score using our model. Remember, our best guess at their score is the mean.

When we cover regression, you will like see additional similarities in ANOVA and regression. All of the analysis we will cover are linked to the general linear model.

12.3.7 Plotting the ANOVA Results

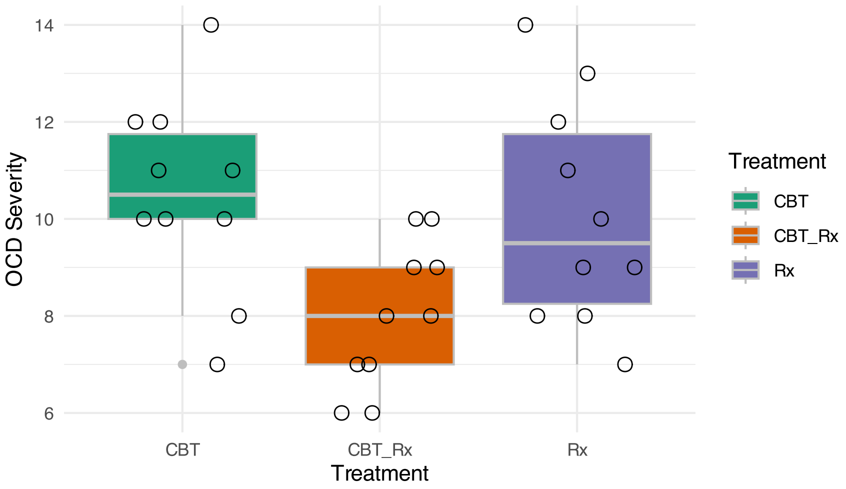
There any many ways you can visualize the results of an ANOVA. One way that is not recommended is the 'dynamite plot'.

Dynamite plot:



Bars represent the 95% CI for each mean.

How should we visualize it? There are many ways, but box plots may be a better method. It gives more details about the data.



12.4 Conclusion

This chapter covered the one-way ANOVA, a statistical method used to determine if there is a significant difference among the means of three or more independent groups. The one-way ANOVA extends the t-test to multiple groups and uses an F-statistic to compare the variance between group means to the variance within groups.

12.5 ANOVA in R

If you are a PSYC3950 student, you are not *required* to know how to calculate ANOVA in R and can skip this section. You will cover this analysis using SPSS in the lab section of the course. If you are so inclined (good for you), continue…

R can easily run the ANOVA and provide an ANOVA summary table. I particularly like the `apa.aov.table()` function from the `apaTables` library. It can quickly provide summary statistics:

```
library(apaTables)
#I named my data dat_ocd
one_way_anova_example <- aov(OCD_Severity~Treatment,
data=dat_ocd)
apa.1way.table(iv=Treatment, dv=OCD_Severity, data=dat_ocd)
```

```


Descriptive statistics for OCD_Severity as a function of
Treatment.

 Treatment     M   SD
       CBT 10.50 2.01
    CBT_Rx  8.00 1.49
        Rx 10.10 2.33

Note. M and SD represent mean and standard deviation,
respectively.
```

and an ANOVA summary table, with effect size:

```
apa.aov.table(one_way_anova_example)
```

```


ANOVA results using OCD_Severity as the dependent variable


   Predictor      SS df      MS      F    p partial_eta2
CI_90_partial_eta2
 (Intercept) 1102.50  1 1102.50
282.42 .000
   Treatment   36.07  2   18.04   4.62 .019          .25
[.03, .42]
       Error  105.40 27
3.90

Note: Values in square brackets indicate the bounds of the 90%
confidence interval for partial eta-squared
```

Tukey's HSD in R

Tukey's HSD can be found in the `stats` library. You simply use your ANOVA model as the main argument.

```
library(stats)
mod_aov <- aov(OCD_Severity ~ Treatment, data=dat_ocd)

TukeyHSD(mod_aov) #I called our ANOVA model 'mod_aov'
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = OCD_Severity ~ Treatment, data = dat_ocd)

$Treatment
           diff        lwr       upr     p adj
CBT_Rx-CBT -2.5 -4.6908019 -0.309198 0.022892
Rx-CBT     -0.4 -2.5908019  1.790802 0.893629
Rx-CBT_Rx   2.1 -0.0908019  4.290802 0.062195
```

You would interpret the last column, *p adj* as usual. The values have been adjusted to account for family wise error rates.

12.6 Practice Question

You are a sport psychologist testing if the type of drink improves performance. You provide three drinks to sprinters and measure the time to run 100m. The following data are obtained.

ID	Drink	Speed
1	Water	9
2	Water	15
3	Water	16
4	Water	18
5	Water	8
6	Milk	19
7	Milk	19

ID	Drink	Speed
8	Milk	18
9	Milk	25
10	Milk	20
11	Gatorade	13
12	Gatorade	12
13	Gatorade	19
14	Gatorade	8
15	Gatorade	15

Conduct an ANOVA to determine if any groups differ from any others. Specifically:

- Write the null and alternative hypotheses
- Conduct appropriate omnibus test
- Conduct post-hoc tests.
- Calculate an effect size.
- Write a results section

12.7 Answers

Hypotheses

$H_0 = \mu_{water} = \mu_{milk} = \mu_{gatorade}$

$H_1 =$ at least one different.

Omnibus

Means and F Table

```


Descriptive statistics for Speed as a function of Drink.
```

```
    Drink     M   SD
 Gatorade 13.40 4.04
     Milk 20.20 2.77
    Water 13.20 4.44

Note. M and SD represent mean and standard deviation,
respectively.
```

```


ANOVA results using Speed as the dependent variable



   Predictor     SS df     MS     F    p partial_eta2
CI_90_partial_eta2
 (Intercept) 897.80  1 897.80
61.63 .000
       Drink 158.80  2  79.40  5.45 .021          .48
[.05, .64]
       Error 174.80 12
14.57

Note: Values in square brackets indicate the bounds of the 90%
confidence interval for partial eta-squared
```

Results

```
The ANOVA (formula: Speed ~ Drink) suggests that:

  - The main effect of Drink is statistically significant and
large (F(2, 12) =
5.45, p = 0.021; Eta2 = 0.48, 95% CI [0.07, 1.00])

Effect sizes were labelled following Field's (2013)
recommendations.
```

Tukey’s HSD

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Speed ~ Drink, data = sprint)

$Drink
               diff        lwr       upr     p adj
Milk-Gatorade   6.8   0.36018 13.23982 0.038411
Water-Gatorade -0.2  -6.63982  6.23982 0.996224
Water-Milk     -7.0 -13.43982 -0.56018 0.033147
```

Practice Question

Solve for individual 12 in the OCD example, who is part of the Rx group and who's OCD Severity score is 12.

Answers

$$y_{12} = \beta_0 + (1)\beta_1 + (0)\beta_2 + e_{12}$$

Therefore:

$$y_{12} = \beta_0 + (1)\beta_1 + e_{12}$$

Filling in our beta's from above"

$$12 = 10.5 + (1)(-0.4) + e_{12}$$

Therefore:

$$e_{12} = 12 - 10.5 + 0.4 = 1.9$$

13 Repeated ANOVA

This chapter will cover the repeated measures ANOVA, a statistical method used to determine if there is a significant difference among the means of three or more *related* or dependent groups. Unlike a one-way ANOVA, which involves independent groups, a repeated measures ANOVA is used when the same participants are measured under different conditions or at different time points. The repeated measures design accounts for the correlation between the measurements from the same participants. For example, if we compare the heights of various children over time, we would assume that children who are tall will grow into adults who are tall, and children who are short will grow into adults who are short. Or, as another example, imagine we test the impact of various sports drink on speed. If we gave the same participant three different sports drinks and tested their running speed after each, that person's scores would be dependent or correlated. Fast people will typically be fast and slower people will typically be slow, regardless of the drink given. This **dependence** or **correlation of observations** is why the working with repeated measures data requires specific assumptions and analysis.

13.1 Some Additional Details

The repeated measures ANOVA is appropriate when there is one independent variable (IV) with three or more levels (e.g., conditions or time points) and one continuous dependent variable (DV). For example, researchers might use a repeated measures ANOVA to compare test scores (DV) across three different time points (IV with three levels) or to assess the effects of a three different treatments (IV with three levels) on pain (DV). Importantly, the same participants are measured at each level of the IV.

💡 Think about it

Sometimes a repeated measures design can be used on different people, but when they are **matched** on various important characteristics. For example, we may match a group of students based on age, gender, IQ, and academic achievement. In this chapter and course we will assume the same people measured across conditions or time.

The null hypothesis for the repeated measures ANOVA posits that all group means are equal across different conditions or time points:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

where k represents the number of conditions or time points. The alternative hypothesis states that at least one group mean is different:

$$H_1 : \text{At least one } \mu_i \text{ differs from another}$$

By rejecting the null hypothesis, researchers conclude that significant differences exist among the conditions or time points–at least one differs from the others. Similar to a one-way ANOVA, post hoc tests are often required to identify where these differences occur.

13.2 Key Assumptions

A repeated measures ANOVA can be conducted under certain assumptions. These include:

1. The data are continuous

The dependent variable should be at the interval or ratio level.

2. Sphericity

Sphericity is a key assumption in repeated measures ANOVA and is closely related to equality of variances in a one-way ANOVA. It refers to the equality of variances of the *differences* between all possible pairs of conditions or time points. For example, consider the following scores representing five people who were all given three different interventions for depressive symptoms.

ID	Treatment.1	Treatment.2	Treatment.3
1	6	10	11
2	8	11	17
3	7	14	15
4	10	16	23
5	13	15	21

We can calculate the differences (one subtracted another) of **each** pair of scores per person. For example, consider treatment 1 and 2:

ID	Treatment.1	Treatment.2	Treatment.3	Difference_1_2
1	6	10	11	−4
2	8	11	17	−3
3	7	14	15	−7
4	10	16	23	−6
5	13	15	21	−2

We can calculate a variance of these difference scores, which would be 4.3. Well, the test of sphericity tests if the variance between ALL possible difference scores (here, $1-2$, $2-3$, and $1-3$) is similar. The repeated measures ANOVA requires these variances be similar; Mauchley's test of sphericity is the formal test for this. For example, the variances of the differences are:

1. Treatment 1 - Treatment 2 = 4.3
2. Treatment 2 - Treatment 3 = 8.7
3. Treatment 1 - Treatment 3 = 8.3

So, while the variances quite obviously are not equally, the sphericity test is a formal test to ensure we meet or do not meet this assumption.

In simpler terms, sphericity assumes that the relationship (or correlation) between scores in different conditions is consistent. This assumption is unique to repeated measures designs because the same participants contribute data across multiple conditions, introducing dependencies in the data. If the assumption of sphericity is violated, the repeated measures ANOVA can lead to inflated Type I error rates, meaning there's a higher chance of incorrectly rejecting the null hypothesis.

Mauchly's Test of Sphericity

Sphericity can be tested using **Mauchly's test of sphericity**. This test evaluates whether the variance of the differences between groups is equal. This test has the following hypotheses:

- **Null hypothesis** (H_0): The variances of the differences between all pairs of conditions are equal (sphericity holds).
- **Alternative hypothesis** (H_1): The variances of the differences are not equal (sphericity is violated).

If Mauchly's test is significant ($p < .05$), sphericity is violated. When this happens, adjustments to the degrees of freedom (df) are required to make the test more robust. Specifically, when sphericity is violated, the **F-statistic** becomes unreliable. To address this, two common corrections can be applied:

1. Greenhouse-Geisser Correction: This conservative adjustment reduces the degrees of freedom based on an estimated **epsilon** (ϵ) value, which quantifies the degree of violation. Lower values of ϵ indicate greater violations of sphericity (with $\epsilon = 1$ indicating perfect sphericity).
2. **Huynh-Feldt Correction:** This is a less conservative adjustment compared to Greenhouse-Geisser. It adjusts the degrees of freedom based on a different estimation of ϵ, which can be slightly larger.

When reporting results of the repeated measures ANOVA, it's standard to use corrected degrees of freedom and note which correction was applied–when applicable. ANOVA results will typically inform you which degrees of freedom go with which correction. Sometimes the degrees of freedom may not be whole numbers. Don't panic; you haven't messed up your analyses!

3. Homogeneity of variances

The variance within each group should be approximately equal. This assumption is less of a concern than in a one-way ANOVA, but it's still important to check. As with one-way ANOVA, Levene's test can be used to assess this.

4. Normality of residuals

As with one-way ANOVA, it's the residuals (the differences between the observed and predicted values) that should be normally distributed. This can be assessed visually using a Q-Q plot or by conducting a Shapiro-Wilks test on the residuals.

We will now go through a comprehensive example of a research project that requires the use of a repeated measures ANOVA.

13.3 Remember Me?

You are hired by the Reach Out Center for Kids (ROCK) as a developmental researcher as part of their cognitive development team. You are tasked with conducting research investigating the changes in memory processes across early childhood. Specifically, you believe that the amount of 'chunks' of memory a child can retain increases as children grow. You decide to develop a memory test to assess any changes over time. After reading the literature, you and your team believe that memory will improve as children develop. Importantly, you believe that notable changes will be evident between children when they are 4, 6, and 8-years old.

13.3.1 1. Generating hypotheses

The main null and alternative hypotheses for this repeated measures ANOVA can be converted into a statistical hypothesis stated as (for the null):

$$H_0 : \mu_{4yo} = \mu_{6yo} = \mu_{8yo}$$

And (for the alternative):

$$H_1 : \text{At least one } \mu_i \text{ differs from the others.}$$

13.3.2 2. Designing a study

You and your team plan out a research study. The method follows:

Participants: Participants were recruited from local schools near ROCK. Posters were created in collaboration with the school board, ensuring all parties agreed on recruitment materials. Eligible participants were children who were typically developing and had no reported neurological or developmental disorders. Children were tested when they are aged 4, 6, and 8 years old.

A power analysis was conducted using your literature review and indicated that a total of 12 children will be needed to achieve a power of $1 - \beta = .8$.

Materials: Tyler's Memory Test (TMT; Pritchard, 2024) was used to assess memory. The TMT is a child-friendly memory that measures children's total overall memory. The tasks involved recalling: items from a story presented with accompanying pictures, series of digits, and abstract shapes through drawing. The test is standardized and compared children's scores to same-aged peers. Memory performance was scored out of 10. The TMT has shown suitable reliability and validity (Pritchard, 2024).

Procedure: Posters advertisements were shared on ROCK's website, in addition to the researchers' social media pages. The poster focused on caregivers of 4-year olds and indicated they could participate in a study on memory. Interested caregivers were provided an informed consent form and, once consenting, completed a brief screener to ensure children did not meet criteria for neurological or other psychological disorder. The resulting participants completed three memory sessions (at age 4, 6, and 8 years old) at ROCK's testing center. All testing was completed by PhD-level psychologists. Parents provided debriefed after each testing session.

The ethics review board at Grenfell Campus reviewed the project and ethics submission and approved the study.

13.3.3 3. Collecting data

The study was completed as described; a final sample size of 8 was used. The following data were obtained:

ID	Age 4	Age 6	Age 8
1	3	5	7
2	4	3	8
3	3	5	5
4	3	3	8
5	4	4	7
6	4	4	9
7	1	4	7
8	2	3	8

Short and long form data

The data above is in short form–think from top to bottom. In most analyses, long form data is preferred. Each analysed unit of data should have a row. In the above, each row has three units bring analysed; there are three memory scores per row. While it might make sense to have each child represented as a row, it's actually better to have each experimental unit (a memory score) be a row. Here's how the data would look in long form:

ID	Age	Memory Score
1	Age 4	3
1	Age 6	5
1	Age 8	7
2	Age 4	4
2	Age 6	3
2	Age 8	8
3	Age 4	3

ID	Age	Memory Score
3	Age 6	5
3	Age 8	5
4	Age 4	3
4	Age 6	3
4	Age 8	8
5	Age 4	4
5	Age 6	4
5	Age 8	7
6	Age 4	4
6	Age 6	4
6	Age 8	9
7	Age 4	1
7	Age 6	4
7	Age 8	7
8	Age 4	2
8	Age 6	3
8	Age 8	8

13.3.4 4. Analyzing data

Our Model

In previous examples of ANOVA, we have had different individuals for each level or condition. Recall that in the one way ANOVA example, each individual received **one** type of therapy. However, sometimes it makes sense to put the same individuals in each condition to assess change or differences **within** the individuals. Repeated measures do just that.

As such, our model will look similar:

$$memory = age + error$$

and for each individual:

$$y_i = age_i + e_i$$

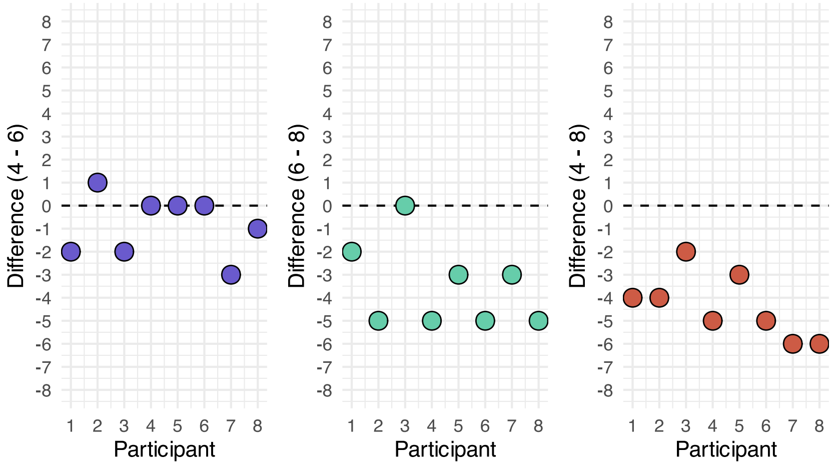
13.3.4.1 Testing Assumptions

Sphericity

To allows us to continue with F-tests, we must test the assumption of sphericity. Recall that this assumption purports that the variance of the *differences* between all conditions is the same.

An easy way to visualize this is by plotting difference scores. In our example, we will have three difference scores (i.e., age 4 - age 6; age 6 - age 8; age 4 - age 8).

ID	Age 4	Age 6	Age 8	four_minus_six	six_minus_eigh	four_minus_eigh
1	3	5	7	−2	−2	−4
2	4	3	8	1	−5	−4
3	3	5	5	−2	0	−2
4	3	3	8	0	−5	−5
5	4	4	7	0	−3	−3
6	4	4	9	0	−5	−5
7	1	4	7	−3	−3	−6
8	2	3	8	−1	−5	−6



In the visualization each person is represented as a dot (the x-axis). The y-axis represents each person's difference score across the various IV levels. We want to look at the dispersion of the dots along the y-axis, not how close they to the lines. The spread of the points should appear similar across the group differences. The variance of each of the differences is:

- Four - Six: 1.839
- Six - Eight: 3.429
- Four - Eight: 1.982

We can test the assumption using **Mauchly's test of Sphericity**, which hypothesizes (for a three condition repeated measures deign):

$$H0 : \sigma^2_{A-B} = \sigma^2_{A-C} = \sigma^2_{B-C}$$

and

$$H1 : \text{at least one var not equal}$$

We will not be concerned with the formal calculations of Mauchly's test; rather, our statistical software can conduct it for us.

For our data:

```
  Effect        W        p p<.05
2    Age 0.823509 0.558478
```

Recall that the null hypothesis is that the variances are equal; thus, we **want** $p > .05$ for Mauchly's test, although it's not a complete deal-breaker if we violate this assumptions. Regardless, our results indicate that we have **not** violated this assumption and can proceed as intended.

Writing up Mauchly's Test

We used Mauchly's test to check the assumption of sphericity and the results indicate that the assumption is not violated, $p = .558$.

The Assumption is Violated: Now What?

You can apply two corrections to the data that account for violations of sphericity. These are the **Greenhouse-Geisser** or **Huynh-Feldt** corrections.

As we have done in the last two chapters, we will partition the various into various sub-components to determine the appropriate F statistic. The following holds:

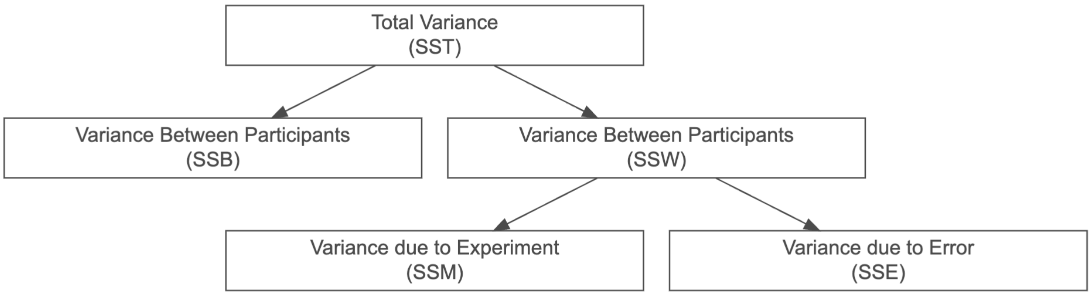

Total Variance
(SST)
Variance Between Participants
(SSB)
Variance Between Participants
(SSW)
Variance due to Experiment
(SSM)
Variance due to Error
(SSE)

Figure 9: Flowchart

You may recall that for independent ANOVAs the individuals in each condition were different. For repeated measures, the individuals will cut across all conditions. So why would they score differently on the same dependent variable? From the figure above, some of the differences may be due to the experiment, while others are just error. It may be helpful to re-conceptualize how we consider variance as the variance between and the variance within an individual. Because all people are in all conditions, changes within an individual can be attributed to the experimental condition and some error.

Let's calculate some of these and it may help them make sense.

13.3.4.2 SST

Our total sum of squares is no different than a one way ANOVA.

$$SST = \sum_{i=1}^{n} \left(x_i - \overline{x}_{grand}\right)^2$$

with $N - 1$ degrees of freedom.

Also, if you know the variance, it can be calculated as:

$$SST = s^2_{overall}(N - 1)$$

Our variance in all scores is 4.717 with $n = 24$. Thus:

$$SST = 4.717(24 - 1) = 108.49$$

13.3.4.3 SSW

Here we will depart from our independent ANOVA method. We will calculate the SSW by looking at the deviations **within** individuals (rather than within groups, which was error in the independent ANOVAs). Recall our data:

ID	Age 4	Age 6	Age 8
1	3	5	7
2	4	3	8
3	3	5	5
4	3	3	8
5	4	4	7
6	4	4	9
7	1	4	7
8	2	3	8

So, let's consider individual 1. Their mean score is $\frac{3+5+7}{3} = 5$. And their deviations are:

$$SS_{x_{i=1}} = (3-5)^2 + (5-5)^2 + (7-5)^2 = 8$$

We do this across **all** individuals! The resulting formula is expressed as:

$$SSW = \sum_{i=1,t=1}^{n} (x_{it} - \overline{x}_i)^2$$

where x_{it} is the score for individual i at time t and $\overline{x}_i$ is the mean for individual i across all conditions. If you can quickly get the variances, you could also use the formula:

$$SSW = \sum_{i=1}^{n} s_i^2(n_t - 1)$$

For us, we have:

ID	Variance
1	4

ID	Variance
2	7
3	1.333
4	8.333
5	3
6	8.333
7	9
8	10.333

and thus, because each individual has three time points:

$$SSW = 4(2) + 7(2) + 1.33(2) + 8.33(2) + 3(2) + 8.33(2) + 9(2) + 10.33(2) = 102.64$$

13.3.4.4 SSM

The variance of the model, SSM, which is **between groups (i.e., experimental conditions)** is calculated the same way as before).

$$SSM = \sum_{j=1}^{n_j} n_j \left(\overline{x}_j - \overline{x}_{overall}\right)^2$$

For us, the means are:

Age	Mean	n
Age 4	3	8
Age 6	3.875	8
Age 8	7.375	8

Therefore, because we know our grand mean is 4.75:

$$SSM = 8(3.00 - 4.74)^2 + 8(3.875 - 4.74)^2 + 8(7.375 - 4.74)^2 = 85.74$$

13.3.4.5 SSE

Our error is calculated by removing the SSM from within individuals. Remember, individual scores vary because of the experimental conditions (i.e., SSM) and due to error (i.e., random individual fluctuations). Thus, the error can be calculated by subtracting SSM from SSW.

$$SSE = SSW - SSB$$

$$SSE = 102.64 - 85.74 = 16.90$$

Perhaps now you see an added benefit to repeated measures designs. We have effectively reduced our error term, which will reduce the mean squared error, which should increase our F statistic.

13.3.4.6 Mean Squares

Our mean squares are calculated the same as before. However, our df_e is calculated by $df_e = df_w - df_b$, where $df_w = n_i(df_b)$. We have eight individuals with $df_b = 2$, therefore $df_w = 8(2) = 16$ and $df_e = 16 - 2 = 14$

$$MSB = \frac{SSB}{df_b}$$

$$= \frac{85.74}{2} = 42.87$$

and

$$MSE = \frac{SSE}{df_e}$$

$$= \frac{16.90}{14} = 1.207$$

13.3.4.7 F Statistic

Our F statistic is calculated the same way as before, a ratio of MSB and MSE.

$$F = \frac{MSB}{MSE} = \frac{42.87}{1.207} = 35.52$$

We can use an F-distribution table to find out our approximate p-value. We determine that $F_{crit}(2, 14) = 3.7389$.

However, remember, an omnibus ANOVA does not tell us *where* the differences are. We have three groups, so we must conduct post-hoc analysis. We looked at this in the one way and factorial ANOVA, so please refer there.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = `Memory Score` ~ Age, data = dat_child_long)

$Age
             diff       lwr     upr    p adj
Age 6-Age 4 0.875 -0.436746 2.18675 0.235571
Age 8-Age 4 4.375  3.063254 5.68675 0.000000
Age 8-Age 6 3.500  2.188254 4.81175 0.000003
```

As you can see, it seems that memory at age eight ($\overline{x}_{age8} = 7.38$, $SD = 1.19$) is higher than both ages four ($\overline{x}_{age4} = 3.00$, $SD = 1.07$, $p < .001$) and six ($\overline{x}_{age4} = 3.88$, $SD = 0.84$, $p < .001$). However, memory at age four did not differ than at age six ($p = .236$).

13.3.4.8 Effect Size

Effect sizes for repeated measures ANOVA are more difficult to calculate by hand. Specifically, we may use *generalized eta squared* (η^2_g) to account for our repeated measures.

We can get this from statistical software. For this example:

```
# Effect Size for ANOVA (Type I)

Group  | Parameter | Eta2 (generalized) |       95% CI
-----------------------------------------------------
Within |       Age |               0.79 | [0.57, 1.00]

- Observed variables: All
- One-sided CIs: upper bound fixed at [1.00].
```

Thus, differences in memory across ages would be classified as a large effect, $\eta^2_g = .79$, $95\% CI[.57, 1.00]$.

13.3.5 5. Write your results/conclusions

Recall your hypothesis from above: children's scores on the [memory] test will improve as they grow over time.

We conducted an ANOVA to test whether age has an affect on a child's memory. We used Mauchly's test to check the assumption of sphericity and the results indicate that the assumption is not violated, $p = .558$. The results of our omnibus ANOVA suggest that age has a strong and statistically significant effect on a child's memory, $F(2, 14) = 35.48$, $\eta_g^2 = .79$, $95\% CI[.63, 1.00]$, $p < .001$.

Post-hoc results indicated that memory at age eight ($\overline{x}_{age8} = 7.38$, $SD = 1.19$) is higher than both ages four ($\overline{x}_{age4} = 3.00$, $SD = 1.07, p < .001$) and six ($\overline{x}_{age4} = 3.88$, $SD = 0.84$, $p < .001$). However, memory at age four did not differ than at age six ($p = .236$).

13.4 Conclusion

This chapter covered the repeated measures ANOVA, a statistical method used to determine if there is a significant difference among the means of three or more *related* or dependent groups. Repeated measures ANOVA is used when the same participants are measured under different conditions or at different time points, which accounts for the correlation between the measurements from the same participants.

13.5 Repeated Measures ANOVA in R

We can use the same `ez` library to conduct our repeated measures ANOVA in R. Our data will need to be in **long** format, with each *measurement* having a row as opposed to each *individual*. The following data is in **long** format.

ID	Age	Memory Score
1	Age 4	3
1	Age 6	5
1	Age 8	7
2	Age 4	4

ID	Age	Memory Score
2	Age 6	3
2	Age 8	8
3	Age 4	3
3	Age 6	5
3	Age 8	5
4	Age 4	3
4	Age 6	3
4	Age 8	8
5	Age 4	4
5	Age 6	4
5	Age 8	7
6	Age 4	4
6	Age 6	4
6	Age 8	9
7	Age 4	1
7	Age 6	4
7	Age 8	7
8	Age 4	2
8	Age 6	3
8	Age 8	8

As you can see, each individual has three rows, one for each time of assessment.

The `ezANOVA()` function will be used. It will automatically conduct Mauchly's test because it picks up we have a 'within' factor:

```
$ANOVA
  Effect DFn DFd       F           p p<.05      ges
2    Age   2  14 35.4828 3.29759e-06     * 0.790323

$`Mauchly's Test for Sphericity`
  Effect        W        p p<.05
2    Age 0.823509 0.558478
```

```
$`Sphericity Corrections`
  Effect       GGe       p[GG] p[GG]<.05     HFe       p[HF]
p[HF]<.05
2    Age 0.849986 1.52556e-05         * 1.09431 3.29759e-06
*
```

13.6 Practice Questions

You are a educational psychologist testing the efficacy of a new reading program for children who are at-risk for developing a reading disorder. Because assessments are time-consuming, expensive, and with a long wait-list, you are asked to implement a program ASAP and determine it's efficacy. You develop a program based in the literature and hypothesize a significant improvement in reading ability. You measure reading ability (a measurement that uses t-scores) prior to starting the program, 1 month after being in place, 2 months after being in place (the conclusion of the program), and 3 months (one month after conclusion).

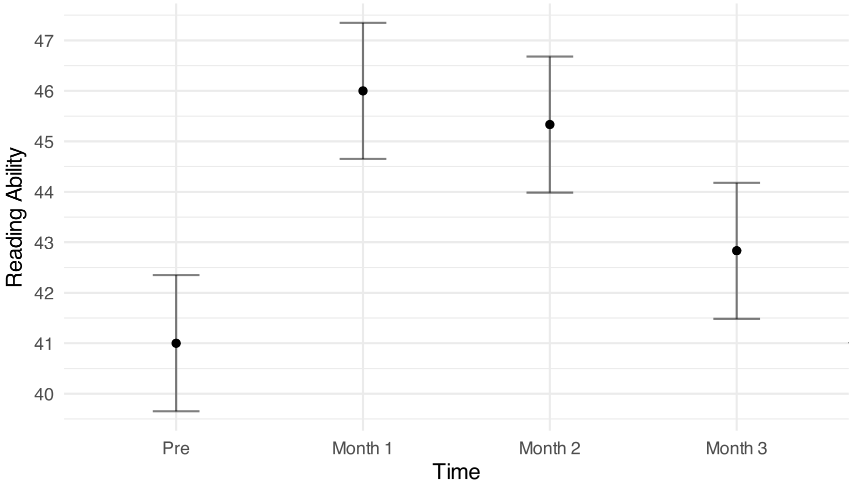
You recruit 6 individual for the program and obtain the following data:

ID	T0_Month	T1_Month	T2_Month	T3_Month
1	38	46	42	42
2	44	51	52	47
3	48	53	50	51
4	39	42	45	35
5	40	42	41	39
6	37	42	42	43

13.7 Answers

```
$ANOVA
  Effect DFn DFd        F           p p<.05      ges
```

```
2   Time   3  15 6.65605 0.00447067     * 0.166897

$`Mauchly's Test for Sphericity`
  Effect        W        p p<.05
2   Time 0.377146 0.615089

$`Sphericity Corrections`
  Effect      GGe     p[GG] p[GG]<.05      HFe      p[HF]
p[HF]<.05
2   Time 0.622618 0.0170215         * 0.979851 0.00479621
*
```



Additional Readings

- Lakens (2013)
- Olejnik & Algina (2003)

14 Factorial ANOVA

This chapter will cover the factorial ANOVA, a statistical method used to determine whether there are differences among group means when there are two or more independent variables (IVs). We have covered the one-way ANOVA, which examines the effect of a single IV on a dependent variable (DV). A factorial ANOVA expand this analyses and allows researchers to study the main effects of each IV and the interaction effect between them. Interactions are when the effect of one IV on the DV depends on the level of another IV. The IVs are also sometimes called **factors**.

Factorial ANOVAs may seem more complex, but this is necessary to model the complexity of the world we live in. Rarely are psychological phenomenon impacted by a single variable. Instead, these phenomenon are likely impacted by multiple interacting factors. For example, imagine we want to examine the impact of teaching method (lecture vs. group discussion) and class size (note that class size could be a ratio variable, but we will consider it as two categories: small [class size less than 20] vs. large [20+]) on student performance. Here, teaching method and class size are two IVs (both categorical), and student performance is the DV. A factorial ANOVA would tell us whether:

- Teaching method alone influences performance (main effect of teaching method),
- Class size alone influences performance (main effect of class size),
- The effect of teaching method depends on class size (interaction effect).

14.1 Some Additional Details

A factorial ANOVA is appropriate when there are two or more categorical/nominal independent variables, each with two or more levels, and one continuous dependent variable. For example, researchers might use a 2×3 factorial ANOVA to examine the effects of diet type (IV1; vegetarian vs. non-vegetarian) and exercise frequency (IV2; none, moderate, high) on weight loss (DV). This design allows for a more comprehensive understanding of how multiple factors work together to influence outcomes.

> Think about it
>
> It is possible to have both between and within (repeated) IVs in ANOVA. When you have 2+ IVs that are between subjects, we call it a *factorial ANOVA*. When there are 2+ within/repeated factors, it is called a *fully repeated measures ANOVA* or a *within-within ANOVA*. When there are both between and within factors, it is called a *mixed ANOVA*.
>
> This chapter covers factorial ANOVA and we will cover mixed ANOVA in the next chapter. Fully repeated measures ANOVA are not in the current edition, but may be added in a later edition.

What the heck is a 2x2 ANOVA?

You will encounter "2x2", "3x2x2", "2x3x4", or any other combination of numbers in the context of ANOVA. It is quite simple to understand. The number of numbers indicates the number of factors, while the numbers themselves represent the number of levels within each factors.

For example, a 2x3 ANOVA has two factors (two IVs, which means two main effects), because there are two numbers: a 2 and a 3. The first factors has two levels and the second factor has three levels.

As another example, consider a 4x3x2 ANOVA. In this analysis there are three factors (three IVs, which means two main effects). The first has four levels, the second has three levels, and the third has two levels.

The null hypothesis for a factorial ANOVA posits that there are no main effects or interaction effects among the independent variables. In other words, the population means are equal across all combinations of factor levels:

$$H_0 : \text{All group means are equal across the levels of Factor A and Factor B}$$

More commonly, and probably easier to understand, a two-way factorial ANOVA typically involves three separate null hypotheses:

1. Main effect of Factor A:

$$H_{0A} : \mu_{A1} = \mu_{A2} = ... = \mu_{An}$$

Here, we are proposing that the means do not differ across levels (from levels 1 to n) of Factor A.

2. Main effect of Factor B:

$$H_{0B} : \mu_{B1} = \mu_{B2} = ... == \mu_{Bm}$$

Here, we are proposing that the means do not differ across levels (from 1 to m) of Factor B.

Think about it

Why does one factor have an n number of levels and another have m levels? Because the IVs can have a different number of levels, the term n and m are used. One IV can have 2 levels ($n = 2$) and the other IV can have a different number of levels such as 3 ($m = 3$).

3. Interaction effect (A × B):

$$H_{0AB} : \text{No interaction between A and B}$$

Here, we are proposing that the effect of Factor A does not depend on Factor B.

The alternative hypotheses state that:

1. At least one main effect exists (Factor A or Factor B influences the DV)
2. There is an interaction effect (the effect of one factor depends on the other).

Rejecting any of these null hypotheses indicates significant differences. When significant effects are found, post-hoc tests or simple effects analyses are often required to determine where these differences occur. More to some.

14.2 Key Assumptions

Many of the assumptions in the factorial ANOVA overlap with our previous chapters. Regardless, a factorial ANOVA can be conducted under the following assumptions:

1. The data are continuous

The dependent variable should be measured at the interval or ratio level.

2. Independence of observations

Each observation should be independent of the others. This is critical because factorial ANOVA typically involves different participants in each group.

3. Homogeneity of variances

The variance within each combination of factor levels should be approximately equal. This can be assessed using Levene's test.

4. Normality of residuals

The residuals (differences between observed and predicted values) should be approximately normally distributed. This can be checked visually (e.g., Q-Q plot) or with tests like Shapiro-Wilk.

We will now go through a comprehensive example of a research project that requires the use of a factorial ANOVA.

14.3 Rate My Physician

You are hired by the regional health authority to conduct research regarding patient-provider satisfaction. You consult the literature and theorize that female physicians are more compassionate and that patients will rate visits with female physicians higher than male physicians. Furthermore, theory suggests that men are more comfortable around female physicians than male physicians and will, thus, rate these visits as more satisfactory.

14.3.1 1. Generating hypotheses

You hypothesize that:

1. Participants will be more satisfied with visits from female physicians compared to male physicians.
2. There will be an interaction between patient and provider gender, such male patients will rate females physicians higher than male physicians, but female patients will not have such a pattern.

Translating Hypotheses

When translating these conceptual hypotheses into statistical hypotheses, it may be helpful to think of each group as a cell on a contingency table. Each group will have a mean. We will consider patient gender as the first subscript below the mean (e.g., $\mu_{m.}$ or $\mu_{f.}$) and provider gender as the second subscript below the mean (e.g., $\mu_{.m}$ or $\mu_{.f}$). A dot will indicate the collapsed grouping of a specific factor. Given this:

- $\mu_{m.}$ is the mean satisfaction rating of all male patients ($n = 20$)
- $\mu_{f.}$ is the mean satisfaction rating of all female patients ($n = 20$)
- $\mu_{.m}$ is the mean satisfaction rating of all patients who saw a male physician ($n = 20$)
- $\mu_{.f}$ is the mean satisfaction rating of all patients who saw a female physician ($n = 20$)
- μ_{mm} is the mean satisfaction rating for male patients who saw a male physician ($n = 10$)
- μ_{mf} is the mean satisfaction rating for male patients who saw a female physician ($n = 10$)
- μ_{fm} is the mean satisfaction rating for female patients who saw a male physician ($n = 10$)
- μ_{ff} is the mean satisfaction rating for female patients who saw a female physician ($n = 10$)

Thus, our statistical hypotheses that align with conceptual hypotheses above are:

1. Participants will be more satisfied with visits from female physicians compared to male physicians.

$$H_0 : \mu_{.f} = \mu_{.m}$$

$$H_1 : \mu_{.f} \neq \mu_{.m}$$

Note we will use a two-tailed test, hence the $\neq$ as opposed to $>$.

2. There will be an interaction between patient and provider gender, such male patients will rate females physicians higher than male physicians, but female patients will not have such a pattern.

$$H_0 : \mu_{mm} - \mu_{mf} = \mu_{fm} - \mu_{ff}$$

Breaking apart this null hypothesis, it means that the difference that male patients rated male versus female physicians is the same as the difference that female patients rated male versus female physicians.

The alternative hypothesis would be

$$H_0 : \mu_{mm} - \mu_{mf} \neq \mu_{fm} - \mu_{ff}$$

Breaking apart this alternative hypothesis, it means that the difference that male patients rated male versus female physicians is not the same as the difference that female patients rated male versus female physicians. So if male patients rated female physicians more positively then $\mu_{mm} - \mu_{mf} < 0$ (i.e., be less than 0). But if female patients did not rate physicians differently, then $\mu_{fm} - \mu_{ff} = 0$ (i.e., should be zero). Thus, the two sides of the above hypotheses are *not equal*.

14.3.2 2. Designing a study

You and your team plan out a research study. The method follows:

Participants: Participants were recruited from the Corner Brook region through advertisements posted in local community centers and online platforms. Recruitment materials were approved by the regional health authority. Eligible participants were adults (18+) who had an upcoming medical appointment with a physician. A total of 40 participants were recruited: 20 male and 20 female patients.

A power analysis based on prior literature indicated that a sample size of 40 participants would achieve a power of

$$1 - \beta = .80$$

for detecting medium-sized effects in a 2×2 factorial design.

Materials: A Patient Satisfaction Questionnaire was developed for this study. The questionnaire included items assessing overall satisfaction with the medical appointment (e.g., communication, empathy, comfort level). Responses were scored on a scale from 1 to 40, with higher scores indicating greater satisfaction. A total sum score was used to represent total patient satisfaction. The questionnaire demonstrated acceptable reliability in pilot testing.

Procedure: Participants were informed about the study through advertisements and provided consent prior to participation. After their next medical appointment, participants completed a short survey that asked patient to:

1. Indicate the gender of their physician (male or female),
2. Complete the Patient Satisfaction Questionnaire.

Thus, we have two independent variables (IVs)–patient gender and provider gender–and one dependent variable (DV)–satisfaction. There was an even split in participants seeing a male versus female physician. This resulted in four groups based on the combination of patient gender and provider gender: * Male patient × Male physician ($n = 10$) * Male patient × Female physician ($n = 10$) * Female patient × Male physician ($n = 10$) * Female patient × Female physician ($n = 10$)

Surveys were completed privately and returned to the research team. All data were anonymized. The study was reviewed and approved by the Grenfell Campus Ethics Review Board in collaboration with the regional health authority.

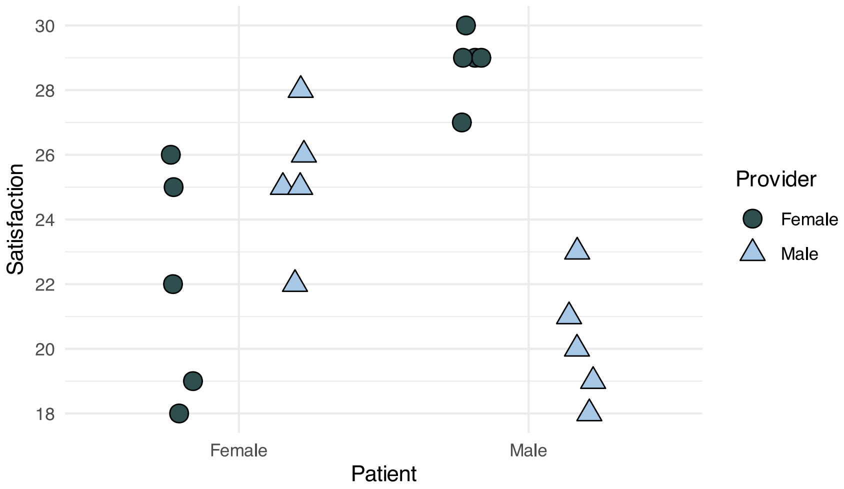
14.3.3 3. Collecting data

The study was completed as described. You obtain the following data:

ID	Patient	Provider	Satisfaction
1	Male	Male	21
2	Male	Male	18
3	Male	Male	20
4	Male	Male	19

ID	Patient	Provider	Satisfaction
5	Male	Male	23
6	Male	Female	29
7	Male	Female	29
8	Male	Female	30
9	Male	Female	27
10	Male	Female	29
11	Female	Male	25
12	Female	Male	26
13	Female	Male	28
14	Female	Male	25
15	Female	Male	22
16	Female	Female	22
17	Female	Female	26
18	Female	Female	25
19	Female	Female	18
20	Female	Female	19

Which can be represented in a figure as:



In the last chapter, we investigated the efficacy of four treatment for obsessive compulsive disorder. In that example, each individual could only be in one group (i.e., someone received CBT *or* Rx). In the example we just introduced regarding visit satisfaction, we have two IVs. That is, an individual has a value on both patient and provider gender. To account for this new model, we must extend our knowledge of the one factor ANOVA into a factorial ANOVA.

The various means of our data can also be summarized in a table:

```


Means and standard deviations for Satisfaction as a function of
a 2(Patient) X 2(Provider) design

          Provider
            Female       Male      Marginal
  Patient        M   SD      M   SD         M   SD
   Female    22.00 3.54 25.20 2.17     23.60 3.24
     Male    28.80 1.10 20.20 1.92     24.50 4.77
 Marginal    25.40 4.35 22.70 3.27

Note. M and SD represent mean and standard deviation,
respectively.
Marginal indicates the means and standard deviations pertaining
to main effects.
```

Marginal Means

A marginal mean refers to the average of cell means across the levels of one factor, ignoring the other factors. For example, the table above shows that the marginal mean of female patients is 23.60. This mean is calculated across all provider genders. We can compare the marginal means to quickly identify potential patterns in the data.

Compare the marginal means of the provider IV (i.e., male vs. female providers); is there potentially a difference?

Our Model

We extend the GLM to represent how we will analyse the data. Recall that the basic structure is:

$$outcome_i = model + error_i$$

And our current model will be:

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{1i}\beta_2 + x_{1i}x_{2i}\beta_3 + e_i$$

Where:

- y_i is the satisfaction (DV) score for individual i,
- x_{1i} is the gender of individual i,
- x_{2i} is the gender of the provider that individual i visited, and
- The βs are the associated coefficients.

Note that $x_{1i}x_{2i}$ is simply individual i's scores on x_1 ND x_2 multiplied together. Some books or references may refer to this as x_3, where $x_{3i} = x_{1i} \times x_{2i}$. Because our IVs are nominal variables, we will use dummy coding. Remember: in dummy coding we have $k - 1$ variables for each level of a factor, where k is the number of levels. Each of our two IVs have two levels, so they each require one variable to represent them. Thus, a patient who is male would score 1 on the patient variable, whereas a female would score 0. Similarly, a provider who was male would score a 1 on the provider variable, whereas a female provider would score a 0. The resulting summary table may be helpful:

Patient	Provider	Mean	SD	x1	x2	x1x2
Female	Female	22	3.54	0	0	0
Female	Male	25.2	2.17	0	1	0
Male	Female	28.8	1.1	1	0	0
Male	Male	20.2	1.92	1	1	1

If you recall from the last chapter, we can determine what each coefficient would be. Using the equation and the table, let's figure out what each coefficient would be. First, for female patients and female providers:

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{1i}\beta_2 + x_{1i}x_{2i}\beta_3 + e_i$$

The above represents the equation for each individual, which is why it includes error. However, for a group of female patients with female providers:

$$FF = \beta_0 + (0)\beta_1 + (0)\beta_2 + (0)(0)\beta_3 = \beta_0$$

Since we know the mean of the female (patient)-female (provider) group is 29.00, we know that $\beta_0 = 22.00$. Second, we can solve the equation for female patients ($x_1 = 0$) with male providers ($x_2 = 1$).

$$FM = \beta_0 + (0)\beta_1 + (1)\beta_2 + (0)(1)\beta_3 = \beta_0 + \beta_2$$

Since we know that $\beta_0 = 22.00$ **and** the mean of female (patient)-male (provider) is 25.2, we can determine that:

$$FM = \beta_0 + \beta_2$$

and

$$\beta_2 = FM - \beta_0 = 25.20 - 22.00 = 3.20$$

Think about it

Before moving, on, try to solve the full equation (i.e., get the values for the other coefficients in the model).

Recall:

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{1i}\beta_2 + x_{1i}x_{2i}\beta_3 + e_i$$

Answer

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{1i}\beta_2 + x_{1i}x_{2i}\beta_3 + e_i$$

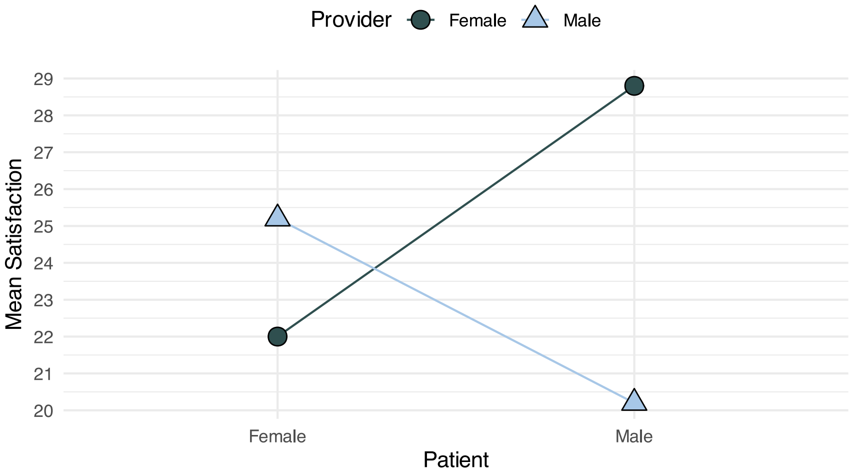
$$y_i = 22.00 + x_{1i}(6.80) + x_{2i}(3.20) + x_{1i}x_{2i}(-11.80) + e_i$$

Understanding this will help you understand regression models in a few chapter.

The Interaction

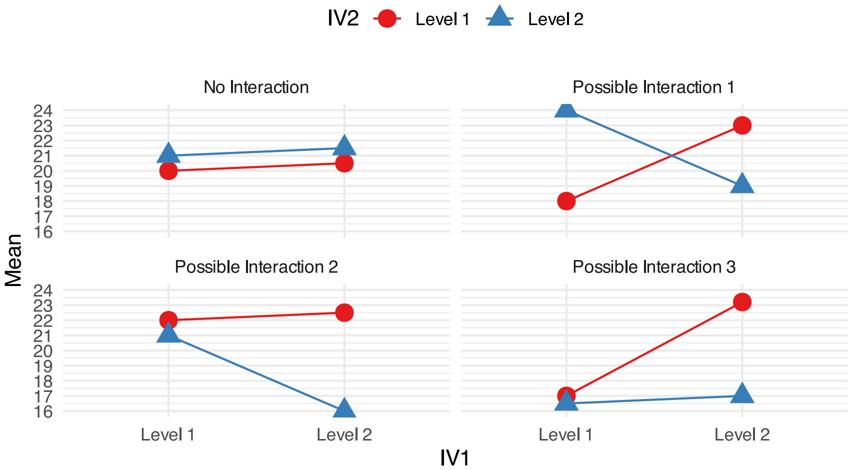
How might you interpret the above? Interactions may seem intimidating, but with some practice you can intuitively understand what they mean. Our interaction term, $\beta_3 = -11.80$, indicates that the relationship between satisfaction and patient/provider differs between men and women. That is, when a **female** patient rates a male versus female physician, we can expect, on average, a 3.20 *increase* in satisfaction. This is represented through $\beta_2 = 3.20$. However, when a **male** patient rates a male instead of female physician, we can expect, on average, a 8.60 *decrease* in satisfaction. This is represented through $\beta_2 + \beta_3 = (3.20) + (-11.80) = -8.60$.

Ultimately, an interaction indicates that the relationship between two variables (typically, an IV and DV) depends on a third variable (typically, another IV). One way to help understand and explain interactions is through figures. Consider the following figure that plots the means of each our our four groups:



In line with the tip above, the relationship between provider gender and satisfaction differs depending on patient gender. Specifically, female patients reported more satisfaction in appointments with male physicians than female physicians, but the reverse occurred for male patients. Male patients reported more satisfaction with female physicians than male physicians. However, we would still need a formal analysis to determine which groups differ from another in a statistically significant way.

Another way to determine whether an interaction is occurring is to view the slopes of the lines. Parallel lines typically indicate no interaction. As the lines become more perpendicular, an interaction is more evident. For example, consider the following figures and whether an interaction likely exists. Note: these are example, not exhaustive. Also, recall that our IVs are categorical/nominal. Therefore, don't let the connecting lines give the illusion of a continuous variable.



Importantly, we need to conduct formal tests to determine if these data indicate a relationship between our IVs and DVs.

14.3.4 4. Analyzing data

In the last chapter, we partitioned the variances and deviations into some sub-components. Specifically, we had:

$$SST = SSB + SSE$$

When we have more than one IV, we need to partition the variances up slightly differently. In the one way ANOVA, we had:

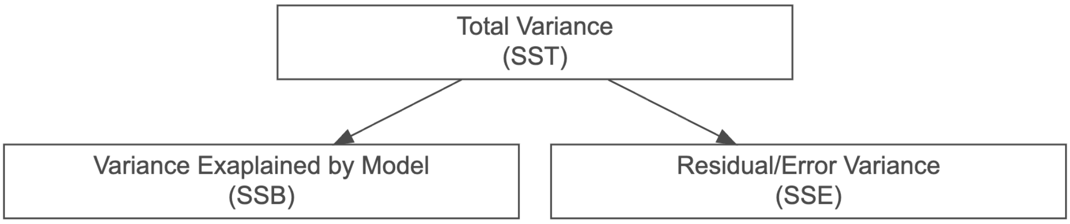


Figure 10: Flowchart

However, in factorial ANOVAs (using the example above), we have:

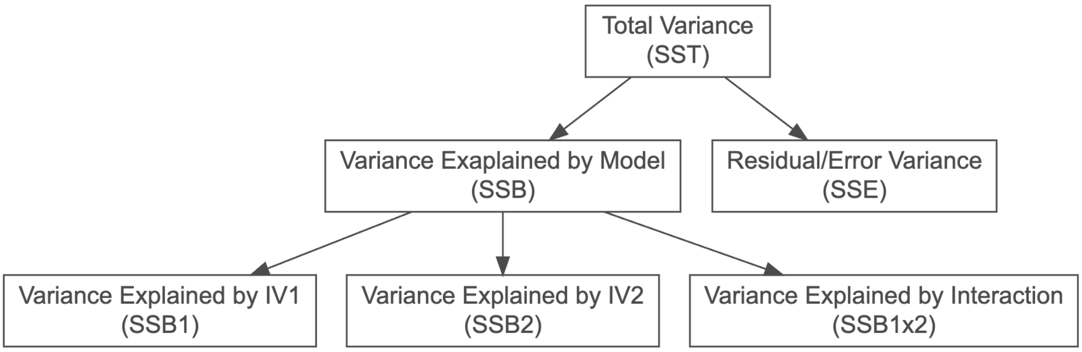

Total Variance
(SST)
Variance Exaplained by Model
(SSB)
Residual/Error Variance
(SSE)
Variance Explained by IV1
(SSB1)
Variance Explained by IV2
(SSB2)
Variance Explained by Interaction
(SSB1x2)

Figure 11: Flowchart

14.3.4.1 SST

Our total sum of squares is no different than a one way ANOVA.

$$SST = \sum_{i=1}^{n} \left(x_i - \overline{x}_{grand}\right)^2$$

with $N - 1$ degrees of freedom. For our example, the mean is 24.05. Thus:

$$SST = (21 - 24.05)^2 + (18 - 24.05)^2 ...(19 - 24.05)^2 = 302.95$$

14.3.4.2 SSB

We can calculate the total SSB the same way we did for a one factor ANOVA:

$$SSB = \sum_{j=1}^{n} n_j \left(\overline{x}_j - \overline{x}_{grand}\right)^2$$

with $k - 1$ degrees of freedom. Although we have two IVs, we want to consider each cell a group. One way to help you visualize this is a table. For our patient-provider example:

		Provider	
		Female	Male
Patient	Female	$\overline{x}_1 - 22.0$	$\overline{x}_2 = 25.2$
	Male	$\overline{x}_3 = 28.8$	$\overline{x}_4 = 20.2$

Thus, we will have four cell to calculate SSB. Recall, our grand mean is 24.05 and the group means are above.

$$SSB = \sum_{j=1}^{n} n_j \big(\overline{x}_j - \overline{x}_{grand}\big)^2$$

$$= 5(22.0 - 24.05)^2 + 5(25.2 - 24.05)^2 + 5(28.8 - 24.05)^2 + 5(20.2 - 24.05)^2 = 214.55$$

Now, we will discuss how to break up the SSB into it's various subcomponents.

14.3.4.2.1 IV1 - Patient Gender

We can calculate the SS for IV1 the same way we would for a one factor ANOVA. Consider only the patient's gender. I have collapsed the table from above to account for this.

Patient	Female	$\overline{x}_{patientf} = 23.60$
	Male	$\overline{x}_{patientm} = 24.5$

In the last chapter, you learned that we could calculate the SSB by:

$$SSB_{patient} = \sum_{n=1}^{j} n_j \big(\overline{x}_j - \overline{x}_{grand}\big)^2$$

with $j - 1$ degrees of freedom (i.e., the number of groups in that factor minus 1). With our data:

$$SSB_{patient} = 10(23.60 - 24.05)^2 + 10(24.50 - 24.05)^2 = 4.05$$

14.3.4.2.2 IV2 - Provider Gender

We can calculate the SS for IV2 the same way we would for a one factor ANOVA. Consider only the provider's gender. I have collapsed the original table from above to account for this.

		Provider
	Female	Male
	$\overline{x}_{providerf} = 25.40$	$\overline{x}_{providerm} = 22.70$

and, thus:

$$SSB_{provider} = \sum_{n=1}^{j} n_j \left(\overline{x}_j - \overline{x}_{grand}\right)^2$$

with $j - 1$ degrees of freedom (i.e., the number of groups in that factor minus 1). With our data:

$$SSB_{provider} = 10(25.40 - 24.05)^2 + 10(22.70 - 24.05)^2 = 36.45$$

14.3.4.2.3 Interaction - Patient X Provider

Now here comes the hard part! Kidding. You know that *SSB = sum of all IV components including interactions*. We can easily calculate the $SSB_{interaction}$ by subtracting our two main factors from our total SSB.

$$SSB_{interaction} = SSB - SSB_{IV1} - SSB_{IV2}$$

$$SSB_{interaction} = 214.55 - 4.05 - 36.45 = 174.05$$

Similarly, the interaction term df are the SSB minus main effect df.

$$df_{interaction} = df_{SSB} - df_{IV1} - df_{IV2}$$

here:

$$df_{interaction} = 3 - 1 - 1 = 1$$

14.3.4.3 SSE

The last thing we need is the SSE for the errors/residuals. You recall that our formula for SSE is:

$$SSE = \sum_{j=1,i=1}^{n} \left(x_{ij} - \overline{x}_j\right)^2$$

A shortcut method may be to use:

$$SSE = \sum_{j=1}^{n} s_j^2 \left(n_j - 1\right)$$

(the sum of the variances multiplied by $n - 1$ for each group). The variances for each of the groups for us is:

Patient	Provider	Variance	Size
Female	Female	12.5	5
Female	Male	4.7	5
Male	Female	1.2	5
Male	Male	3.7	5

Therefore, our SSE is:

$$SSE = 12.5(4) + 4.7(4) + 1.2(4) + 3.7(4) = 88.40$$

14.3.4.4 Mean Squares

Our mean squares are calculated the same as before, matching each SS with their respective df.

$$MS_{patient} = \frac{SS_{patient}}{df_{patient}} = \frac{4.095}{1} = 4.05$$

$$MS_{provider} = \frac{SS_{provider}}{df_{provider}} = \frac{36.45}{1} = 36.45$$

$$MS_{interaction} = \frac{SS_{interaction}}{df_{interaction}} = \frac{174.05}{1} = 174.05$$

and for our error:

$$MS_{error} = \frac{SSE}{df_{error}} = \frac{88.40}{16} = 5.525$$

14.3.4.5 F Statistic

The F statistics will be calculated the same way as a one way ANOVA. However, we will now have **three** tests: one for each main effect and interaction.

$$F_{patient} = \frac{MS_{patient}}{MSE} = \frac{4.05}{5.525} = 0.733$$

$$F_{provider} = \frac{MS_{provider}}{MSE} = \frac{36.45}{5.525} = 6.60$$

$$F_{interaction} = \frac{MS_{interaction}}{MSE} = \frac{174.05}{5.525} = 31.50$$

The p-value for each statistic can be calculated for each F statistic using an F-Distribution table. Find a table here and a calculator here. However, most statistical software/programs will provide the appropriate statistics and p-values. For example, in R:

```
The ANOVA (formula: Satisfaction ~ Patient * Provider) suggests
that:

  - The main effect of Patient is statistically not significant
and small (F(1,
16) = 0.73, p = 0.405; Eta2 (partial) = 0.04, 95% CI [0.00,
1.00])
  - The main effect of Provider is statistically significant
and large (F(1, 16)
= 6.60, p = 0.021; Eta2 (partial) = 0.29, 95% CI [0.03, 1.00])
  - The interaction between Patient and Provider is
statistically significant and
large (F(1, 16) = 31.50, p < .001; Eta2 (partial) = 0.66, 95%
CI [0.40, 1.00])

Effect sizes were labelled following Field's (2013)
recommendations.
```

14.3.4.6 Effect Size

We will use an effect size similar to the one we used for one way ANOVA. However, we will adjust the formula to account for only the ratio of variance explained by a factor relative to the unexplained variance. This effect size is called **partial eta squared** (η_p^2).

$$\eta_p^2 = \frac{SS_{factor}}{SS_{factor} + SSE}$$

For patients:

$$\eta_p^2 = \frac{SS_{patient}}{SS_{patient} + SSE} = 4.05/(4.05 + 88.40) = .043$$

For providers:

$$\eta_p^2 = \frac{SS_{provider}}{SS_{provider} + SSE} = 36.54/(36.45 + 88.40) = .293$$

For the interaction:

$$\eta_p^2 = \frac{SS_{interaction}}{SS_{interaction} + SSE} = 174.05/(174.05 + 88.40) = .663$$

14.3.4.7 Post-hoc Analysis

Remember, the omnibus ANOVA tells us that the group differ, but not how. Furthermore, you should always be mindful that main effects may not tell the full story when the interaction is statistically significant. Thus, the interaction should be the main foci of interpretation when it is statistically significant. While there are many ways to conduct post-hoc analyses (e.g., contrast), I would focus on Bonferroni corrected t-test on the *a priori* analyses of interest.

Specifically, we would conduct individual analyses of one IV –> DV on other the various levels of the other IV. These are called **simple effects**. For example, we could look at the effect of Provider Gender on Satisfaction separately for female and male patients.

14.3.5 5. Write your results/conclusions

Typically, we would report the following in order:

1. Main effect 1
2. Main effect 2
3. Main effect… n
4. Interactions
5. Post-hoc/simple effects

Be sure to address your research questions/hypotheses.

Recall our hypotheses:

1. Participants will be more satisfied with visits from female physicians compared to male physicians.

Results: The results suggest that our data unlikely given a true null that satisfaction is equal for both male and female providers, $F(1, 16) = 6.60, p = .021, \eta_p^2 = .29, 95\% CI[0.03, 1.00]$. Thus, our hypothesis were supported.

2. There will be an interaction between patient and provider gender, such male patients will rate females physicians higher than male physicians, but female patients will not have such a pattern.

Results: There will be an interaction between patient and provider gender, such male patients will rate females physicians higher than male physicians, but female patients will not have such a pattern.

The interaction between Patient and Provider is statistically significant and large, $F(1, 16) = 31.50, p < .001, \eta_p^2 = 0.66, 95\% CI[0.40, 1.00]$.

And now, our post-hoc analyses investigating the effect of provider effect on satisfaction for male and female patients.

14.3.5.1 Male Patients

We can conduct a t-test using only the male data. Please see the t-test section for details.

```

	Welch Two Sample t-test

data:  Satisfaction by Provider
t = 8.687, df = 6.348, p-value = 9.4e-05
alternative hypothesis: true difference in means between group
Female and group Male is not equal to 0
95 percent confidence interval:
  6.20948 10.99052
sample estimates:
mean in group Female   mean in group Male 
                28.8                 20.2 
```

Effect sizes were labelled following Cohen's (1988) recommendations.

The Welch Two Sample t-test testing the difference of Satisfaction by Provider (mean in group Female = 28.80, mean in group Male = 20.20) suggests that the effect is positive, statistically significant, and large (difference = 8.60, 95% CI [6.21, 10.99], t(6.35) = 8.69, p < .001; Cohen's d = 6.90, 95% CI [2.86, 10.88]).

14.3.5.2 Female Patients

```

	Welch Two Sample t-test

data:  Satisfaction by Provider
t = -1.725, df = 6.635, p-value = 0.13
alternative hypothesis: true difference in means between group
Female and group Male is not equal to 0
95 percent confidence interval:
 -7.63505  1.23505
sample estimates:
mean in group Female   mean in group Male 
                22.0                 25.2 
```

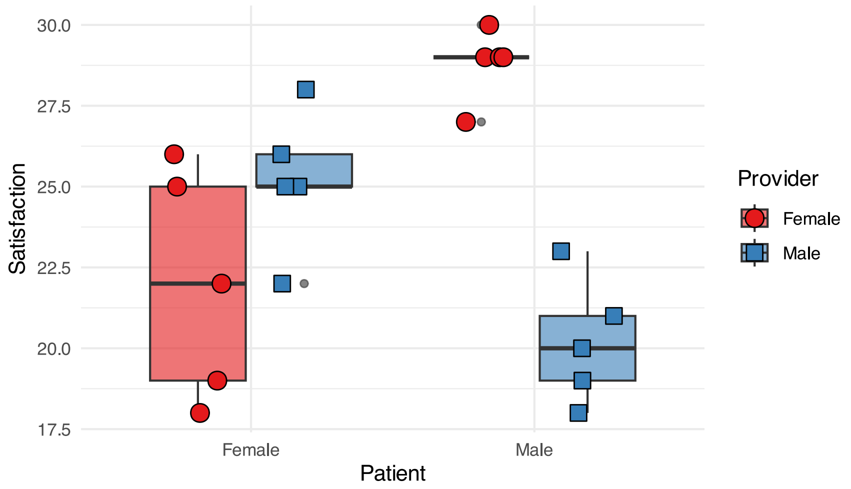
Effect sizes were labelled following Cohen's (1988) recommendations.

The Welch Two Sample t-test testing the difference of Satisfaction by Provider (mean in group Female = 22.00, mean in group Male = 25.20) suggests that the effect is negative, statistically not significant, and large (difference = −3.20, 95% CI [−7.64, 1.24], t(6.64) = −1.73, p = 0.130; Cohen's d = −1.34, 95% CI [−2.98, 0.38]).

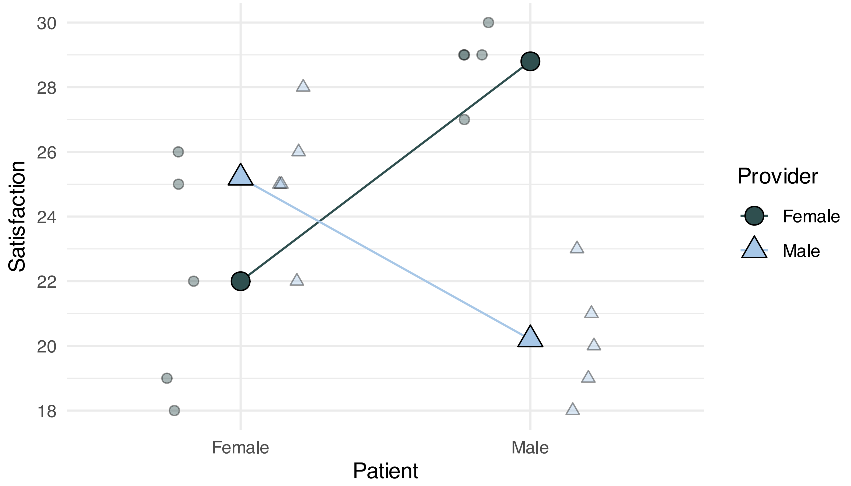
Please note that these results are automated using an R package and you must adjust for APA formatting.

14.3.6 Visualizing the ANOVA

Much like the last chapter, we can visualize the ANOVA using box plots. However, I prefer the following methods:

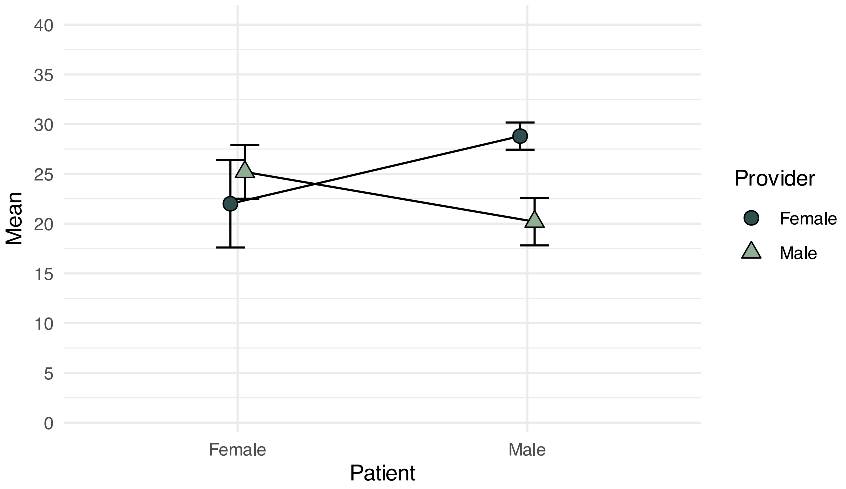


OR:



OR:

We can plot the means and SEs.



Think about it

Look at the error bars above, which represent the 95% confidence intervals of each mean. Which groups would you expect to be statistically significant from one another?

14.4 Conclusion

This chapter covered the factorial ANOVA, a statistical method used to determine whether there are differences among group means when there are two or more independent variables (IVs). This analysis expands. the previous types of ANOVAs we have explored and allows researchers to study the main effects of each IV and the interaction effect between them.

14.5 Factorial ANOVA in R

ANOVAs in R are 'ez'. The `ez` library allows for easily running various types of ANOVA, including mixed, factorial, and one way. It also includes relevant assumption tests.

```
library(ez)
# and the main function
ezANOVA(data = dat_pp, # this is what I named our main data
file
        dv = Satisfaction, # specify the dependent variable
        between = c(Provider, Patient), # concatenated list of
IVs
        wid = ID) #participant ID
```

```
$ANOVA
            Effect DFn DFd         F           p p<.05
ges
1         Provider   1  16  6.597285 2.06161e-02     *
0.2919503
2          Patient   1  16  0.733032 4.04543e-01
0.0438075
3 Provider:Patient   1  16 31.502262 3.89012e-05     *
0.6631739

$`Levene's Test for Homogeneity of Variance`
  DFn DFd   SSn  SSd       F        p p<.05
1   3  16 12.55 28.4 2.35681 0.110246
```

14.6 Practice Question

You work with Disney+ and are responsible for researching audience reception for prospective shows. You are testing the effect of having various main characters for a new teen cartoon. Specifically, the show revolves around a superhero and a sidekick. The superhero can be a dog, cat, or rabbit. The sidekick can be a teen boy or girl.

Based on previous shows, you hypothesize that the cat would be more popular than the dog and bird as main characters. Furthermore, you believe a girl sidekick would be more popular than a boy sidekick. However, you think that differences in animal as a main superhero depend on the sidekick. Specifically, there will be no difference in animal popularity when the sidekick is male, but there cats will be more popular than dogs or birds when the sidekick is female.

After showing six difference groups ($n = 5$ for each group) a pilot episode featuring different combination of superheros and sidekicks, you measure their rating of the show (0-100).

You decide to test the data using a 3 (Cat/Dog/Bird) x 2 (Male/Female Sidekick) ANOVA. Your data are as follows:

ID	Superhero	Sidekick	Rating
1	Cat	Male	52
2	Cat	Male	44
3	Cat	Male	48
4	Cat	Male	48
5	Cat	Male	46
6	Dog	Male	43
7	Dog	Male	49
8	Dog	Male	50
9	Dog	Male	50
10	Dog	Male	48
11	Bird	Male	48
12	Bird	Male	43
13	Bird	Male	44
14	Bird	Male	45
15	Bird	Male	45
16	Cat	Female	69
17	Cat	Female	68
18	Cat	Female	70
19	Cat	Female	73
20	Cat	Female	69
21	Dog	Female	54
22	Dog	Female	63
23	Dog	Female	58
24	Dog	Female	59

ID	Superhero	Sidekick	Rating
25	Dog	Female	59
26	Bird	Female	45
27	Bird	Female	52
28	Bird	Female	47
29	Bird	Female	42
30	Bird	Female	52

14.7 Answers

The ANOVA (formula: Rating ~ Superhero + Sidekick + Superhero * Sidekick) suggests that:

- The main effect of Superhero is statistically significant and large (F(2, 24) = 42.87, p < .001; Eta2 (partial) = 0.78, 95% CI [0.63, 1.00])
- The main effect of Sidekick is statistically significant and large (F(1, 24) = 115.82, p < .001; Eta2 (partial) = 0.83, 95% CI [0.71, 1.00])
- The interaction between Superhero and Sidekick is statistically significant and large (F(2, 24) = 26.93, p < .001; Eta2 (partial) = 0.69, 95% CI [0.49, 1.00])

Effect sizes were labelled following Field's (2013) recommendations.

For female sidekicks:

The ANOVA (formula: Rating ~ Superhero) suggests that:

- The main effect of Superhero is statistically significant and large (F(2, 12) = 55.50, p < .001; Eta2 = 0.90, 95% CI [0.78, 1.00])

Effect sizes were labelled following Field's (2013) recommendations.

Subsequent ttests:

Cat v Dog

Effect sizes were labelled following Cohen's (1988) recommendations.

The Welch Two Sample t-test testing the difference between sh_female_cat and sh_female_dog (mean of x = 69.80, mean of y = 58.60) suggests that the effect is positive, statistically significant, and medium (difference = 11.20, 95% CI [7.19, 15.21], t(6.55) = 6.69, p < .001; Cohen's d = 0.55, 95% CI [−0.74, 1.81])

Bird v Dog

Effect sizes were labelled following Cohen's (1988) recommendations.

The Welch Two Sample t-test testing the difference between sh_female_bird and sh_female_dog (mean of x = 47.60, mean of y = 58.60) suggests that the effect is negative, statistically significant, and very small (difference = −11.00, 95% CI [−16.70, −5.30], t(7.32) = −4.52, p = 0.002; Cohen's d = 0.03, 95% CI [−1.21, 1.27])

Cat v Bird

Effect sizes were labelled following Cohen's (1988) recommendations.

The Welch Two Sample t-test testing the difference between sh_female_cat and sh_female_bird (mean of x = 69.80, mean of y = 47.60) suggests that the effect is positive, statistically significant, and small (difference = 22.20, 95% CI [16.83, 27.57], t(5.48) = 10.35, p < .001; Cohen's d = 0.27, 95% CI [−0.99, 1.51])

For male sidekicks:

The ANOVA (formula: Rating ~ Superhero) suggests that:

- The main effect of Superhero is statistically not significant and large (F(2, 12) = 1.91, p = 0.190; Eta2 = 0.24, 95% CI [0.00, 1.00])

Effect sizes were labelled following Field's (2013) recommendations.

15 Mixed ANOVA

This chapter will cover the mixed ANOVA, a statistical method used when a study includes both between-subjects and within-subjects factors. Unlike a one-way or factorial ANOVA, which only involve between-subjects factors, and unlike a repeated measures ANOVA, which only involves within-subjects factors, a mixed ANOVA combines the two. This design is useful when researchers want to examine how one factor varies across different groups while also considering changes within the same participants over time or across conditions.

For example, imagine a study on memory performance where researchers compare men and women, and measure their memory scores at three different time points. Gender is a between-subjects factor (participants belong to one group only), while time is a within-subjects factor (each participant is measured repeatedly). A mixed ANOVA allows us to test the:

- Main effect of the between-subjects factor (e.g., age group),
- Main effect of the within-subjects factor (e.g., time),
- Interaction effect between the two (e.g., whether changes over time differ by age group).

15.1 Some Additional Details

A mixed ANOVA is appropriate when there is at least one between-subjects IV (e.g., group membership) and at least one within-subjects IV (e.g., repeated measurements). Additionally, it's used when the DV is continuous (interval or ratio scale). It is most commonly used in longitudinal studies that compare different groups across multiple time points or conditions.

Hypotheses in a mixed ANOVA are analogous to those used in either one-way, repeated, and/or factorial ANOVAs. As an example, consider a mixed ANOVA with one between-subjects factor and one within-subjects factor. The mixed ANOVA will have three hypotheses:

1. Main effect of the between-subjects factor

$$H_{0A} : \mu_{A1} = \mu_{A2} = ... = \mu_{An}$$

Where n is the number of levels in the between-subjects factor A.

2. Main effect of the within-subjects factor

$$H_{0B} : \mu_{B1} = \mu_{B2} = ... == \mu_{Bm}$$

Where m is the number of levels in the within-subjects factor B.

3. Interaction effect

$$H_{0AB} : \text{No interaction between A and B}$$

The alternative hypotheses state that at least one main effect or interaction exists. If any null hypothesis is rejected, follow-up tests (e.g., post hoc or simple effects) are needed to locate differences.

15.2 Key Assumptions

A mixed ANOVA requires the following assumptions, which are similar to those presented in both one-way and repeated measures ANOVAs (i.e., it combines the assumptions of both):

1. Continuous dependent variable:

The DV should be measured at the interval or ratio level.

2. Independence of observations

Observations for the between-subjects factor must be independent (e.g., different participants in each group).

3. Normality of residuals

Residuals should be approximately normally distributed for each combination of factors. This can be checked with Q-Q plots or tests like Shapiro-Wilk.

4. Homogeneity of variances

Variances should be roughly equal across groups defined by the between-subjects factor. Levene's test is commonly used.

5. Sphericity (for within-subjects factor)

The variances of the differences between all pairs of repeated measures should be equal. Mauchly's test can assess this; if violated, corrections like Greenhouse-Geisser or Huynh-Feldt are applied.

Think about it

We will not focus on calculating test statistics by hand in this chapter. They are slightly more complex than the factorial ANOVA. However, you should know when to use this design, and how to interpret the results.

15.3 Our New Drug

We have developed a new anti-depressant drug that we believe will be exceptionally effective at reducing major depressive symptoms. This drug has shown promising effects in rats, with rats reported less depressive symptoms (we have a squeak translator) for up to 12-months later. We decide that measuring depression pre-treatment, post-treatment, and 12-month followup would be best. However, the main mechanism by which the drug works is through binding to existing testosterone in the body. Theoretically, this drug should be more effective for men than women. We hope to test this through empirical research.

15.3.1 1. Generating hypotheses

We hypothesize that a main effect and an interaction:

1. There will be main effect of time (a reduction in symptoms)

Null hypothesis:

$$H_0 : \mu_{pre} = \mu_{post} = \mu_{followup}$$

Alternative hypothesis (note: we use a two-tailed, non-directional test):

$$H_0 : \mu_{pre} \neq \mu_{post} \neq \mu_{followup}$$

Think about it

Note that a change in some value can be depicted as 'delta' (Δ). So a change in the mean depression score for participants from pre- to post-treatment can be expressed as:

$$\Delta\mu = \mu_{pre} - \mu_{post}$$

If pre-treatment mean was higher than the post-treatment mean (i.e., they went down), then $\Delta\mu$ should be a positive number. If they got worse, it should be negative.

2. The new drug will be more effective for men and the old drug will be equally effective for men and women (an interaction).

Null hypothesis:

$$H_0 : \Delta\mu_{(newdrug,men)} = \Delta\mu_{(newdrug,women)}$$

Alternative hypothesis:

$$H_A : \Delta\mu_{(newdrug,men)} \neq \Delta\mu_{(newdrug,women)}$$

15.3.2 2. Designing a study

You and your team plan out a research study. The method follows:

Power Analysis: A power analysis based on prior literature indicated that a sample size of 20 participants would achieve a power of

$$1 - \beta = .80$$

for detecting medium-sized effects in a 2 (Gender: Men vs. Women) × 2 (Treatment Type: SSRI vs. New Drug) × 3 (Time: Pre-treatment, Post-treatment [6 months], 12-month follow-up) mixed ANOVA design.

Participants: Participants were recruited from the Corner Brook region through advertisements posted in local community centers and online platforms. Recruitment materials were approved by the regional health authority. Eligible participants were adults (18+) who reported having been diagnosed with major depressive disorder. A total of 20 participants (10 men and 10 women) were recruited. Within each gender group, an equal number of participants were randomly assigned to one of two treatment conditions (standard SSRI or new testosterone antidepressant drug.

Materials: Depressive symptoms were assessed using the Beck Depression Inventory-II (BDI-II), a widely used self-report measure of depressive symptom severity. The BDI-II consists of 21 items scored on a 0–3 scale, with higher scores indicating greater depressive symptoms. Total scores were used as the primary outcome variable. The measure has demonstrated strong reliability and validity in clinical populations.

Procedure: Participants provided informed consent prior to participation. Each participant completed the BDI-II at three time points: 1. Pre-treatment (baseline), 2. Post-treatment (6 months after treatment initiation), and 3. 12-month follow-up.

Participants were randomly assigned to one of two treatment conditions (SSRI vs. new drug) and stratified by gender to ensure equal representation. Treatment protocols were standardized to ensure consistency. Participants and research teams members who provided the medications were both blind to the medication provided. All participants were informed they would be taking an ‘anti-depressant medication’.

Thus, the study included two between-subjects factors, one within0subject factor, and a DV:

Between Factors:

1. Gender (Men vs. Women)
2. Treatment Type (SSRI vs. New Drug)

Within factor:

1. Time (Pre-treatment, Post-treatment, 12-month follow-up)

DV:

1. Depressive symptom severity (BDI-II total score)

Data collection occurred privately, and all responses were anonymized. The study was reviewed and approved by the Grenfell Campus Ethics Review Board in collaboration with the regional health authority.

15.3.3 3. Collecting data

The study was completed as described. You obtain the following data:

ID	Drug	Sex	Pre	Post	Follow
1	New	Male	14	8	18
2	New	Male	12	0	7
3	New	Male	16	6	12
4	New	Male	16	8	14
5	New	Male	3	0	5
6	New	Female	17	17	2
7	New	Female	13	14	10
8	New	Female	14	13	8
9	New	Female	8	7	3
10	New	Female	16	19	14
11	TAU	Male	12	9	12
12	TAU	Male	16	13	10
13	TAU	Male	16	11	7
14	TAU	Male	17	12	10

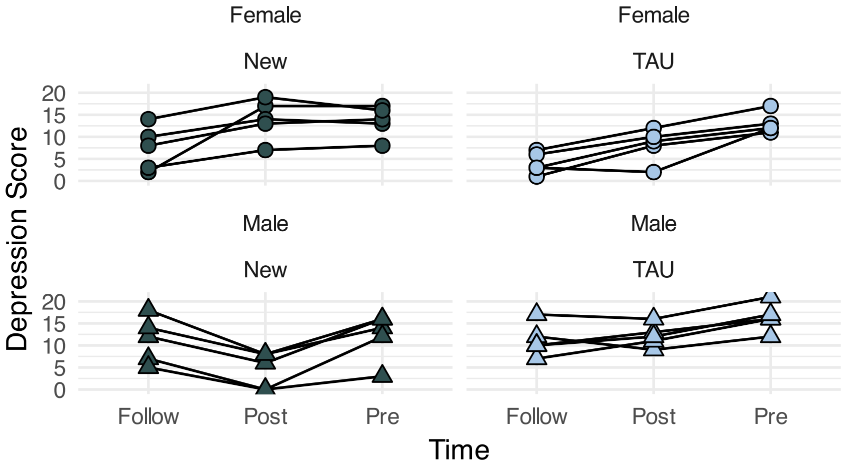
ID	Drug	Sex	Pre	Post	Follow
15	TAU	Male	21	16	17
16	TAU	Female	17	12	7
17	TAU	Female	11	8	1
18	TAU	Female	12	9	3
19	TAU	Female	13	10	6
20	TAU	Female	12	2	3

Thus, we have a 2 (sex) x 2 (pre/post) x 2 (New Drug vs TAU) design, with two between factors (sex and drug) and one within factor (pre/post).

Recall that data should be in long form; it's currently in short form. Long form would like like (note that not all data is presented because the table is 60 rows long; 20 participants X 3 time points = 60 experimental units):

ID	Drug	Sex	Time	BDIScore	Depression
1	New	Male	Pre	14	14
1	New	Male	Post	8	8
1	New	Male	Follow	18	18
2	New	Male	Pre	12	12
2	New	Male	Post	0	0
2	New	Male	Follow	7	7
3	New	Male	Pre	16	16
3	New	Male	Post	6	6
3	New	Male	Follow	12	12
4	New	Male	Pre	16	16

Our data can be represented graphically as (this is the raw data for visual inspection and would not be placed in a paper as it potentially identified individuals; we would present means and CIs/SEs):



15.3.3.1 Our Model

Our models are getting quite lengthy. Here is the full model for this study, which builds on the general linear model:

$$y_i = \beta_0 + \beta_{(drug)}(x_{1i}) + \beta_{(time)}(x_{2i}) + \beta_{(sex)}(x_{3i}) + \beta_{(d\times t)}(x_{1i})(x_{2i}) + \beta_{(d\times s)}(x_{1i})(x_{3i}) + \beta_{(s\times t)}(x_{2i})(x_{3i}) + \beta_{(d\times t\times s)}(x_{1i})(x_{2i})(x_{3i}) + e_i$$

This may look complex, but we have a β for each main effect and interaction. In total we have three main effects, three 2-way interactions and one 3-way interaction.

15.3.4 4. Analyzing data

Contrasts

We will not explore contrasts explicitly in this class. But they are imperative to meaningfully conducting your analyses and interpreting your results. Contrasts are specific comparisons between means that go beyond the overall F-tests. While the ANOVA tells you whether there is a statistically significant effect for a factor or interaction, it doesn't tell you where the differences are. Contrasts allow you to test targeted hypotheses. You should be familiar with them for potential honours projects in the future. For a detailed exploration of contrasts, go here.

By default, R uses dummy coding. However, dummy coding doesn't work well with type III sums of squares, which is what we want to model an interaction. We must use an orthogonal contrast (we will use effects coding).

15.3.4.1 Assumptions

Sphericity

ezANOVA automatically provides Mauchley's tests for each repeated value:

Effect	W	p
Time	0.8273	0.2413
Sex:Time	0.8273	0.2413
Drug:Time	0.8273	0.2413
Sex:Drug:Time	0.8273	0.2413

Based on the results of Mauchley's test, we have not violated this assumption.

Normality

The `rstatix` package can easily test normality for call groups of our analysis:

Drug	Sex	Time	variable	statistic	p
New	Female	Follow	BDIScore	0.9417	0.67793
TAU	Female	Follow	BDIScore	0.9251	0.56329
New	Male	Follow	BDIScore	0.963	0.82901
TAU	Male	Follow	BDIScore	0.9274	0.57909
New	Female	Post	BDIScore	0.9517	0.74938
TAU	Female	Post	BDIScore	0.8948	0.38178
New	Male	Post	BDIScore	0.7817	0.05705
TAU	Male	Post	BDIScore	0.984	0.95464
New	Female	Pre	BDIScore	0.914	0.49192
TAU	Female	Pre	BDIScore	0.813	0.10299
New	Male	Pre	BDIScore	0.79	0.06698
TAU	Male	Pre	BDIScore	0.94	0.66579

We have not violated this assumption.

Homogeneity of Variance

For Sex:

Time	df1	df2	statistic	p
Follow	1	18	0.0083	0.9283
Post	1	18	0	1
Pre	1	18	0.3276	0.5741

For Drug:

Time	df1	df2	statistic	p
Follow	1	18	0.3052	0.5874
Post	1	18	2.555	0.1274
Pre	1	18	0.0304	0.8636

Thus, all of our major assumptions are fine, so let’s move along.

Note: we could set this up as a multi-level model. Although I recommend this, it is beyond the scope of this class.

Note that we do our assumptions prior to our main analyses, when possible. Your results are not interpretable if you have violated the assumptions in a major way.

To analyze, we can use `ezANOVA()` from the `ez` package.

Which gives the following output:

Effect	DFn	DFd	SSn	SSd	F	p	ges
(Intercept)	1	16	6805.35	586.8	185.558	0	0.9
Sex	1	16	22.817	586.8	0.622	0.442	0.029
Drug	1	16	2.017	586.8	0.055	0.818	0.003
Time	2	32	313.3	166	30.198	0	0.294
Sex:Drug	1	16	198.017	586.8	5.399	0.034	0.208
Sex:Time	2	32	172.633	166	16.639	0	0.187
Drug:Time	2	32	33.633	166	3.242	0.052	0.043
Sex:Drug:Time	2	32	76.433	166	7.367	0.002	0.092

15.3.4.2 Main Effects

15.3.4.2.1 Hypothesis 1 - Symptoms will decrease over time**

We will explore all main effects for the purposes of learning, but note that we are interested particularly in the main effect of time (see hypotheses).

Before looking at the main effects, it's important to understand that main effects, significant or not, have little interpretation value when interactions are present. Thus, while we can report these, please do not put to much weight into them.

Sex

Based on our output above, we know there was no effect of sex on response to the drug, $F(1, 16) = 0.622$, $p = 0.442$, $\eta_g^2 = 0.029$. If we ignored all other variables in the model and looked only at the differences between men and women, there would not be an effect.

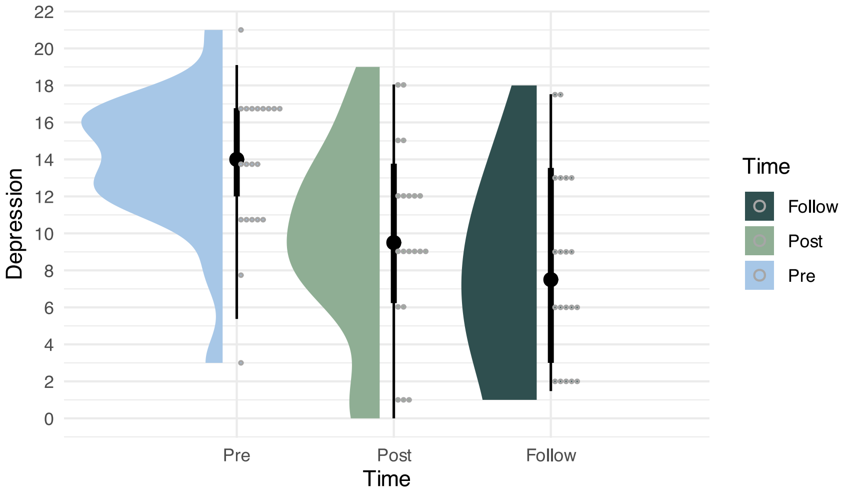
Drug

Furthermore, there seem to be no main effect of drug, $F(1, 16) = 0.055$, $p = 0.818$, $\eta_g^2 = 0.003$. If we ignored sex and time, all other variables in the model and looked only at the differences between TAU and the new drug, there would not be an effect.

Time

There was a statistically significant main effect of time, $F(2,32) = 30.20$, $p < .001$, $\eta_g^2 = 0.294$. If we ignored sex and drug time, depression scores would vary across time.

Let's look at these difference in more detail. This first figure focuses on the changes over time:



You may notice that there seems to be a downward trend, such that depression scores go down from pre, to post, to followup. We can complete post-hoc analyses by running a Tukey's test for the within-subject variable:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Depression ~ Time, data = data_long)

$Time
            diff      lwr     upr    p adj
Post-Follow 1.25 -2.32548 4.82548 0.679101
Pre-Follow  5.35  1.77452 8.92548 0.001899
Pre-Post    4.10  0.52452 7.67548 0.020903
```

Please see the repeated measures ANOVA section of this companion for additional details on reported this output. However, there is a statistically significant reduction in depressive symptoms from the Pre to Post periods, and the Pre to Follow-up periods. However, Post and Follow-Up did not differ.

NOTE: this effect is qualified by significant interactions, which requires additional in-depth exploration.

15.3.4.3 Two-Way Interactions

15.3.4.3.1 Hypothesis 2 - New drug more effective for men**

Sex x Drug

The output suggests a significant two-way interaction between sex and drug, $F(1,16) = 5.399$, $p = .034$, $\eta_g^2 = .208$.

We can investigate this like we did a factorial ANOVA. Our output is as follows:

Drug	.y.	group1	group2	n1	n2	p
New	Depression	Female	Male	15	15	0.241
TAU	Depression	Female	Male	15	15	0.00396

The above uses a Bonferroni adjust p-values. The results suggest that males and females did not differ in response to the new drug, $p = .241$. However, females did respond more favorably to the treatment as usual, $p = .004$. Please see the factorial ANOVA chapter for more details on conducting and writing up a two-way interaction.

Drug x Time

There was no statistically significant drug x time interaction, $F(2,32) = 3.24$, $p = .052$, $\eta_g^2 = .043$.

Sex x Time

The output suggests a significant two-way interaction between sex and time, $F(2,32) = 7.37$, $p = .002$, $\eta_g^2 = .092$. We will explore this in detail; note that this is exploratory analyses versus planned analyses.

We can investigate this like we did a factorial ANOVA. Our output is as follows:

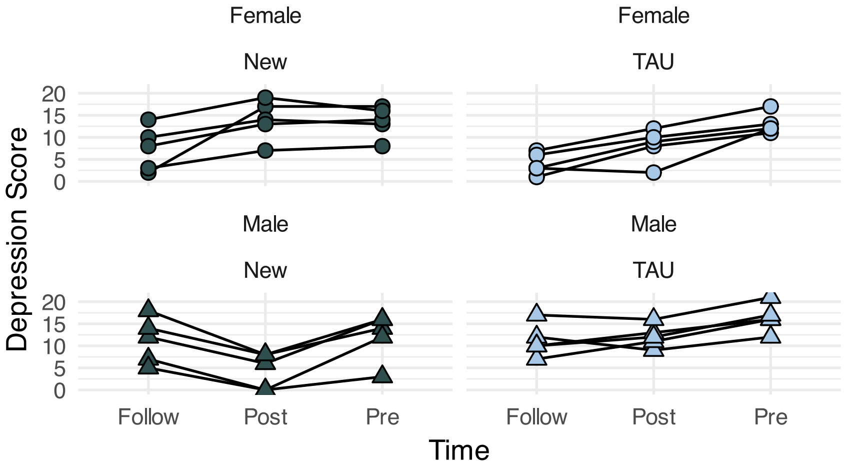
Time	.y.	group1	group2	n1	n2	p
Follow	Depression	Female	Male	10	10	0.009
Post	Depression	Female	Male	10	10	0.237
Pre	Depression	Female	Male	10	10	0.574

The above uses a Bonferroni adjust p-values. The results suggest that males and females did not differ in response during the 'Pre' stage, $p = .574$ nor the 'Post' stage, $p = .237$. However, females did respond move favorably during the 'Follow-up' stage of treatment, $p = .009$. Please see the factorial ANOVA chapter for more details on conducting and writing up a two-way interaction.

15.3.4.4 Three-Way Interaction

The three-way interaction will help clarify the complete picture of the results. Remember, main effects are largely uninterpretable the context of interactions. Well, higher-order interactions may better explain a lower-order interaction. Remember, we had main effects of Time, but males and females only differed in the Follow-up (two-way interaction above).

The following figure will make a reappearance.



In essence, we will be asking if any differences in depression scores for Sex x Time depend on the drug. Or, similarly, if any differences in Drug x Sex depend on time.

Sex x Time for New Drug

Effect	DFn	DFd	F	p	ges
Sex	1	8	0.829	0.389	0.076
Time	2	16	6.447	0.009	0.144
Sex:Time	2	16	16.556	0	0.302

So, for the new drug, we have a sex by drug interaction. Let's tease this apart with post-hoc pairwise comparisons.

Time	.y.	group1	group2	n1	n2	p
Follow	Depression	Female	Male	5	5	0.275
Post	Depression	Female	Male	5	5	0.00818
Pre	Depression	Female	Male	5	5	0.64

Thus, it seems that males and females only differed at the post time for the new drug, with females having higher depression scores. Note that you will need to write up each in proper t-test style.

Let's determine if the changes over time differed for males and females.

Sex	.y.	group1	group2	n1	n2	statistic	df	p
Female	Depression	Follow	Post	5	5	−3.1252	4	0.035
Female	Depression	Follow	Pre	5	5	−2.683	4	0.055
Female	Depression	Post	Pre	5	5	0.5345	4	0.621
Male	Depression	Follow	Post	5	5	7.9048	4	0.001
Male	Depression	Follow	Pre	5	5	−0.5774	4	0.595
Male	Depression	Post	Pre	5	5	−4.9934	4	0.008

Thus, females had no statistically significant changes in depressive symptoms across any time points. However, males had a significant reduction in symptoms from pre to post, but an increase from post to follow.

Sex x Time for TAU

Effect	DFn	DFd	F	p	ges
Sex	1	8	8.366	0.02	0.441
Time	2	16	37.043	0	0.534
Sex:Time	2	16	2.995	0.079	0.085

So, for TAU, we have a main effect of sex and time, but no interaction. We can conduct post hoc tests to determine the nature of these main effects.

.y.	group1	group2	n1	n2	p
Depression	Female	Male	15	15	0.00396

Thus, the means of males ($\overline{x} = 13.30$) was higher than females ($\overline{x} = 8.40$).

For the main effect of time, we can conduct post-hoc analyses.

.y.	group1	group2	n1	n2	statistic	df	p	p.adj
Depression	Follow	Post	10	10	−2.487	9	0.035	0.104
Depression	Follow	Pre	10	10	−7.144	9	0.000054	0.000162
Depression	Post	Pre	10	10	−6.548	9	0.000105	0.000315

We can see that depressive score were lower for the Pre time when compared to the Post and Follow-up time. However, the Post and Follow-up up times were not statistically significant when accounting for the Bonferroni correction.

We now have enough information to answer our initial hypotheses.

15.3.5 5. Write your results/conclusions

All tests are tests are reported as significant at $p < .05$; Bonferroni corrections were used for multiple comparisons.

We first hypothesized a main effect of Time on depressive symptoms, such that depressive symptoms would decrease over time. Indeed, the main effect of time was statistically significant, $F(2, 32) = 30.20, p < .001, \eta^2_g = .294$. Specifically, depressive symptoms were lower at the Pre time ($\overline{x} = 13.8, SD = 3.83$) when compared to the Post ($\overline{x} = 9.7, SD = 5.18, p = .021$) and Follow-up ($\overline{x} = 8.45, SD = 4.97, p = .002$) times.

Second, we hypothesized that the new drug would be more effective for men in long term, while the old drug would not vary over time between men and women. For the new drug, while there was a significant main effect for time, $F(2, 16) = 6.45, p = .009, \eta_g^2 = .144$, females had no statistically significant changes in depressive symptoms across time point, while males experiences a significant decrease in symptoms from Pre to Post and Increase from Post to Follow-up. The Pre and Follow-up scores did not differ for males. males experience lower depressive symptoms when compared to women at the Post time, while other differences existed.

For TAU, there was a main effect of sex, $F(1, 8) = 8.37, p = .020, \eta_g^2 = .440$, with females ($\overline{x} = 8.40$) having significantly lower depressive symptoms than males ($\overline{x} = 13.30$). There was a main effect of time on depressive symptoms, $F(2, 16) = 37.04, p < .001, \eta_g^2 = .534.$ Here, individuals experiences a reduction in symptoms from the Pre time ($\overline{x} = 14.79$) to the Post time ($\overline{x} = 10.20$) and Follow-up time ($\overline{x} = 7.60$). The Post and Follow-up times did not differ.

Thus, while depressive symptoms did decrease, there were some sex and drug differences. Overall, the TAU works equally for men and woman at decreasing symptoms, with most notable benefits from Pre to Post time. There were no addition benefits or downsides to depressive symptoms at follow-up.

However, the new drug seems to have no benefit for reducing depressive symptoms in females. However, for males, it appears to have a significant impact of reducing depressive symptoms in the short term (Pre to Post), but that symptoms increase again in the long-term (from Post to Follow-up).

15.4 Conclusion

Mixed ANOVAs combine both independent and repeated designs, allowing research to model the complexity of psychology phenomenon that change over time and context. While it may feel intimidating, your knowledge of previous designs will help you conduct this analysis.

16 ANCOVA

Analysis of Covariance (ANCOVA) focuses on accounting for variance explained by a continuous variable that may not be the manipulated or pseudo-manipulate variable of interest. Thus, we expend our ANOVA to account for another variable. Typically, we call this other variable a **covariate**.

Simply, we include a covariate in model predicting the DV, without the IV. Then, we add the IV to determine if it has an effect above and beyond the covariate. Furthermore, adding a covariate can help reduce the error variance (i.e., it explains what would otherwise be unexplained error variance). Last, adding a covariate can reduce the impact of confounds on the DV by including them as covariates.

16.1 Our Model

An ANCOVA is essentially a model with a continuous covariate and a dummy coded IV.

$$y_i = (model) + e_i$$

$$y_i = b_0 + b_1(x_{1i}) + b_2(x_{2i}) + e_i$$

Where b_1 is the regression coefficient for the covariate, x_{1i} is individual i's score on the covariate, b_2 is the coefficient for the dummy coded IV, x_{2i} is individual i's score on the IV (0 or 1), and e_i is the error.

Note that there will be more dummy coefficients for more levels of the IV (number of dummy coded variables = levels - 1). For example, if we were conducting a study with one covariate and an IV with three levels (low, medium, and high), our resulting model would be:

$y_i = b_0 + b_1(x_{covariate}) + b_2(x_{medium}) + b_3(x_{high})$

16.2 Our Assumptions

ANCOVA has two assumptions that we have not encountered.

1. Independence between the covariate and the IV: When the IV and the covariate are related/dependent, it makes the interpretation of the model difficult. Furthermore, it can result in inflated estimates of the effects. This is analogous to multicollinearity, which will be discussed in multiple regression.

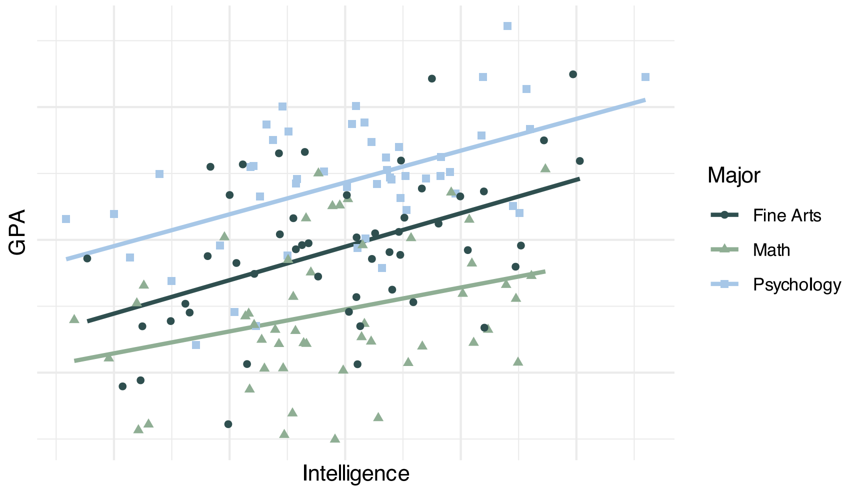
We can test this by running an ANOVA using the IV as the IV and the covariate as the DV. Think about what this would tell us.

Think about it

If the ANOVA model, when `covariate ~ IV` is statistically significant, it would indicate the the covariate differs between IV levels to a degree that is unlikely under a true null hypothesis (i.e., that the covariate is equal between levels of the IVs). Thus, the covariate and the IV are not independent.

We want a non-statistically significant result for this test.

2. Homogeneity of Regression Slopes: We want the relationship between the covariate and the DV to be the same across all levels of the IV. For example, imagine an ANCOVA investigating the effect of university major (IV) on GPA (DV), but using intelligence as a covariate. We would want the association between intelligence and GPA to be the same across majors. Said another way, we would likely expect a positive association between intelligence and GPA, and want these to be the same across disciplines. This is represented in the following figure:



Notice how the line of best fit for each major is similar.

We can test this assumption by modelling an interaction. A significant interaction would indicate that the slopes vary based on levels of the IV and that we have violated this assumption.

3. Homogeneity of Variance: This is the same as one-way ANOVA and, thus, we can use Levene's test.

16.3 Family dietary changes

You are hired by an organization seeking to promote healthy families in the community. They are implementing a course for families around healthy nutrition and balanced diets. They are interested in the impact of participating in a full version of this course, a brief version, and no treatment (three-level IV) on eating habits (DV). Specifically, we will measure eating habits as the total number of servings of fruits/vegetables per family over a one week period, divided by the number of people in that family. Thus, a DV score of 12 would indicate that the family ate total 12 servings of fruits or vegetables per person over the week following their intervention. They ask you to conduct the analyses. Importantly, you know that income is a strong predictor of eating habits, because healthy foods are typically more expensive. Thus, you treat income as a covariate (family income/year).

Practice Question

1. What are the IV, DV, and covariate?
2. What could be another covariate?
3. What is the model (i.e., write out the regression equation)?

Answers

1. IV: eating intervention. Three levels (full, brief, none). DV: average fruit and vegetables intake per person in a family over one week. Covariate: income.
2. Many possibilities.
3.

$$y_{fruitvegetable,i} = b_0 + b_i(x_{income,i}) + b_2(x_{full,i}) + b_3(x_{brief,i}) + e_i$$

16.4 Power Analysis

You review the literature and determine find a nice meta-analysis on eating habit interventions. This study suggests that a brief intervention was able to effectively reduce junk food consumption and resulted in a effect size if $\eta^2 = .54$, $95\%CI[.35, .62]$. Bring brilliant, you use the lower bound estimate. Your power analysis suggests:

```
     Balanced one-way analysis of variance power calculation 

              k = 3
              n = 7.06532
              f = 0.733799
      sig.level = 0.05
          power = 0.8
```

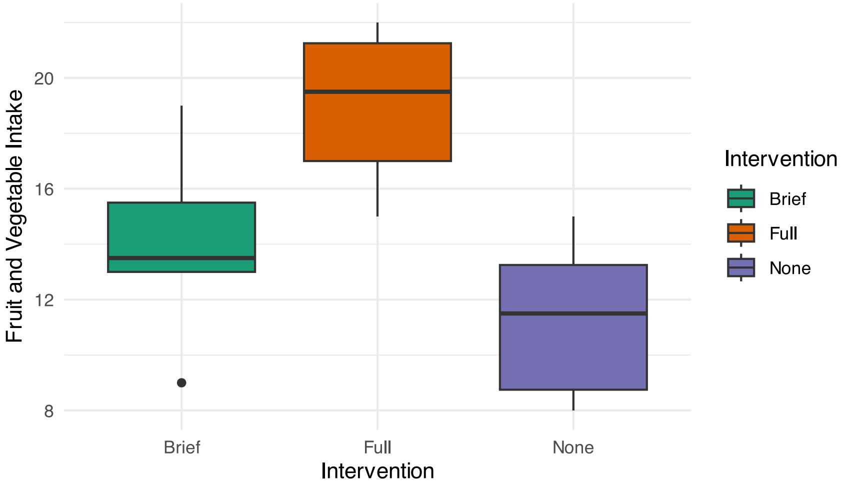
```
NOTE: n is number in each group
```

Thus, we need eight families per group ($n = 7.06$, rounded up). You successfully recruit 24 families. You obtain the following data:

Family	Intervention	FV	Income
1	Full	18	57200
2	Full	15	57900
3	Full	21	62100
4	Full	21	55200
5	Full	22	50700
6	Full	17	62900
7	Full	17	60500
8	Full	22	51500
9	Brief	14	50500
10	Brief	13	59500
11	Brief	13	44200
12	Brief	17	61900
13	Brief	13	47400
14	Brief	9	55500
15	Brief	19	50500
16	Brief	15	57800
17	None	9	52100
18	None	8	48100
19	None	12	57200
20	None	11	46000
21	None	13	65300
22	None	8	49400
23	None	15	50400
24	None	14	50000

Which gives us the following descriptive statistics:

```
Descriptive statistics for FV as a function of Intervention.

 Intervention     M       M_95%_CI   SD
        Brief 14.12 [11.62, 16.63] 3.00
         Full 19.12 [16.87, 21.38] 2.70
         None 11.25  [8.98, 13.52] 2.71

Note. M and SD represent mean and standard deviation,
respectively.
LL and UL indicate the lower and upper limits of the 95%
confidence interval
for the mean, respectively.
The confidence interval is a plausible range of population
means that could
have caused a sample mean (Cumming, 2014).
```



16.5 1. Generating hypotheses

Our hypothesis will be analogous to one-way ANOVA.

$$H_0 : \mu_{none} = \mu_{brief} = \mu_{full}$$

or

$$H_0 : \text{all } \mu \text{ equal}$$

and:

$$H_A = \text{at least one } \mu \text{ different}$$

16.6 4. Analyzing data

16.6.1 Assumptions

First, let’s check our assumptions.

1. Independence of IV and covariate: For this, we would run a one-way ANOVA using our IV as our IV and our covariate as the DV.

```


ANOVA results using Income as the dependent variable


    Predictor             SS df             MS      F    p
partial_eta2
  (Intercept) 22823161250.00  1 22823161250.00
699.71 .000
 Intervention   107507500.00  2    53753750.00
1.65 .216          .14
        Error   684977500.00 21
32617976.19
 CI_95_partial_eta2

         [.00, .36]


Note: Values in square brackets indicate the bounds of the 95%
confidence interval for partial eta-squared
```

Thus, we have not violated the assumptions. The income level did not differ based on treatment groups, $F(2, 21) = 1.65$, $p = 0.216$; $\eta^2 = 0.14$, $95\% CI[0.00, 0.31]$.

2. Homogeneity of Regression Slopes: We want to test a model including both covariate, DV, and their interaction. We will interpret our main analyses later, but for now, focus on the interaction term in a full ANOVA model. We will specify an interaction by simply multiplying the DV and covariate in a model (in addition to our regular model). We must specify a type 3 sums of squares. We will the the `apa.aov.table()` function from the `apaTables` package:

```


ANOVA results using FV as the dependent variable


             Predictor     SS df    MS    F    p partial_eta2
           (Intercept)  12.11  1 12.11 1.55 .230
          Intervention  39.10  2 19.55 2.50 .110          .22
                Income   0.72  1  0.72 0.09 .765          .01
 Intervention x Income  24.34  2 12.17 1.55 .238          .15
                 Error 140.91 18  7.83
 CI_95_partial_eta2

         [.00, .45]
         [.00, .19]
         [.00, .38]


Note: Values in square brackets indicate the bounds of the 95%
confidence interval for partial eta-squared
```

We want to focus on the interaction term. Here the interaction (Intervention x Income) is not statistically significant. This indicates that the relationship between income and FV consumed does not depend on intervention type. We have not violated the assumption.

3. Homogeneity of Variance: Here we use Levene's test.

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2   0.054  0.947
      21
```

We conducted Levene's test to asses the homogeneity of variances and the results suggest that we did not violate this assumption, $F(2, 21) = .054$, $p = .9472$.

We then conduct our main analysis, using our IV, DV, and covariate. Most statistical software programs allow you to readily specify these variables. Here are our results:

```
$ANOVA
        Effect DFn DFd     SSn     SSd       F           p
p<.05      ges
1 Intervention   2  21 191.211 194.307 10.3327 0.000751026
* 0.495985

$`Levene's Test for Homogeneity of Variance`
  DFn DFd     SSn     SSd        F       p p<.05
1   2  21 3.99751 68.0933 0.616416 0.54936
```

16.6.2 Effect Size

Although many statistical software programs will calculate our effect size for us, you should know what it means. We can calculate η_p^2 (partial eta squared). η_p^2 can be calculated as:

$$\eta_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error/residual}}$$

For our intervention:

$$\eta_p^2 = \frac{220.27}{220.27 + 165.25} = .571$$

This represent a ratio of what is explained by the IV compared to the residual error.

16.6.3 Post-Hoc Tests

Much like a one-way ANOVA, we will need to conduct post-hoc tests. However, we will need to adjust the groups to account for any differences in covariates; remember, we want to control for these differences in the

DV, which is why we are doing this in th first place. We can still use Tukey's HSD and get a comparison of each group.

```

     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = FV ~ Income + Intervention, data = dat_nutr)

Linear Hypotheses:
                  Estimate Std. Error t value Pr(>|t|)
Full - Brief == 0     5.00       1.50    3.34   0.0087 **
None - Brief == 0    -2.87       1.44   -1.99   0.1398
None - Full == 0     -7.87       1.54   -5.12   <0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

And the associated confidence intervals:

```

     Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = FV ~ Income + Intervention, data = dat_nutr)

Quantile = 2.53
95% family-wise confidence level


Linear Hypotheses:
                  Estimate lwr     upr
Full - Brief == 0   4.998    1.208   8.787
None - Brief == 0  -2.874   -6.524   0.775
None - Full == 0   -7.872  -11.759  -3.985
```

It is also important to get effect size estimates for each comparison. But remember, we want to adjust the means using the covariate (i.e., account for their different scores on the covariate). We will also need the adjusted means and standard deviations.

16.6.3.1 Adjusted Means

```

 Intervention effect
Intervention
  Brief    Full    None
14.1255 19.1232 11.2512

 Lower 95 Percent Confidence Limits
Intervention
   Brief     Full     None
11.99536 16.89994  9.08174

 Upper 95 Percent Confidence Limits
Intervention
  Brief    Full    None
16.2557 21.3465 13.4207
```

We will also need adjusted standard deviations. We have standard errors:

```
[1] 1.02120 1.06584 1.04003
```

Note: the order aligns with the previous adjusted means (Brief, Full, None). You may recall from an earlier chapter that $\sigma_{\overline{x}=\frac{s}{\sqrt{N}}}$. Thus, we can estimate:

$$s = \sigma_{\overline{x}}\sqrt{N}$$

We have eight per group, thus the following will return the adjusted SD for each group:

```
[1] 2.88840 3.01465 2.94165
```

Again, the order aligns with the output from the adjusted means (Brief, Full, None). The `mes()` function can then calculate each standardized difference (Cohen’s d) for each group difference.

```
Mean Differences ES:

 d [ 95 %CI] = -1.69 [ -2.83 , -0.55 ]
  var(d) = 0.34
  p-value(d) = 0.01
  U3(d) = 4.52 %
  CLES(d) = 11.56 %
  Cliff's Delta = -0.77

 g [ 95 %CI] = -1.6 [ -2.68 , -0.52 ]
  var(g) = 0.3
  p-value(g) = 0.01
  U3(g) = 5.47 %
  CLES(g) = 12.89 %

 Correlation ES:

 r [ 95 %CI] = -0.67 [ -0.88 , -0.26 ]
  var(r) = 0.02
  p-value(r) = 0.01

 z [ 95 %CI] = -0.81 [ -1.36 , -0.27 ]
  var(z) = 0.08
  p-value(z) = 0.01

 Odds Ratio ES:

 OR [ 95 %CI] = 0.05 [ 0.01 , 0.37 ]
  p-value(OR) = 0.01

 Log OR [ 95 %CI] = -3.07 [ -5.14 , -1 ]
  var(lOR) = 1.12
  p-value(Log OR) = 0.01

 Other:
```

```
 NNT = -5.14
 Total N = 16
```

```
Mean Differences ES:

 d [ 95 %CI] = 0.99 [ -0.05 , 2.02 ]
  var(d) = 0.28
  p-value(d) = 0.08
  U3(d) = 83.79 %
  CLES(d) = 75.72 %
  Cliff's Delta = 0.51

 g [ 95 %CI] = 0.93 [ -0.05 , 1.91 ]
  var(g) = 0.25
  p-value(g) = 0.08
  U3(g) = 82.44 %
  CLES(g) = 74.51 %

 Correlation ES:

 r [ 95 %CI] = 0.47 [ -0.04 , 0.78 ]
  var(r) = 0.04
  p-value(r) = 0.09

 z [ 95 %CI] = 0.51 [ -0.04 , 1.05 ]
  var(z) = 0.08
  p-value(z) = 0.09

 Odds Ratio ES:

 OR [ 95 %CI] = 5.98 [ 0.91 , 39.28 ]
  p-value(OR) = 0.08

 Log OR [ 95 %CI] = 1.79 [ -0.09 , 3.67 ]
  var(lOR) = 0.92
  p-value(Log OR) = 0.08

 Other:

 NNT = 2.8
 Total N = 16
```

```
Mean Differences ES:

 d [ 95 %CI] = 2.64 [ 1.3 , 3.98 ]
  var(d) = 0.47
  p-value(d) = 0
  U3(d) = 99.59 %
  CLES(d) = 96.92 %
  Cliff's Delta = 0.94

 g [ 95 %CI] = 2.5 [ 1.23 , 3.77 ]
  var(g) = 0.42
  p-value(g) = 0
  U3(g) = 99.38 %
  CLES(g) = 96.14 %

 Correlation ES:

 r [ 95 %CI] = 0.82 [ 0.54 , 0.93 ]
  var(r) = 0.01
  p-value(r) = 0

 z [ 95 %CI] = 1.15 [ 0.6 , 1.69 ]
  var(z) = 0.08
  p-value(z) = 0

 Odds Ratio ES:

 OR [ 95 %CI] = 120.78 [ 10.6 , 1375.76 ]
  p-value(OR) = 0

 Log OR [ 95 %CI] = 4.79 [ 2.36 , 7.23 ]
  var(lOR) = 1.54
  p-value(Log OR) = 0

 Other:

 NNT = 1.31
 Total N = 16
```

16.7 5. Write your results/conclusions

We investigated the impact of a healthy eating intervention on the average number of fruits and vegetables consumed per person, per family over the course of one week. Additionally, we controlled for family income. We hypothesized that intervention would impact the amount of F&V consumed and explored potential differences using post-hoc comparisons. We tested several assumptions to determine the suitability of ANCOVA to test our hypothesis.

The assumption of independence of intervention and our covariate, income was held. The income level did not differ based on treatment groups, $F(2, 21) = 1.65$, $p = 0.216$; $\eta^2 = 0.14$, $95\% CI[0.00, 0.31]$. Second, we conducted Levene's test to assess the homogeneity of variances of treatment groups; the results suggest that we did not violate this assumption, $F(2, 21) = .054$, $p = .9472$. Last, we did not violate the homogeneity of regression slope assumption, $F(2, 18) = 1.55$, $p = .238$.

We conducted an ANCOVA to determine the impact of intervention on consumption of fruits and vegetables, using family income as a covariate. The covariate, income, was not related to the consumption of fruits or vegetables, $F(1, 20) = 0.00$, $p = .996$, $\eta_p^2 = 0$. However, the results suggest the amount of fruit and vegetables consumed did vary by treatment type to a proportion that is unlikely given a true null hypothesis $F(2, 20) = 110.14$, $p < .001$, $\eta_p^2 = .57$, $95\% CI[.26, .69]$.

16.7.1 Post-hoc test

We conducted post-hoc tests using Tukey's LSD to determine which interventions differed. We had no a priori hypothesis; thus, we analyses were purely exploratory. First, our results suggest that families in the Full intervention ($M = 19.12$) consumed statistically significant more fruits and vegetables compared to the Brief intervention ($M = 14.13$), difference $= 4.99$, $d = 1.69$, $95\% CI = [0.55, 2.83]$, $p = .001$.

Second, our results suggest that families in the Full intervention ($M = 19.12$) consumed statistically significant more fruits and vegetables compared to no intervention ($M = 11.25121$), difference $= 7.87$, $d = 2.64$, $95\% CI = [1.30, 3.98]$, $p < .001$.

16.8 Conclusion

THE ANCOVA is a powerful analysis that allows researchers to account for covariate, a variable they believe to be related to the DV, in the analysis. This allows them to parse out the unique variance in the DV explained by the IV, which can demonstrate the importance of a hypothesized variable of interest.

16.9 ANCOVA in R

16.9.1 `ezANOVA()`

We can use a number of functions to calculate the ANCOVA; however, the `ezANOVA()` function from the `ez` package is, not to sound lame, ez. We will need to specify the type of sum of squares; we will use type 2 because we don't care about an interaction. Note that this function will run Levene's test on the adjusted data.

```
$ANOVA
        Effect DFn DFd     SSn     SSd       F           p
p<.05      ges
1 Intervention   2  21 191.211 194.307 10.3327 0.000751026
* 0.495985

$`Levene's Test for Homogeneity of Variance`
  DFn DFd     SSn     SSd        F       p p<.05
1   2  21 3.99751 68.0933 0.616416 0.54936
```

16.9.2 aov()

We could use the default `aov()` to conduct our analyses. The lovely `apa.aov.table()` function let's us specify the types of sums of squares. **We will use the `apa.aov.table()` output for our class.**

```


ANOVA results using FV as the dependent variable



    Predictor     SS df     MS     F    p partial_eta2
CI_90_partial_eta2
       Income   0.00  1   0.00
0.00 .996          .00
 Intervention 220.27  2 110.14 13.33 .000          .57
[.26, .69]
        Error 165.25 20
8.26

Note: Values in square brackets indicate the bounds of the 90%
confidence interval for partial eta-squared
```

Thus, it seems the covariate, income, is not linked to the amount of FV consumed. However, there is a main effect of intervention on number of F&V consumed. Third, our results suggest that families in the Full intervention ($M = 11.25121$) consumed statistically significant more fruits and vegetables compared to the Brief intervention ($M = 14.13$), difference $= 2.87$, $d = 0.99$, $95\% CI = [-0.05, 2.02]$, $p = .139$.

17 Correlation

Correlation is a statistical technique used to measure the strength and direction of the relationship between two continuous variables. Unlike ANOVA, which compares group means, correlation focuses on whether changes in one variable are associated with changes in another. Importantly, for this courser, a correlation involves two continuous variables.

The most common measure is Pearson's correlation coefficient (r), which ranges from −1 to +1: - $r = +1$: Perfect positive relationship (as one variable increases, the other increases). - $r = -1$: Perfect negative relationship (as one variable increases, the other decreases). - $r = 0$: No linear relationship.

17.1 Some Additional Details

Correlation describes two key aspects of the relationship between variables:

1. Direction

Direction involves that nature of the relationship between the variables. A **positive correlation** indicates that as one variable increases, the other also increases (e.g., hours studied and exam scores). Positive correlations are above 0 ($r > 0$); however, rarely is the "+" placed in from of the correlation. Thus, assume that a correlation with no symbol means it is positive. A **negative correlation** indicates that as one variable increases, the other decreases (e.g., stress level and sleep duration). A negative correlation is between -1 and 0. There will be a "-" symbol in front of a negative correlation. A correlation of zero (0) indicates that there is no linear relationship between the variables.

2. Strength The strength of a correlation is indicated by the absolute value of r. That means you can ignore whether it is positive of negative. For example, a correlation of $r = .3$ and $r = -.3$ have the same strength/magnitude, but different directions. There are some typical cut-offs to apply a qualitative descriptor to correlations. For example, Cohen (2013, p. 116) described:

- Weak: $|r| \approx 0.10 - 0.29$
- Moderate: $|r| \approx 0.30 - 0.49$
- Strong: $|r| \geq 0.50$

The closer $|r|$ is to 1, the stronger the linear relationship. However, correlation does not imply causation; two variables may be related without one causing/influencing the other.

17.2 Key Assumptions

When using Pearson's correlation, the following assumptions should be met:

1. Continuous Variables: Both variables should be measured at the interval or ratio level.

2. Linearity: The relationship between the two variables should be approximately linear. This can be checked visually using a scatterplot.

3. Normality: Both variables should be approximately normally distributed, especially for significance testing. This can be assessed with histograms, Q-Q plots, or tests like Shapiro-Wilks.

4. Homoscedasticity: The variability of one variable should be similar across all values of the other variable. This can be checked visually in a scatterplot.

5. Independence of Observations: Each observation should be independent of the others.

17.3 Do Looks Matter?

You are hired by Instagram to research their trending posts. They ask if people like more posts based on the poster's attractiveness. You decide that you will collect posts and determine if there is a relationship between the number of likes a posts receives and how attractive the poster is rated.

17.3.1 1. Generating hypotheses

We hypothesize that there is a relationship between attractiveness and likes of a post. We will use a two-tailed test (i.e., not specify a direction). Thus, our hypotheses are as follows under NHST. Note: ρ ('row') is the population correlation.

$$H_0 : \rho_{attr,likes} = 0$$

$$H_A : \rho_{attr,likes} \neq 0$$

17.3.2 2. Designing a study

Power Analysis: You review the literature and determine that the there is a strong link between attractiveness and popularity. You determine the best estimate to of a population parameter to be $\rho = .75$. You power analysis (see chapter related to power) results in:

```
     approximate correlation power calculation (arctangh
transformation)

              n = 10.725
              r = 0.75
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Thus, you require 11 posts.

Materials: The study used Instagram profile photos from 11 randomly selected accounts. Photos had to contain a full face and body view of the user. For each photo, the number of likes on the Instagram post (continuous variable) was collected. Next, four anonymous individuals rated the profile photos' attractiveness on a 10-point Likert scale (1 = very unattractive, 10 = very attractive). The attractiveness ratings were averaged across raters to create a single attractiveness score per photo.

Procedure: Raters were were shown a series of Instagram photos in randomized order and asked to rate each photo's attractiveness using the 10-point scale. The number of likes for each photo was recorded directly from Instagram at the time of data collection.

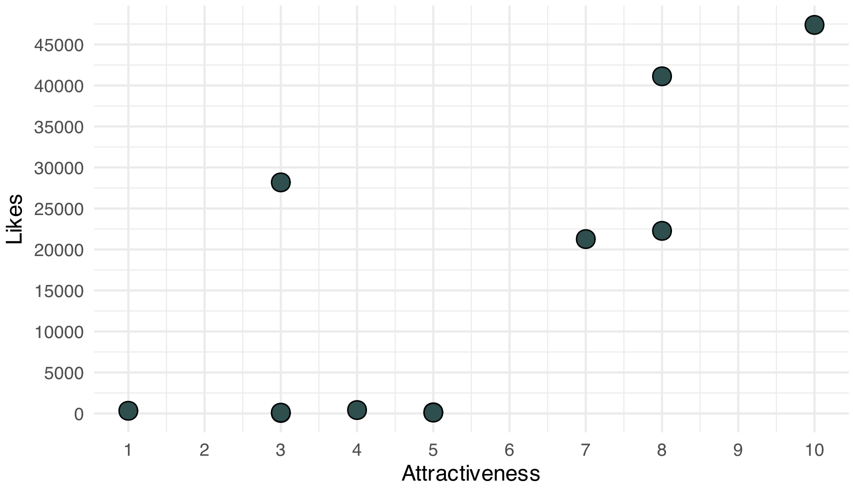
The primary analysis examined the correlation between Instagram likes and attractiveness ratings. All data were anonymized, and no identifying information was shared. The study was reviewed and approved by the Grenfell Campus Ethics Review Board.

17.3.3 3. Collecting data

The study was completed as described. You obtain the following data:

Likes	Attractiveness
22272	8
47387	10
65	3
417	4
99	3
143	5
41123	8
108	5
28183	3
330	1
21268	7

Which we can plot as a scatterplot:



17.3.4 4. Analyzing data

We typically estimate the associations with variables using covariances and correlations. Prior to looking at correlations, let's review covariance. The covariance measures the cross-product (multiplication of two variables) of deviations from the respective variables' means to determine their average deviations. Consider the mean for likes 1.467227^{4} and attractiveness 5.181818. We would calculate how much each variables deviates from the mean, multiple each, then average them.

$$cov_{(x,y)} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

Where i is an individual. For out data, we need to calculate the deviation from likes and attractiveness:

Likes	Attractiveness	Deviation from Like	Deviation from Attractiveness
22272	8	7600	2.8182
47387	10	32715	4.8182
65	3	−14607	−2.1818
417	4	−14255	−1.1818
99	3	−14573	−2.1818
143	5	−14529	−0.1818

Likes	Attractiveness	Deviation from Like	Deviation from Attractiveness
41123	8	26451	2.8182
108	5	−14564	−0.1818
28183	3	13511	−2.1818
330	1	−14342	−4.1818
21268	7	6596	1.8182

To calculate the sum of the products (numerator of covariance), we would multiple the third and fourth column for each row and add them up. This is known as the sum of the products (SP):

$$SP = \sum (x_i - \bar{x}) \times (y_i - \bar{y})$$

For our data:

$$SP = (7599.727)(2.8181818) + (32714.727)(4.8181818) + ...$$

$$= 381880.5$$

We then divide by $n - 1$

$$cov_{(likes,attractivess)} = \frac{381880.5}{11 - 1} = 38188.05$$

Much like a correlation, a positive covariance indicates that the variables tend to associate in the same direction. So here, posts with more likes tend to be rated as more attractiveness. A negative covariance would indicate the opposite: higher scores on one variable are associated with lower scores on others.

A major issue with covariance is that it is not standardized. That is, it is difficult to compare covariances with one another, making them difficult to interpret. One cannot readily interpret a covariance of 10, 100, 1000, or 10000, because they depend on the original metric of the variables used to calculate it. For example, imagine that I wanted to calculate the covariance of height and weight (kg). I could use cm or m as a metric of height. When I use height in cm I get a covariance of 189.44. When I use height in m, I get 1.89. *This is despite the strength of the association being identical*. How might we resolve this?

17.3.4.1 Correlation Coefficient

The **correlation coefficient** is a standardized covariance. What does standardization do? It converts a variable into a standard unit that facilitates comparisons. We can scale our variables considering the standard deviations. Specifically, we would adjust our formula to be:

$$r = \frac{cov_{x,y}}{s_x s_y}$$

Which, using some maths, can be re-written as:

$$r = \frac{SP}{\sqrt{(SS_x)(SS_y)}}$$

Where SS_x is the sum of squared deviations of x and SS_y is the sum of squared deviations of y. Thus, using our data above, we need the standard deviation of Likes (18197.74) and Attractiveness (2.75). Using covariance and standard deviations:

$$r = \frac{38188.05}{(18197.74)(2.75)} = .763$$

Or, if we used sum of squared deviations (the second way to calculate it; see above):

$$r = \frac{381880.5}{\sqrt{(3311577666)(75.63636)}} = \frac{381880.5}{500475.5} = .763$$

Recall that correlations range from −1 to 1. A correlation of −1 indicates a perfect negative relationship. A correlation of 1 indicates a perfect positive relationship. A correlation of 0 indicates no relationship. Thus, correlation helps us understand the **direction** (+, -) and **magnitude** (absolute size of the number) of a relationship. For example, $r = .4$ indicates a positive relationship, but $r = -.6$ indicates a stronger relationship that is negative.

For our research, seems that Likes and Attractiveness have a strong positive relationship. However, you know that we must determine if this data is unlikely given a true null hypothesis.

Distribution of the r test statistic

Correlations have a distribution that is related to the t distribution. Simply, r can be converted to a t statistic:

$$t_r = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$$

We could then use a t-distribution to determine the likelihood of our or more extreme data given a true null. For our example:

$$t = \frac{.7630\sqrt{11-2}}{\sqrt{1-.7630^2}} = 3.541699$$

Much like the t-distribution, the r-distribution has $n-2$ degrees of freedom. For our study this means that:

$$df = n - 2 = 11 - 2 = 9$$

Looking up the in our t-distribution table, we get a p-value of .00629. Therefore, our probability of getting this large or larger of a correlation under a true null is 0.006.

In r, the output we get is (my data is called `dat` and the variables are `Likes` and `Attractiveness`:

```

	Pearson's product-moment correlation

data:  dat$Likes and dat$Attractiveness
t = 3.542, df = 9, p-value = 0.0063
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.300882 0.934957
sample estimates:
     cor
0.763035
```

Most researcher will present a correlation matrix for all continuous variables used in their study, regardless of whether their *main* analyses was a correlation. This provides a nice summary of the means, SDs, and correlations between all variables. The following is an example of a correlation matrix.

```


Means, standard deviations, and correlations with confidence
intervals
 

  Variable          M        SD       1          
  1. Likes          14672.27 18197.74            
                                                 
  2. Attractiveness 5.18     2.75     .76**      
                                      [.30, .93] 
                                                 

Note. M and SD are used to represent mean and standard
deviation, respectively.
Values in square brackets indicate the 95% confidence interval.
The confidence interval is a plausible range of population
correlations
that could have caused the sample correlation (Cumming, 2014).
 * indicates p < .05. ** indicates p < .01.
```

17.3.4.2 Effect Size

The effect size for correlation is known as the coefficient of determination (R^2). It is simply the correlation squared and it tells us the proportion of variance in one variable that is accounted for by another variables; it is usually expressed as a percent (i.e., $R^2 = .348$ would be written as 34.8%. Sometimes you may encounter it expressed as the amount of variance in one variable that can be 'explained' by another variable. Note that this does not mean that one variable causes another.

$$R^2 = (r)^2$$

17.3.5 5. Write your results/conclusions

We hypothesized that the number of likes on an Instagram post would be correlated with the rated Attractiveness of the poster. The results suggest that our data are unlikely given a true null hypothesis, $r = .763$, $95\% CI[.30, .93]$, $df = 9$, $p = .006$, $R^2 = .582$. Approximately 58.2% of the variance in Likes can be explained by Attractiveness, indicating a strong effect size.

17.4 Conclusion

You will encounter correlations a lot in psychological research. They are used in both descriptive and inferential statistics. They are used to describe *your* data set, and to make inferences about population parameters. They will be imperative in future chapters about regression and related designs.

17.5 r in R

We can use the `cor.test()` function, where we specify the two variables we wish to correlate.

```
cor.test(dat$Likes, dat$Attractiveness) #our data was called
'dat'
```

```

    Pearson's product-moment correlation

data:  dat$Likes and dat$Attractiveness
t = 3.542, df = 9, p-value = 0.0063
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.300882 0.934957
sample estimates:
     cor
0.763035
```

Note that R provides many useful pieces of information: r, t, and p. It also gives us the 95%CI (which is for r). You can use r as an effect size, but R^2 (squaring r), the **coefficient of determination**, also indicates the proportion in one variable that is explained by the other. Here, 58.22% of the variance in Likes can be explained by knowing the Attractiveness rating of the poster. Note: again, this does not mean causes. It means by knowing someones Attractiveness/Likes, we have a fairly reliable guess at what the other would be.

Plotting

The standard way to plot two continuous variables is through a scatterplot.

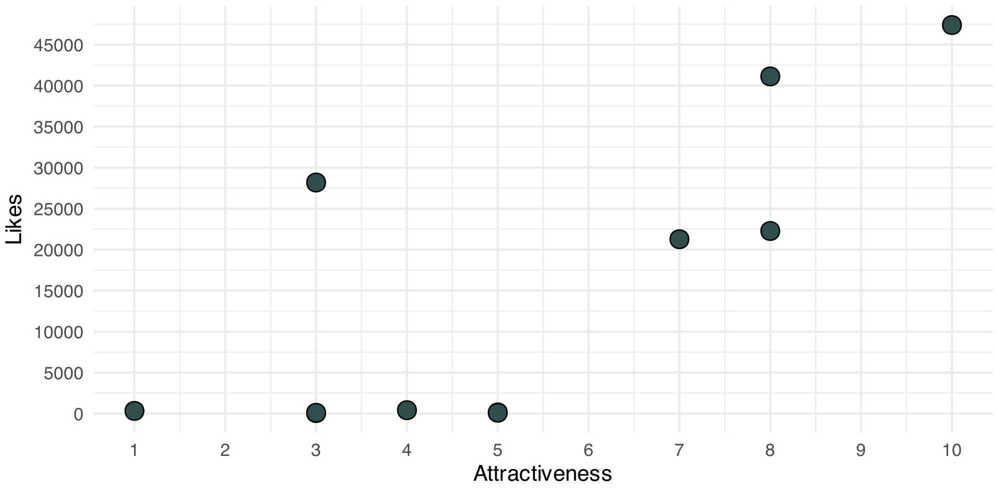
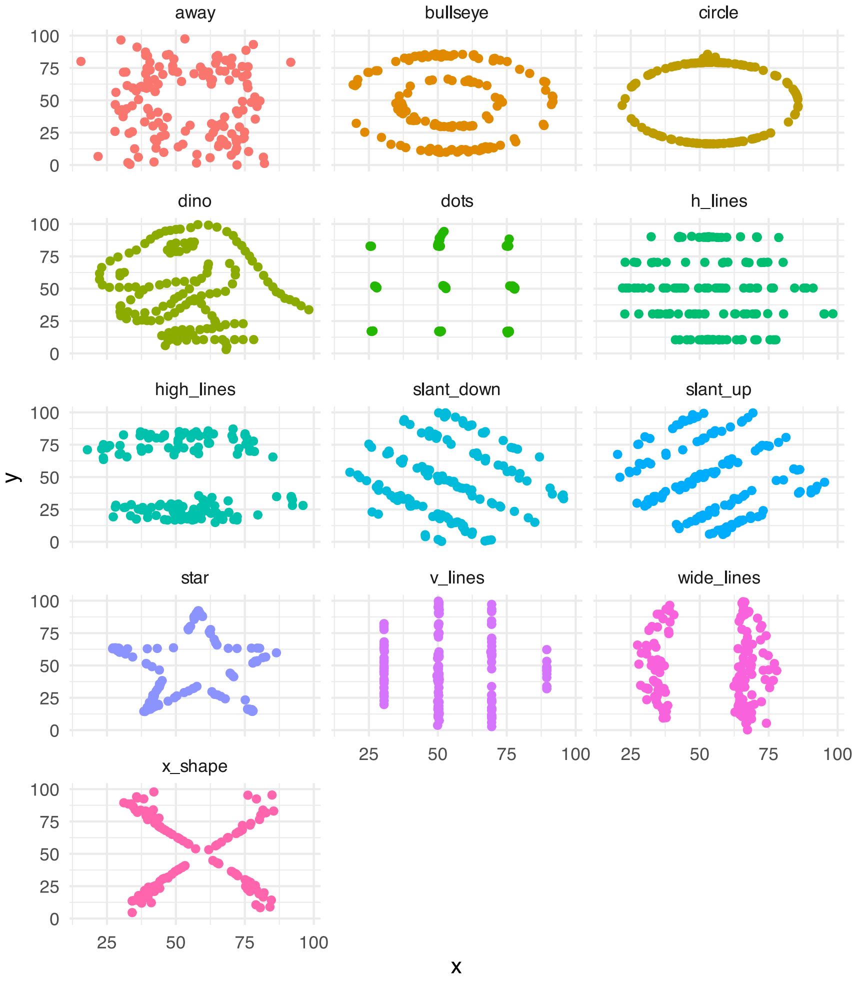


Figure 12: Scatterplot of data.

It's important to visually inspect your data. For example, the `datasauRus` package (Gillespie et al., 2025) shows a quick demonstration how two variables' relationship can be vastly different even with the exact same means, standard deviations, and correlations. As is the case for each of the following data sets:



17.6 Practice Questions

Using the following data, which were measurements of sizes of children's books:

Height	Width
10	4
9	11
18	16
4	6
15	20

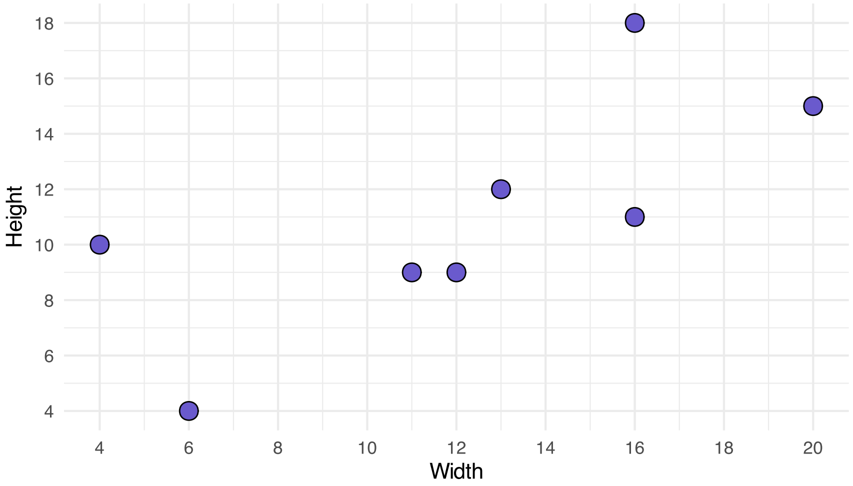
Height	Width
12	13
11	16
9	12

Calculate the correlation, and write the results up (including r, p, and the CI if you use R).

Draw a quick scatterplot of the data (put width on the x-axis).

17.7 Answers

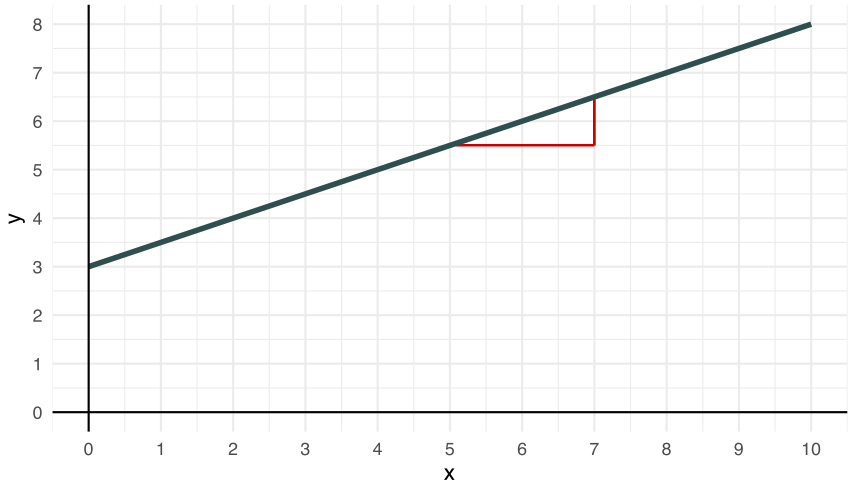
```
Effect sizes were labelled following Funder's (2019)
recommendations.

The Pearson's product-moment correlation between df_prac$Height
and
df_prac$Width is positive, statistically significant, and very
large (r = 0.72,
95% CI [0.04, 0.95], t(6) = 2.56, p = 0.043)
```



18 Simple Regression

18.1 Some Additional Details

Prior to diving into the data, let's revisit some high school math! We have already encountered this in sections referring to our models, but let's go back to basics. You likely recall $y = mx + b$, the function of a line. Recall that we can determine the y-position of a line by knowing the x position, the slope, and the y-intercept of that line. Consider the following:



Hopefully, you can see that the line crosses the y-axis at 3. Furthermore, the slope can be calculated by $\frac{y_2 - y_1}{x_2 - x_1}$ or, *rise over run*. This is represented by the red lines. Using rise over run, we get: $\frac{6.5 - 5.5}{7 - 5} = \frac{1}{2} = 0.5$. Thus, the above line is represented by $y = 0.5x + 3$. We can readily predict what the value of y would be by knowing the value of x. For example, by knowing that $x = 3$, we could determine that y is $y = 0.5x + 3 = 0.5(3) + 3 = 4.5$.

Practice Question

Determine the equation for the following line:

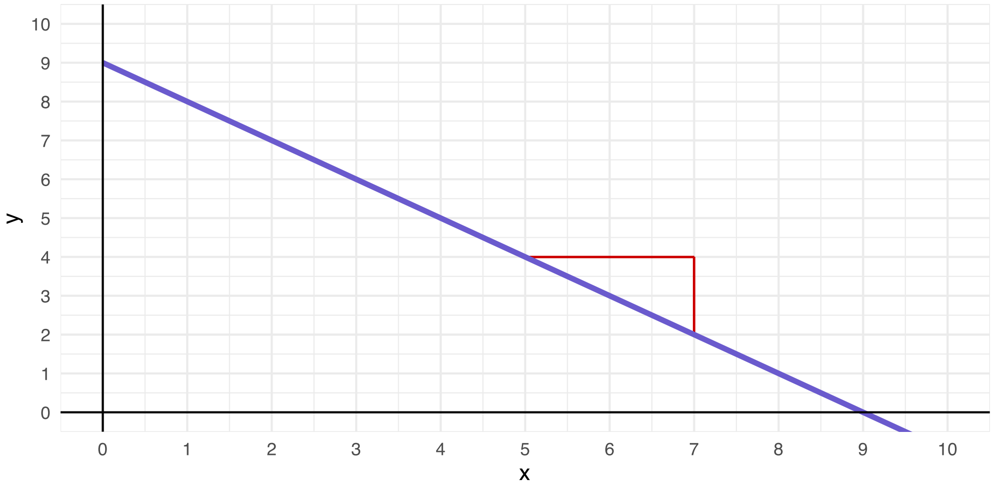


Figure 13: Line graph.

Answer

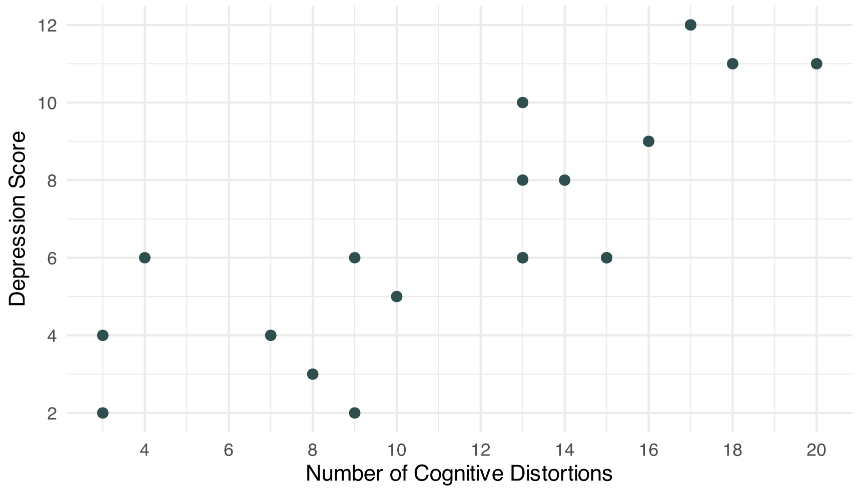
$y = -1x + 9$

You may be thinking, "STOP TYLER", but this is relevant. This equation maps nicely onto our more general linear models that we have been using in our analyses:

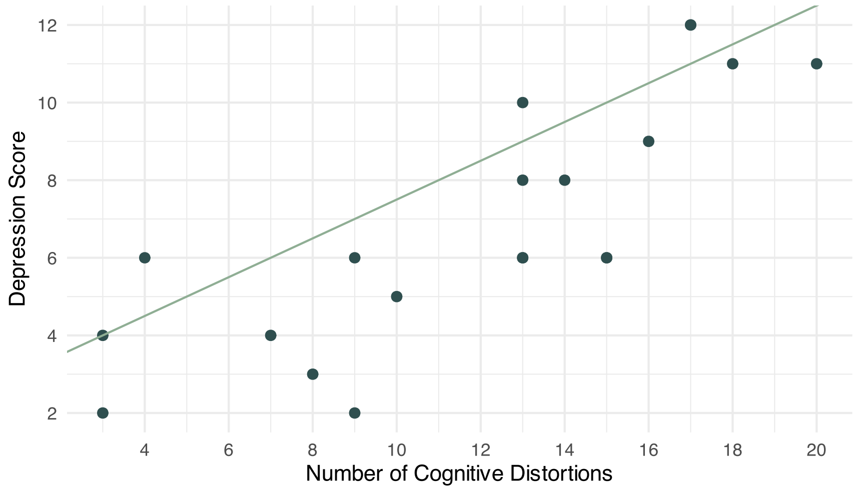
$$y_i = model + e_i$$

Where y_i is the outcome for individual *i*, the model is based on our hypotheses and resulting analyses, and e_i is the error for individual *i* (i.e., what the model does not explain).

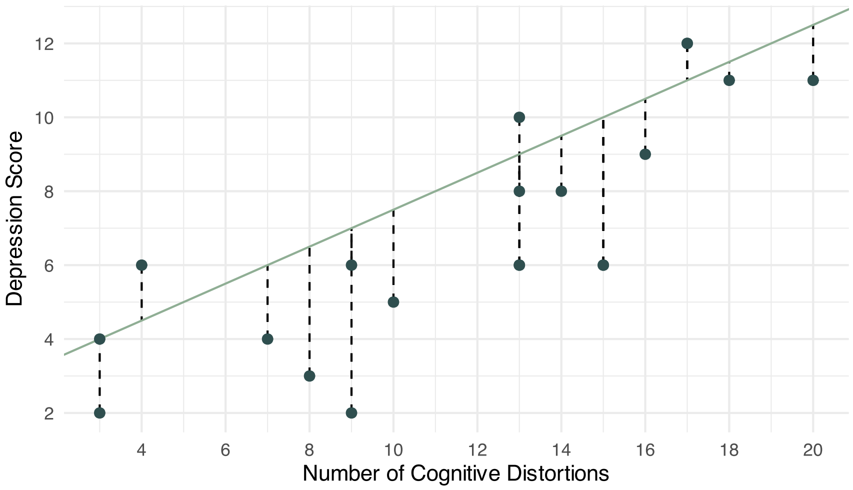
Imagine that instead of x and y, we had a independent and a dependent variable. We wanted to predict someones depression score (y; and measured on a scale of $1 - 14$) by knowing the number of cognitive distortions they have on average each day. It might look like the following:



We can try to fine a straight line that fits *all* those points, but that's impossible. For example, maybe we can guess that that y-intercept is around 2.5, and the slope is about 0.5. This would result in:



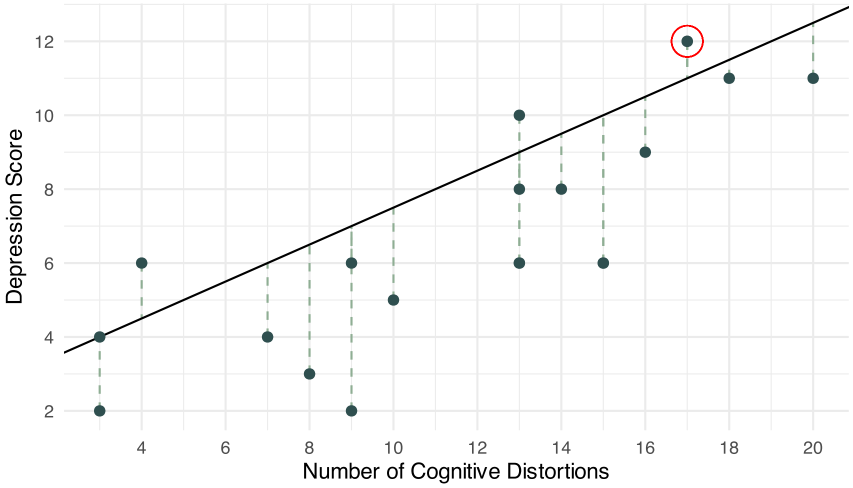
That's not *too bad* of a guess…but what it seems to have a lot of error. Our line doesn't do a great job fitting on all the points. We can measure how much error it has by measuring the distance from the points to the line:



Here, error is represented by the dotted lines. That is, we guessed that:

$$dep = 2.5 + 0.5(Distortions)$$

That someone's depression score would be 2.5 plus 0.5 times the number of distortions they have. But that would mean that the points fall directly on the line. Thus, the distance between each point and the line is the error. Let's consider person 1 (circled below):



This person had, on average, 17 cognitive distortions per day and had a depression score of 12. Our line would not predict a depression score of 12.

$$depression = 2.5 + 0.5(distortions) = 2.5 + 0.5(17) = 11$$

The difference between 12 and 11 is the error for individual 1. We call this a **residual**. The residual for individual 1 is 1.

$$y_i = 2.5 + 0.5(x_i) + e_i$$

For person 1:

$$12 = 2.5 + 0.5(17) + 1$$

Perhaps there is a better line that would minimize the errors across all the observations? That is, if we calculated the error for every person, as we did for person 1, the total of the squared errors would be 115.75. Let's try to lower this number.

18.2 Ordinary Least Squares

Ordinary Least Squares (OLS) is an algebraic way to get the best possible solution for a regression line. It minimizes the error of the line. Typically, in psychology we write a simple regression as the following, where β are referred to as coefficients:

$$y_i = b_0 + b_1 x_i + e_i$$

- Where $\hat{y}$ is the predicted score on y, the dependent variable,
- b_0 is the intercept coefficient,
- b_1 is the slope coefficient,
- x_1 is the score of individual i on the independent variable, and
- e_i is the residual for individual i.

Math people have figured out the optimal solution to find these coefficients. The following are solutions to OLS simple regression:

$$b_1 = \frac{cov_{(x,y)}}{s_x^2}$$

Where $cov_{(x,y)}$ is the covariance between x and y, and s_x^2 is the variance of x. For us:

- Covariance = 12.871053
- Variance = 25.292105

- Mean of depression = 11.85
- Mean of distortions = 6.85

Thus:

$$b_1 = \frac{12.87105}{25.29211} = 0.50889$$

And the intercept solution is:

$$b_0 = \overline{y} - b_1\overline{x}$$

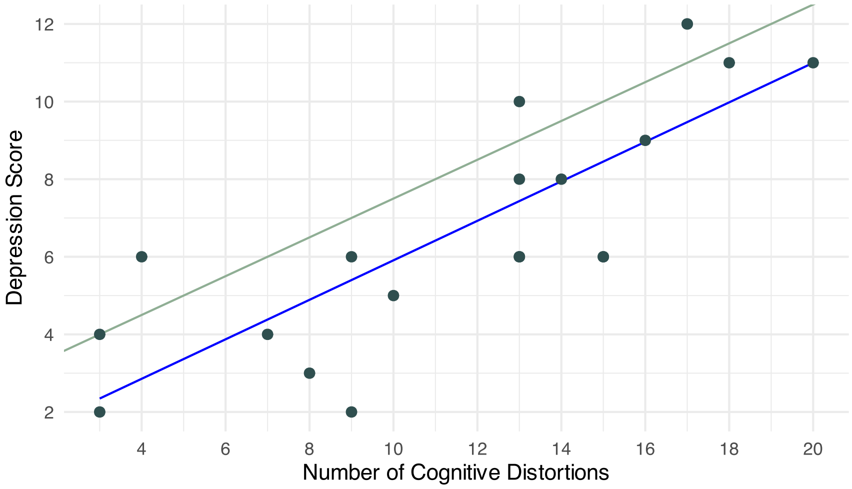
Therefore:

$$b_0 = 6.85 - 0.50889(11.85) = 0.81965$$

You did it! Our best possible fitting line is:

$$y_i = 0.81965 + 0.50889(x_i) + e_i$$

If we were to calculate the sum or the squared residuals for each person, we would get 66.1. This is lower than the line that was built on our best guess. In fact, this is the smallest possible value. We can see compare our original line with the line of best fit:



Sometimes in psychological researcher, people try to determine if knowing an individual's score on one variable will allow us to predict their score on another variable. For example, by knowing someone is severely depressed, can we predict the likelihood they will attempt suicide? By knowing a student's social support, can we predict how many classes they will attend in university? By knowing how many classes someone attends, can we predict their grade in the course? In all of these example, the predictor (or as you will learn in the next chapter, set of predictors) is often termed the independent variable(s) (x), and the outcome/criterion variable is the dependent variable (y). When we have only one predictor, we refer to this as a *simple regression*. **Note: this is not way implies that x causes y. Consider prediction much like a relationship or association, as discussed in the last chapter on correlations**.

Tip

Note that a regression is not necessarily an experiment, despite sometimes using the terms independent and dependent variables.

18.3 Key Assumptions

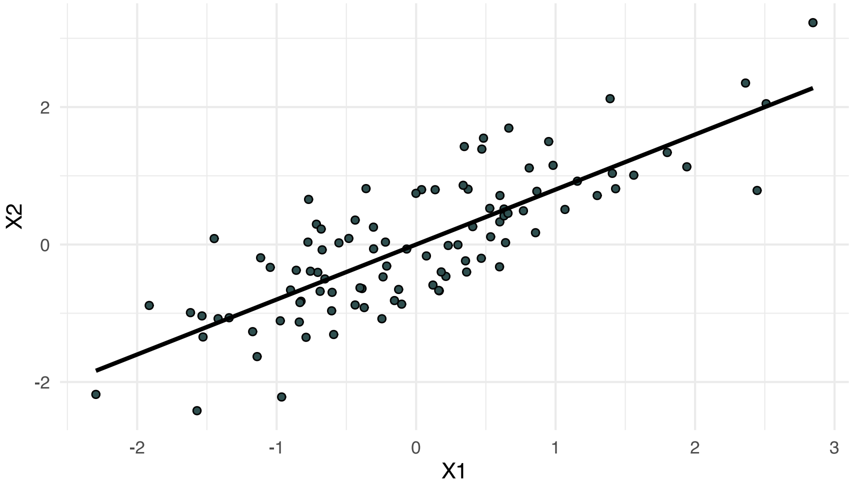
There are a few basic assumptions for regression analyses:

1. Homoscedasticity: The variance of the residuals (i.e., error terms)is constant across different levels/values of the IV. That is, the spread of data around the line of best fit should be similar all along the line.

It can be hard to visualize with a diagonal line, so a formal analysis can be done. We can plot the residuals on the y axis and the predicted values (sometimes terms fitted values) on the x-axis. Here, we want a relatively straight line around 0, indicating an mean residual of 0. Furthermore, we want the dots to be dispersed equally around each predicted value. It's hard to determine with our data because there are so few points.

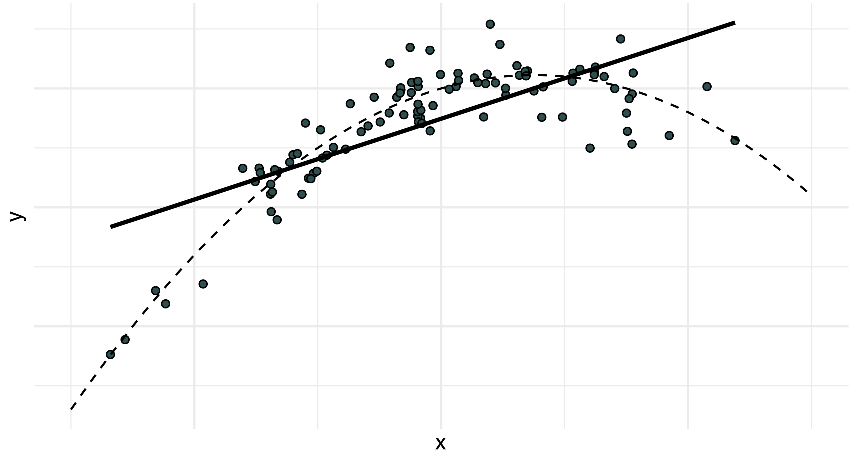
2. Independence: Each observation is independent; thus, each residual is independent. You must ensure this as a researcher. For example, if you had repeated measures (e.g., two observation from each person), then these would not be independent.

3. Linearity: The relationship between IV and DV is linear. We can visually assess this using a scatterplot. We hope that the points seem to follow a straight line. We can fit our line of best fit from OLS to help with this. Consider the following, which depicts a linear relationship:



Our data appear quite linear. For your own reference, here is an example of a non-linear relationship. It is a quadratic relationship (yes, like you remember from math $y = x^2$):

Quadratic



Dotted line is true line of best fit. Solid line is line of best fit resulting from a OLS regression.

4. Normality of residuals: We can asses using Q-Q plots and Shapiro-Wilk, which was covered in a previous chapter. *Remember, the null hypothesis of the SW test is that the data are normally distributed.*

18.4 Get some shut eye

You are a psychologist investigating the impact of technology use at night and sleep quality. You conduct a literature review and believe that the amount of screen time within two hours before 'bedtime' will negatively impact the total time in REM sleep during for that night.

18.4.1 1. Generating hypotheses

Regression hypothesis reference βs (betas), which are regression coefficients. The population parameter will be called β, while that sample statistic will be b. Thus, our hypothesis (we will use a two-tailed test) is:

Null:

$$H_0 : \beta = 0$$

Alternative:

$$H_A : \beta \neq 0$$

Recall that we started by relating our regression to $y = mx + b$, specifically:

$$outcome_i = (model) + error_i$$

And we are hypothesizing that the outcome is the function of some variables, so we can now say:

$$y_i = b_0 + b_1(x_{1i}) + e_i$$

Where y_i is the DV for individual i, b_0 is the intercept, b_1 is the coefficient for the IV, and e_i is the residual for individual i. So, our best guess at an individual's REM sleep will be a function of two coefficients. Any differences between our guess (i.e., predicted value) and the actual REM sleep (i.e., observed value) is error.

18.4.2 2. Designing a study

You decide to recruit students and ask them to measure both screen time before bed (IV) and access their Apple Watch data to assess the amount of time in REM sleep during the night (DV). Specifically:

Sample Size Determination: You review the literature and believe that the link between screen time and sleep is negative. Specifically, your best guess at the population parameter is $R^2 = .25$, which can be converted to $f^2 = .3333$ ($f^2 = \frac{R^2}{1-R^2}$. We will conduct a power analysis. For simple regression, the degrees of freedom for the numerator is $n_b - 1$ (number of coefficients, including intercept, minus 1). You must include the intercept! So our $df_{numerator} = 2 - 1 = 1$. The power analysis is as follows:

```

     Multiple regression power calculation

              u = 1
              v = 23.6195
             f2 = 0.3333
      sig.level = 0.05
          power = 0.8
```

So, the results suggest that $v = 23.62$. We will round to 24. But what does this mean? The degrees of freedom, v, is: $df_{denominator} = N - n_b$ (total sample size minus total coefficients, including intercept). Thus, $24 = N - 2$ can be rearranged to $N = 24 + 2 = 26$. Thus, we recruit 26 individuals.

Participants: Participants will be undergraduate students aged 18–25 recruited from a Grenfell Campus through email invitations and campus advertisements. Inclusion criteria require that participants own an Apple Watch capable of tracking sleep stages. All participants will provide informed consent prior to participation.

Measures: Screen time before bed will serve as the independent variable and will be measured in minutes prior to sleep using each participant's device-based screen time tracking feature (measured for two hours prior to students self-imposed bedtime; range: 0-120mins). The dependent variable, REM sleep duration, will be measured in minutes using Apple Watch sleep tracking data. Participants will also complete a brief demographic questionnaire and report general sleep habits to account for potential confounding factors.

Procedure: After consent and baseline data collection, participants will record their screen time and wear their Apple Watch overnight. Participants will submit their screen time logs and Apple Watch sleep summaries via a secure online form. All data will be anonymized prior to analysis. Statistical analyses will include Pearson correlation and, if appropriate, multiple regression to examine the relationship between screen time and REM sleep duration while controlling for confounders.

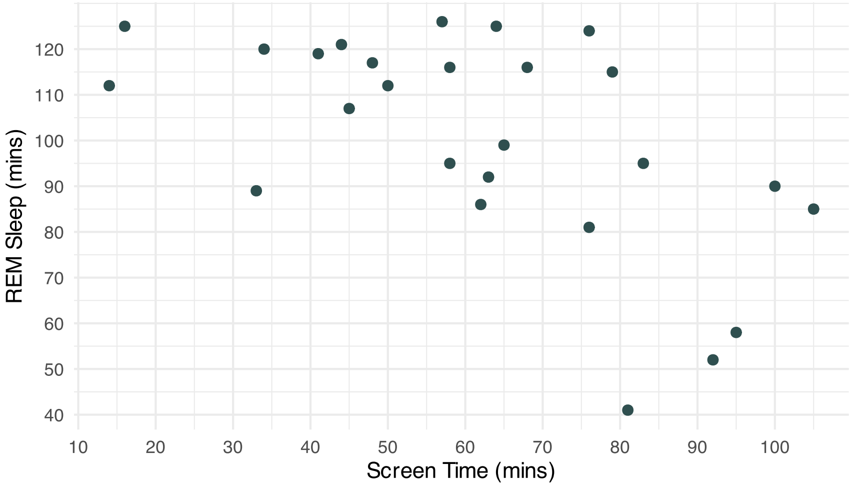
18.4.3 3. Collecting data

Our data is as follows:

ScreenTime	REM
64	125
79	115
50	112
83	95
48	117
45	107
63	92
14	112
57	126
92	52
62	86
16	125
34	120
68	116
76	124
100	90
41	119
76	81

ScreenTime	REM
105	85
58	116
33	89
81	41
65	99
44	121
58	95
95	58

And we can represent is as a scatterplot:



18.4.4 4. Analyzing data

We can use the formulas above to solve the regression equation. We will need the mean of the IV (Screen Time), mean of the DV (REM Sleep), their covariance, and the variances. These are as follows:

Mean_Screen	Mean_REM	Var_Screen	Var_REM	Cov
61.81	100.7	572.4	548.9	−321.3

Thus:

$$b_1 = \frac{cov_{(ST,REM)}}{s^2_{ST}} = \frac{-321.3015}{572.4015} = -0.5613$$

We interpret this as, for every 1-unit change in Screen Time (which was in minutes), we would predict a 0.5613 unit decrease in minutes of REM sleep. Thus, for every minute more of screen time, we would predict 0.5613 less minutes of REM sleep.

We must also solve for b_0.

$$b_0 = \overline{y} - b_1\overline{x} = 100.6923 - 61.80769(-0.5613) = 135.385$$

Interpreting intercept coefficients is relatively straight forward. We would predict someone with NO screen time before bed ($x = 0$) to get 135.34 minutes of REM sleep. Note: sometimes it makes *no sense* to interpret the intercept. For example, imagine a regression equation that predicts height using weight ($y_h = b_0 + b_w x_w + e_i$). We would interpret b_o, the intercept, as 'we would predict a height of XXX for someone with NO weight; that doesn't make sense!

We have our final equation!

$$y_i = 135.39 + (-0.561)(x_i) + e_i$$

18.4.4.1 Effect Size

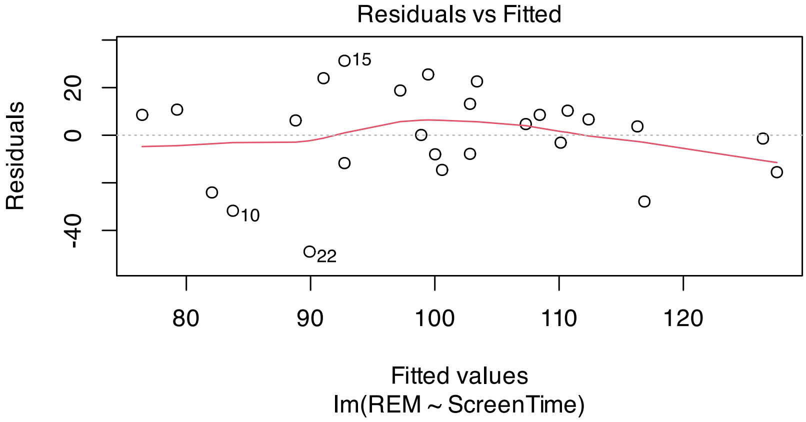
Effect size for simple regression is R^2, which is interpreted as the amount of variance in the outcome that is explained by the model. Since we have only one predictor, R^2 is simply the squared correlation between the one IV and the DV. The correlation between Screen Time and REM Sleep is −0.573233. Thus:

$$R^2 = (r)^2 = (-0.5732328)^2 = 0.329$$

Therefore, the model explains 32.9% of the variance in REM Sleep.

18.4.4.2 Our Assumptions

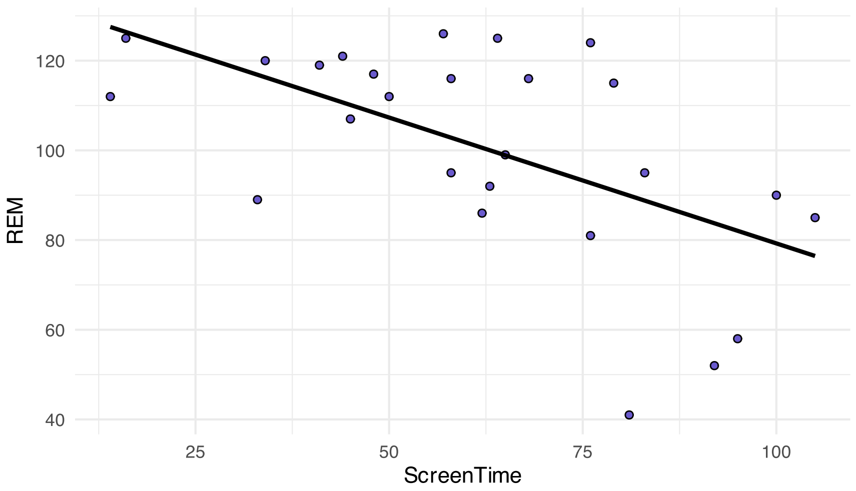
1. Homoscedasticity: The residual variance is constant across different levels/values of the IV. R can produce a plot of residuals across each fitted value of y.



Here, we want a relatively straight line around 0, indicating an mean residual of 0. Furthermore, we want the dots to be dispersed equally around each fitted value. It's hard to determine with our data because there are so few data points. However, we can be relatively confident that they are homoscedastic.

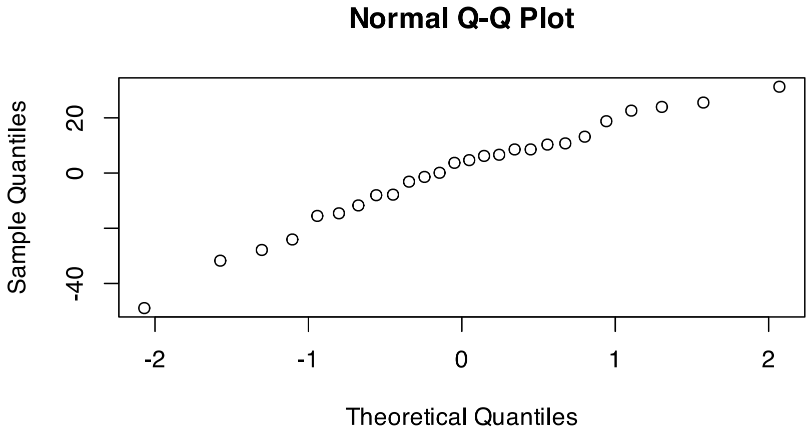
2. Independence: Each observation is independent; thus, each residual is independent. You must ensure this as a researcher. For example, if you had repeated measures (e.g., two observation from each person), then these would not be independent.

3. Linearity: The relationship between IV and DV is linear. We can visually assess this using a scatterplot. We hope that the points seem to follow a straight line. We can fit our line of best fit from OLS to help with this.



Our data appear linear.

4. Normality of residuals: We can asses using Q-Q plots and Shapiro-Wilk, which was covered in a previous chapter. *Remember, the null hypothesis of the SW test is that the data are normally distributed.*



```

	Shapiro-Wilk normality test

data:  our_model$residuals
W = 0.9659, p-value = 0.521
```

18.4.5 5. Write your results/conclusions

We fitted a linear model to predict REM Sleep with Screen Time. The model explains a statistically significant and substantial proportion of variance, $R^2 = 0.33$, $F(1, 24) = 11.75$, $p = 0.002$. Screen Time was a statistically significant and negative predictor of REM sleep, $b = -0.56$, $95\% \, CI[-0.90, -0.22]$, $t(24) = -3.43$, $p = 0.002$.

18.5 Conclusion

Simple regression takes a predictor and tries to explain the variance in an outcome. Often times researchers want to predict the value on one variable, knowing another. Consider these example: by knowing a student's social support, can we predict how many classes they will attend in university? By knowing how many classes someone attends, can we predict their grade in the course? In the upcoming chapter, we will add to simple regression to help answer more complex research questions. We will have more than one predictor and begin to model complex interactions among predictors.

18.6 Analysis in R

Regression and ANOVAs fall under the 'general linear model', which indicates that an outcome (e.g., y_i, DV) is the function of some linear combination of predictors (e.g., $b_1(x_i)$). We can use the `lm()` (linear model) function to write out our regression equation.

```
lm(REM ~ ScreenTime, data=sr_data)
```

Note that here, I have a data frame called *sr_dat* with two variables called *ScreenTime* and *REM*. The ~ symbol is the same as 'equal' or **predicted by**. So, we have REM is predicted by ScreenTime. R will automatically include an intercept and the error term.

The results of `lm(REM ~ ScreenTime, data=sr_data)` should be passed into a `summary()` argument. So, first, let's create our model!

And then pass that into the summary function:

```

Call:
lm(formula = REM ~ ScreenTime, data = sr_dat)

Residuals:
   Min     1Q Median     3Q    Max 
-48.92 -10.80   4.19  10.64  31.27 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  135.386     10.828   12.50  5.3e-12 ***
ScreenTime    -0.561      0.164   -3.43   0.0022 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.6 on 24 degrees of freedom
Multiple R-squared:  0.329,	Adjusted R-squared:  0.301 
F-statistic: 11.7 on 1 and 24 DF,  p-value: 0.0022
```

Another Way

The `apaTables` package also has a lovely output, and can save a word document with the output.

```


Regression results using REM as the criterion
 

   Predictor        b         b_95%_CI  beta    beta_95%_CI sr2
sr2_95%_CI
 (Intercept) 135.39** [113.04,
157.73]
  ScreenTime  -0.56**   [-0.90, -0.22] -0.57 [-0.92, -0.23] .33
[.05, .55]


```

```
      r              Fit

 -.57**
           R2 = .329**
       95% CI[.05,.55]


Note. A significant b-weight indicates the beta-weight and
semi-partial correlation are also significant.
b represents unstandardized regression weights. beta indicates
the standardized regression weights.
sr2 represents the semi-partial correlation squared. r
represents the zero-order correlation.
Square brackets are used to enclose the lower and upper limits
of a confidence interval.
* indicates p < .05. ** indicates p < .01.
```

While the regular `lm()` function gives exact p-values, the `apa.reg.table()` function gives more info such as CIs, r, sr, and effect size.

18.7 Practice Question

1. Generate the regression equation for the following data that investigating the Graduate Record Exams ability to predict GPA in graduate school.
2. Interpret the intercept and coefficient for GRE.
3. Write the hypotheses.
4. Write up the results.

Student	GRE	GPA
1	163	1.6
2	171	1.9
3	173	1.8

Student	GRE	GPA
4	139	3.1
5	174	3.9
6	139	1.7
7	162	1.6
8	141	3.6

18.8 Answers

1.

```
Call:
lm(formula = GPA ~ GRE, data = dat_prac)

Residuals:
   Min     1Q Median     3Q    Max 
-0.910 -0.744 -0.390  0.620  1.682 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.1708     3.9497    1.06     0.33
GRE          -0.0112     0.0249   -0.45     0.67

Residual standard error: 1.03 on 6 degrees of freedom
Multiple R-squared:  0.0327,    Adjusted R-squared:  -0.129 
F-statistic: 0.203 on 1 and 6 DF,  p-value: 0.668
```

$y_i = b_0 + b_1(x_{1i}) + e_i$

$y_i = -4.17 + (-0.011)(x_{1i}) + e_i$

2.

Intercept: Someone with a score of 0 on the GRE would be predicted to have a score GPA of 4.17 (this is impossible).

Slope: For every one unit increase in GRE score, we would predict a 0.011 unit decrease in GPA.

3.

$H_0 : b_1 = 0$

$H_A : b_1 \neq 0$

4.

We fitted a linear model to predict GPA with GRE. The model did not explain a statistically significant proportion of variance $R^2 = 0.03$, $95\% CI[.00, .41]$, $F(1, 6) = 0.20, p = 0.668$. The effect of GRE is statistically non-significant, $b = -0.01$, $95\% CI[-0.07, 0.05]$, $t(6) = -0.45$, $p = 0.668$.

19 Multiple Regression

This chapter will cover multiple regression, a statistical method used to examine the relationship between a dependent (or outcome/criterion) variable and multiple independent (predictor) variables. Unlike simple regression, which involves only one predictor, multiple regression allows researchers to assess the combined influence of several predictors on an outcome. But why do researchers need multiple predictors?

Using multiple predictors in regression allows researchers to better understand complex relationships between variables. Real-world psychological phenomenon are rarely caused or influenced by a single factor; instead, they result from multiple interacting influences. Furthermore, by including multiple predictors, we can control for confounding variables, improve the accuracy of our predictions, and gain a more comprehensive understanding of how different factors contribute to an outcome.

19.1 Some Additional Details

Multiple regression is useful in situations where we expect multiple factors to influence an outcome. For example, a researcher might want to predict job performance based on cognitive ability, motivation, and job experience.

The general form of the multiple regression equation is:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + e_i$$

where:

- Y is the dependent variable (outcome of interest),

- $X_1, X_2, ..., X_n$ are independent variables (predictors),
- β_0 is the intercept (value of Y when all predictors are 0),
- $\beta_1, \beta_2, ..., \beta_n$ are regression coefficients representing the effect of each predictor on Y,
- e is the error term.

19.2 Key Assumptions

Like all of our analyses thus far, a multiple regression analysis is valid model under the following assumptions (many we have already explored):

1. Linearity: The relationship between each predictor and the dependent variable should be linear.

2. Independence of Errors: Observations should be independent, and errors should not be correlated.

3. Homoscedasticity: The variance of errors should be constant across all levels of the independent variables.

4. Normality of Residuals: The residuals (errors) should be normally distributed.

5. No Multicollinearity: Predictor variables should not be highly correlated with one another. More to come.

Prior to further exploring our hypotheses and conducting a formal analysis, an explanation of various types of correlations is needed. Correlations help us understand the relationships between variables and are particularly important in multiple regression, where we assess the contribution of multiple predictors to an outcome variable. We have explored some of these, but revisit them so they are fresh in your mind.

19.3 Types of Correlations

19.3.1 Pearson Correlation Coefficient

We have already explored correlation, r, in a previous chapter. Recall that when we square the correlation, we obtain the **coefficient of determination** (R^2), which indicates the proportion of variance in one variable that is accounted for/explained by (**not to be confused with CAUSED BY**) the other. This provides insight into how strongly two variables are related, but it does not imply causality. Recall that one formula for the Pearson correlation coefficient is:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

where:

- X_i and Y_i are individual data points,
- $\bar{X}$ and $\bar{Y}$ are the means of the variables,
- The numerator represents the covariance between X and Y,
- The denominator standardizes the covariance by dividing by the product of the standard deviations.

To visualize the coefficient of determination, consider the following two variables: x_1 and y.

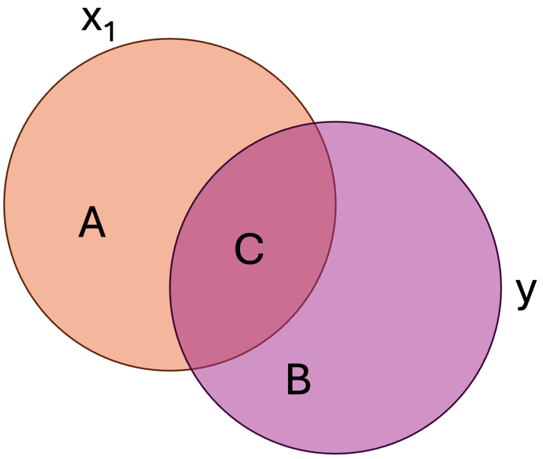


Figure 14: Venn diagram.

Here:

- $B + C$ represents the total variance in y, or σ_y^2
- C represents the variance in y that is accounted for/explained by x_1, or R^2
- B represents the variance that is unaccounted for in the model

19.3.2 Semi-Partial Correlation (Part Correlation)

The semi-partial correlation, denoted as $r_{x1(y.x2)}$ (where we have three variables: x_1, x_2, and y), measures the *unique* relationship between one predictor and the outcome variable **after controlling for the effects of another predictor on that predictor** (but not on the outcome variable). This is particularly useful when we want to understand how much unique variance a predictor contributes to the dependent variable **without adjusting for other predictors' influence on the outcome**.

For example, if we are examining the relationship between study hours (x_1) and exam scores (y), while controlling for prior GPA (x_2), the semi-partial correlation tells us how much variance in exam scores is *uniquely* explained by study hours that is **not shared with prior GPA**.

In short, it is the variance uniquely explained relative to **all** of criterion. Let's visualize this regression model, wherein we have two predictors, x_1 and x_2, predicting the criterion, y:

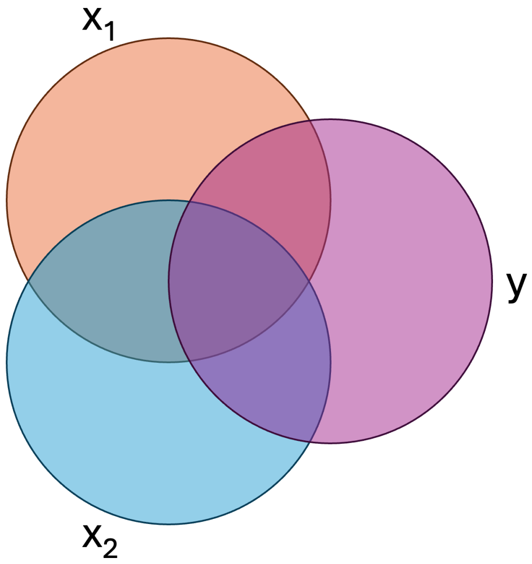


Figure 15: Semi-partial correlation.

We can assign each region of the above figure a letter:

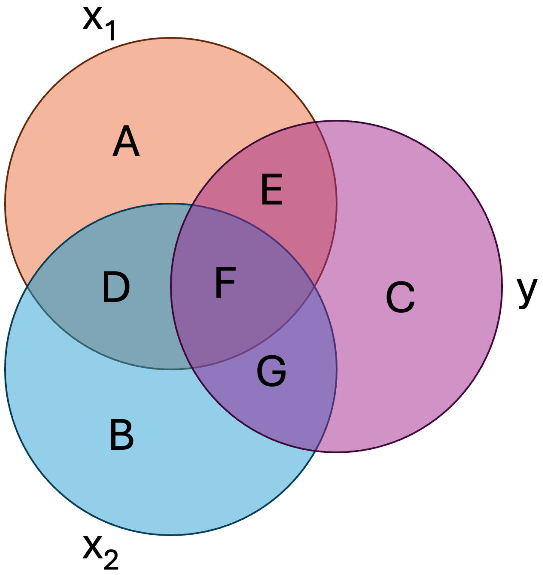


Figure 16: Sections of semi-partial correlation.

In this figure:

- $E + F + G + C$ represents the **total** variance in y, or σ_y^2
- $E + F + G$ represents the **total** variance explained by the model, or R^2

- C represents the **unaccounted** for/unexplained variance in y, or $1 - R^2$
- $E + F$ represents the **total** variance in y explained by x_1, or $r^2_{x_1 y}$
- E represents the **unique** variance in y explained by x_1
- $F + G$ represents the **total** variance in y explained by x_2, or $r^2_{x_2 y}$
- G represents the **unique** variance in y explained by x_2
- F represents the **shared** variance in y explained by x_1 and x_2

In the regular Pearson correlation, $E + F$ would have been considered the variance in y explained by x_1. However, some of this variance is also explained by x_2. One way to represent this is by the squared semi-partial or part correlation (note: we are squaring it to give us the 'proportion of variance', just as we did in correlation). The squared semi-partial/part correlation of x_1 and y would be:

$$r^2_{x_1(y.x_2)} = \frac{E}{C + E + F + G}$$

Or, simply:

$$r^2_{x_1(y.x_2)} = R^2 - r^2_{x_2 y}$$

In the above formula we are, essentially, saying that $r^2_{x_1(y.x_2)}$ is the difference between R^2, the total variance in y and r^2_{x2y}, the total variance in y explained by x_2. Logically, this means that in this two predictor model, any variance in y that is not explain by x_2 is uniquely explained by x_1. Note the differences in the notation between the Pearson correlation between the two, $r_{x_1 y}$, and the part correlation that accounts for $x2$, $r_{x_1(y.x_2)}$.

! Practice

1. What regions would represent the squared semi-partial/part correlation of x_2 and y?
2. What would be the mathematical formula?

! Answers

1. $r^2_{x_2(y.x_1)} = \frac{G}{C+E+F+G}$
2. $r^2_{x_2(y.x_1)} = R^2 - r^2_{x_1 y}$

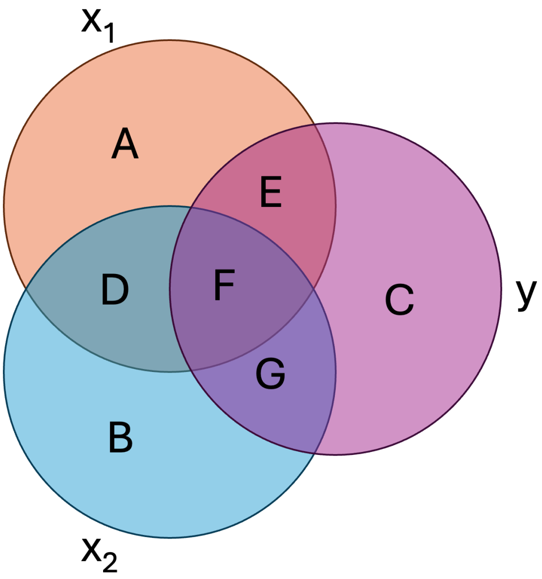
19.3.3 Partial Correlation

The partial correlation, denoted as or $r_{x1y.x_2}$, assesses the direct relationship between a predictor and the outcome variable **after controlling for the influence of other predictors on both the predictor and the outcome**. This differs from the semi-partial correlation because it removes the effect of control variables or other predictors (in our above example, x_2) from both the predictor of interest and the outcome.

Mathematically, the squared partial correlation, $r^2_{x1y.x_2}$, tells us the proportion of variance in y that is uniquely explained by x_1 after removing the influence of all other predictors. **In short, it is the variance uniquely explained relative to the unexplained variance of the criterion.**

The squared partial correlation of x_1 and y would be:

$$r^2_{x_1 y.x_2} = \frac{E}{C+E}$$



Or, mathematically:

$$r^2_{x_1 y.x_2} = \frac{R^2 - r^2_{x_2 y}}{1 - r^2_{x_2 y}}$$

Both partial and semi-partial correlations help us understand how an independent variable (IV) relates to the dependent variable (DV) while accounting for other variables in a regression model. However, they answer slightly different questions. Here is a quick reference to help you.

1. Partial Correlation → "What is the pure relationship between this predictor and the outcome?"

- It tells you how much an IV is related to the DV after removing the influence of other IVs from both the predictor and the outcome.
- Example: If you're studying how stress affects exam scores while controlling for sleep, the partial correlation tells you the direct relationship between stress and scores as if sleep was completely removed from the equation for both stress and scores.

2. Semi-Partial (Part) Correlation → "How much does this predictor add to the model's ability to predict the outcome?"

- It tells you how much an IV uniquely contributes to explaining the DV without adjusting the DV itself.
- Example: If you add stress as a predictor to your exam scores model (which already includes sleep), the semi-partial correlation tells you how much extra variance in exam scores is explained just by stress (after removing overlap with sleep in stress but not in the scores).

In this class we will primarily use the semi-partial/part correlation–mostly the squared semi-partial correlation–in our regression analyses. With this in mind, let's continue with a practical example involving our favorite musician, Taylor Swift.

19.4 Regression…you can do it with a broken heart.

Taylor Swift and her team are consulting you, a research expert, to help determine what features of music determine the popularity it achieves. She hopes to use your findings to write new music. Specifically, she is interested in knowing whether certain characteristics of music are more likely to get played on Spotify. Taylor and her team have a theory that they have called the **"Rhythmic Positivity Theory"**. This proposes that songs with higher danceability and happier tones are more popular because they elicit positive emotions and encourage social engagement. Taylor also has specific hypotheses: *that both both positively valenced (i.e., happy) and danceable songs will be more popular*.

19.5 1. Generating hypotheses

In regression you hypothesize about coefficients, typically referred to as βs (beta). Other times you may hypothesize about the full model (i.e., variance explained in the outcome; R^2). Thus, we could have two different sets of hypotheses. The most common will refer to coefficients. Here, we could convert our text-based hypotheses to statistical hypotheses:

$$H_0 : \beta_{dance} = 0 \text{ and } \beta_{valence} = 0$$

and

$$H_A : \beta_{dance} \neq 0 \text{ and } \beta_{valence} \neq 0$$

ℹ None

More generally, you would simply have:

$$H_0 : \beta_s = 0$$

$$H_A : \beta_s \neq 0$$

We will use $\neq$ for these alternative hypotheses because they could be positive or negative and we are doing a two-tailed test. The second type of hypotheses we may propose have to do with the *full model*, and how well it accounts for variance in the outcome (i.e., DV/criterion):

$$H_0 : R^2 = 0$$

$$H_A : R^2 > 0$$

We use $>$ for this hypothesis because R^2 cannot be negative.

While these are our main hypotheses, we should also try to conceptualize our study's model. Our model can be represented as follows:

$$y_i = \beta_0 + \beta_{dance}(x_{dance,i}) + \beta_{valence}(x_{valence,i}) + e_i$$

Where:

- $x_{dance,i}$ is individual i's score on danceability
- $x_{valence,i}$ is individual i's score on valence
- β_0 is the intercept of the model
- β_{dance} is the coefficient for danceability
- $\beta_{valence}$ is the coefficient for valence
- e_i is individual i's error

19.6 2. Designing a study

While Taylor has given you a $3,000,000 budget, you decide to put that money in you RRSP, cheap out, and collect publicly-available data from Spotify. You decide that you will collect a random sample of songs from Spotify and use a computer to estimate the valence and danceability of the songs. These are both measured as continuous variables. You decide to use a regression to determine the effects of both variables on a song's popularity (number of plays on Spotify in 2025, in millions). All variables are continuous (although regression can handle most variable types; ANOVA is just a special case of regression).

You do a power analysis and determine you need a sample of approximately 50 songs.Prior to conducting your research, you submit your research plan to the Grenfell Campus research ethics board, which approves your study and classified it as low-risk.

19.7 3. Collecting data

You follow through with your research plan and get the following data:

track_artist	track_name	Popularity	Valence	Danceability
Camila Cabello	My Oh My (feat. DaBaby)	208	13	2
Tyga	Ayy Macarena	95	17	7
Maroon 5	Memories	200	23	4
Harry Styles	Adore You	62	9	9
Sam Smith	How Do You Sleep?	189	22	4
Tones and I	Dance Monkey	87	19	7
Lil Uzi Vert	Futsal Shuffle 2020	140	1	5
J Balvin	LA CANCIÓN	144	14	5
Billie Eilish	bad guy	129	6	7
Dan + Shay	10,000 Hours (with Justin Bieber)	141	8	5
Regard	Ride It	191	13	3
Billie Eilish	bad guy	175	19	5
The Weeknd	Heartless	126	13	4
Y2K	Lalala	112	18	5
Future	Life Is Good (feat. Drake)	36	16	8
Lewis Capaldi	Someone You Loved	76	17	6

track_artist	track_name	Popularity	Valence	Danceability
Anuel AA	China	185	22	4
Regard	Ride It	162	17	5
Dua Lipa	Don’t Start Now	144	7	4
Anuel AA	China	144	12	5
Regard	Ride It	133	13	5
Bad Bunny	Vete	107	13	6
Roddy Ricch	The Box	142	7	2
Juice WRLD	Bandit (with YoungBoy Never Broke Again)	109	12	7
Roddy Ricch	The Box	247	25	4
Regard	Ride It	133	14	5
Trevor Daniel	Falling	149	11	4
Anuel AA	China	190	15	3
Shawn Mendes	Señorita	140	8	4
Travis Scott	HIGHEST IN THE ROOM	186	10	3
Juice WRLD	Bandit (with YoungBoy Never Broke Again)	164	21	4
Camila Cabello	My Oh My (feat. DaBaby)	135	22	6
Sam Smith	How Do You Sleep?	113	12	5
Harry Styles	Adore You	129	10	4
Don Toliver	No Idea	53	20	7
Billie Eilish	everything i wanted	133	20	5

track_artist	track_name	Popularity	Valence	Danceability
Lil Uzi Vert	Futsal Shuffle 2020	65	21	7
DaBaby	BOP	111	16	5
Lil Uzi Vert	Futsal Shuffle 2020	18	23	8
blackbear	hot girl bummer	166	17	4
Tones and I	Dance Monkey	198	13	2
Tyga	Ayy Macarena	87	13	6
Selena Gomez	Lose You To Love Me	113	11	5
Dalex	Hola - Remix	106	15	5
The Black Eyed Peas	RITMO (Bad Boys For Life)	100	11	8
Arizona Zervas	ROXANNE	116	11	6
The Black Eyed Peas	RITMO (Bad Boys For Life)	111	4	6
Arizona Zervas	ROXANNE	101	14	6
Roddy Ricch	The Box	57	13	8
MEDUZA	Lose Control	119	23	6

19.8 4. Analyzing data

19.8.1 Matrix Algebra

Matrix algebra can be used to 'solve' our regression equation. However, we will not use matrix algebra to solve our regression coefficients in this class. For those interested, we *could* using the following (see here for more information):

$$(X'X)^{-1}X'Y$$

Where X is a n (number of observations) by v (number of predictors, including intercept) matrix of scores. The score on the 'intercept' is 1 for all observations. Y is a n by 1 vector of scores on the DV.

The results from our matrix algebra would work out to:

$$\begin{bmatrix} 237.42 \\ 1.09 \\ -23.79 \\ \\ \end{bmatrix}$$

Where each row is β_0 to β_2. Thus, the equation would be:

$$y_i = 237.42 + 1.09(x_{valence,i}) + (-23.79)(x_{dance,i}) + e_i$$

When we had one variable, we could effectively visualize a line of best fit. We can visualize a 'plane' of best fit when we have two predictors.

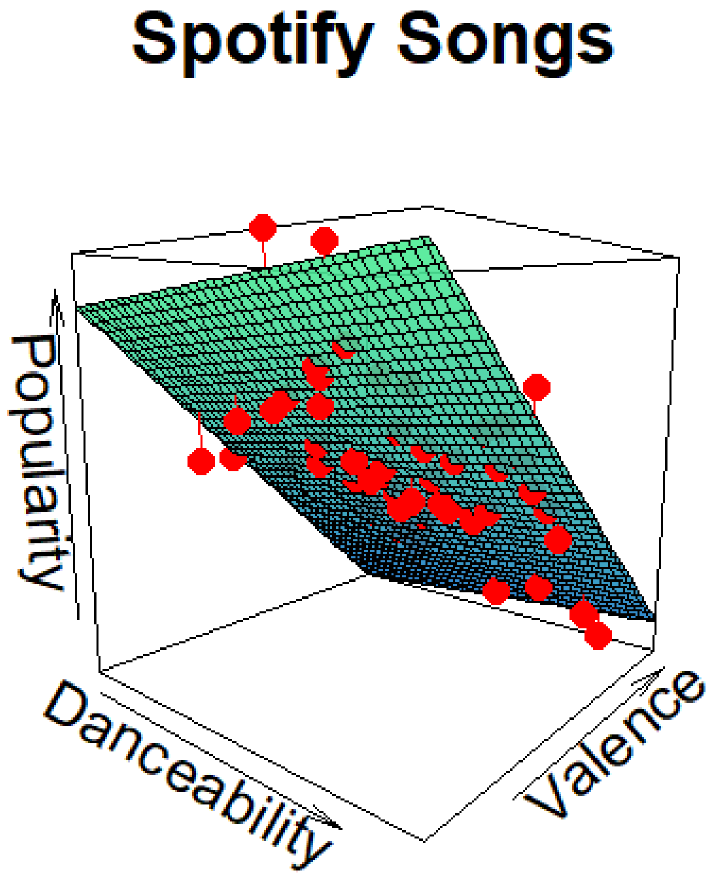


Figure 17: 3D Scatterplot.

As we now have more variables, the visualization becomes difficult. We struggle to interpret anything beyond 3D!

19.9 SST

Like simple regression, sum of squares total (SST) represents the difference between the observed scores on the outcome/criterion and the mean of the outcome/criterion.

$$SST = \sum (y_i - \overline{y})^2$$

19.10 SSE

Like simple regression, the sum of squares error/residual (SSE) represents the difference between the observed scores on the outcome/criterion and the predicted values of the outcome/criterion.

$$SSR = \sum (y_i - \hat{y_i})^2$$

19.11 SSR

Like simple regression, the sum of squares regression/model (SSR) represents the difference between the predicted values on the outcome/ criterion and the mean of the outcome/criterion

$$SSM = \sum (\hat{y_i} - \overline{y})^2$$

For us, .

Given these, we can calculate the MSR (mean square of the regression; with $p-1$ degrees of freedom; p being the number of b coefficients) and MSE (mean square error; with $n-p$ degrees of freedom) and calculate the appropriate F-statistic.

$$MSR = \frac{74669.74}{2} = 37334.87$$

and

$$MSE = \frac{33118.68}{47} = 704.65$$

and

$$F = \frac{MSR}{MSE} = \frac{37334.87}{704.65} = 52.98$$

And you can look up the associated p-value in any standard critical F table. Or R can calculate it for us using `pf(q=52.98, df1=2, df2=47)` (the probability of F with our given).

19.12 Effect Size - R^2

Like simple regression, we can calculate an effect size (R^2). We can calculate this using:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{33118.68}{107788.42} = .69$$

So how might we interpret this in the context of our original hypotheses? First, consider R^2. We can conclude that the model explains a statistically significant and substantial proportion of variance in popularity $R^2 = 0.69$, $95\% CI[.52, .77]$, $F(2, 47) = 52.98$, $p < .001$, $R^2_{adj} = 0.68$).

Second, consider the hypotheses regarding the unique predictive ability of each individual predictor, which concerns each's sr^2. We can conclude that Valence is not a statistically significant predictor of song popularity, $b = 1.09, p = .126, sr^2 = .02, 95\%\, CI[-.02, .05]$. However, Danceability was a statistically significant predictor of song popularity, $b = -23.79, p =< .001, sr^2 = .69, 95\%\, CI[.54, .83]$. Thus, for every 1-unit change in Danceability, a song's popularity is expected to decrease by 23.79, *while holding all other predictors constant.*

This later piece is important for interpreting regression models. A predictor's impact is dependent on holding all other aspects of the model constant. If I added a new predictor, the whole model would likely change, including the Danceability coefficient. If I removed the Valence predictor from the model, even though it was not statistically significant, I would expect the Danceability regression coefficient to change.

19.12.1 Measures of Fit

19.12.1.1 R^2

Our effect size is similar to simple regression and represents the proportion of variance the model explains in the outcome. It represents the total contribution of all predictors and is multiple R^2 (multiple given multiple predictors).

As discussed, R^2 can be calculated as:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{33118.68}{107788.42} = .69$$

R^2 can never go down when we add more predictors. Thus, getting large values for models with lots of variables in unsurprising. However, it does not indicate that any single predictor is doing a good job at uniquely predicting the outcome. More to come on this.

19.12.1.2 Adjusted R^2

While R^2 measures the proportion of variance explained by the model, it has a known limitation: it always increases (or stays the same) when more predictors are added, even if those predictors do not meaningfully contribute to explaining the outcome. To account for this, **Adjusted** R^2 adjusts for the number of predictors in the model, penalizing excessive complexity.

The formula for Adjusted R^2 is:

$$R^2_{\text{adj}} = 1 - \left(\frac{SSE/(n-p-1)}{SST/(n-1)} \right)$$

Applying this formula:

$$R^2_{\text{adj}} = 1 - \left(\frac{33118.68/(50-2-1)}{107788.42/(50-1)} \right) = .68$$

Unlike R^2, Adjusted R^2 can **decrease** if a new predictor does not improve model fit beyond what would be expected by chance. This makes it a more reliable metric when comparing models with different numbers of predictors.

19.12.1.3 AIC

Akaike Information Criterion (AIC) is a fit statistic we can use for regression models (and more). The major benefit of AIC is that is penalizes models with many predictors.

$$AIC = n \ln\left(\frac{SSE}{n} \right) + 2k$$

For our above model:

$$AIC = 50 \ln \frac{16224}{50} + 2(4) = 297.11$$

Is this good? Bad? Medium? Hard to say. The smaller the number the better.

19.12.2 Assumptions

The assumptions for multiple regression are similar to those in simple regression, with one key addition: **multicollinearity**.

19.12.2.1 Multicollinearity

Multicollinearity occurs when two or more predictors in the model are highly correlated, making it difficult to determine their unique contribution to the outcome. This can inflate standard errors, leading to unstable estimates and misleading significance tests.

A common way to check for multicollinearity is by calculating the **Variance Inflation Factor (VIF)**. Most statistical software will provide VIFs, such as this:

Observations	50
Dependent variable	Popularity
Type	OLS linear regression

F(2,47)	52.98
R²	0.69
Adj. R²	0.68

	Est.	S.E.	t val.	p	VIF
(Intercept)	237.42	15.61	15.21	0.00	NA
Valence	1.09	0.70	1.56	0.13	1.01
Danceability	−23.79	2.32	−10.26	0.00	1.01
Standard errors: OLS					

A **VIF > 10** suggests severe multicollinearity, though some researchers use a lower threshold (e.g., **VIF > 5**). If multicollinearity is detected, possible solutions include **removing redundant predictors**, **combining highly correlated variables**, or **using ridge regression** to stabilize estimates.

19.13 5. Write your results/conclusions

We conducted a multiple regression analysis to examine the association between Valence, Danceability, and Popularity. The overall model was statistically significant, suggesting that Valence and Danceability explain a substantial proportion of the variance in Popularity, $R^2 = 0.69$, $95\% CI[0.52, 0.77]$, $F(2, 47) = 52.98$, $p < .001$.

Examining individual predictors, the effect of Valence on Popularity was positive but not statistically significant, $b = 1.09, 95\% CI[-0.32, 2.50]$, $t(47) = 1.56$, $p = 0.126$. The squared semi-partial correlation (sr^2) was small and non-significant, $sr^2 = 0.02$, $95\% CI - 0.02, 0.05]$.

Conversely, Danceability had a statistically significant negative effect on Popularity, $b = -23.79$, $95\% CI[-28.45, -19.12]$, $\beta = -0.83$, $95\% CI[-1.00, -0.67]$, $t(47) = -10.26$, $p < .001$. Thus, for every 1-unit increase in Danceability, we would expect a −23.79-unit decrease in popularity. Additionally, the squared semi-partial correlation was substantial, $sr^2 = 0.69$, $95\% CI[0.54, 0.83]$, indicating Danceability uniquely explained a large proportion of the variance in Popularity.

These results suggest that Danceability is a strong negative predictor of Popularity, while Valence does not significantly contribute to the prediction of Popularity when controlling for Danceability.

19.14 Conclusion

Multiple regression is a power tool for your statistical toolbox. Using numerous predictors, with different types of predictors, we can explain variance in an outcome of interest. Allowing for multiple predictors can increase the complexity of phenomenon we study; indeed, many phenomenon have multiple interacting influences in the real world.

In the next chapter, we extend regression to difference variable types and their interactions.

19.15 Regression in R

There are multiple ways we can run a regression in R. We will use the basic `lm()` function that we used in the simple regression chapter.

```
taylors_model <- lm(Popularity ~ Valence +Danceability,
data=taylor)
```

and the summary of that model:

Observations	50
Dependent variable	Popularity
Type	OLS linear regression

F(2,47)	52.98
R²	0.69
Adj. R²	0.68

	Est.	S.E.	t val.	p
(Intercept)	237.42	15.61	15.21	0.00
Valence	1.09	0.70	1.56	0.13
Danceability	−23.79	2.32	−10.26	0.00
Standard errors: OLS				

Also recall that the `apaTables()` package provides some additional information that is useful for our interpretation.

```
Regression results using Popularity as the criterion


    Predictor        b         b_95%_CI  beta    beta_95%_CI
sr2  sr2_95%_CI
  (Intercept) 237.42** [206.02,
268.83]
      Valence     1.09    [-0.32, 2.50]  0.13  [-0.04,
0.29] .02 [-.02, .05]
 Danceability -23.79** [-28.45, -19.12] -0.83 [-1.00,
-0.67] .69  [.54, .83]




      r             Fit

    .06
 -.82**
            R2 = .693**
        95% CI[.52,.77]



Note. A significant b-weight indicates the beta-weight and
semi-partial correlation are also significant.
b represents unstandardized regression weights. beta indicates
the standardized regression weights.
sr2 represents the semi-partial correlation squared. r
represents the zero-order correlation.
Square brackets are used to enclose the lower and upper limits
of a confidence interval.
* indicates p < .05. ** indicates p < .01.
```

20 Multiple Regression - Additional Considerations

This chapter will cover a few additional considerations for multiple regression analyses. First, we will cover some diagnostic tests to assess the validity of your analyses, including adjustments to your data–if necessary. Last, we will cover types of variable entry methods for multiple regression. These concepts will further your understanding of multiple regression and increase the breadth of hypotheses you can test.

20.1 Diagnostics

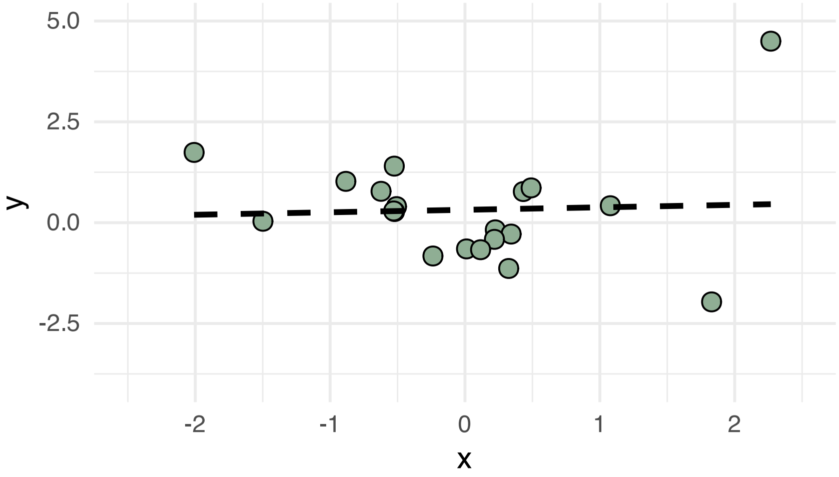
A Diagnostics assess whether our regression model meets key assumptions, ensuring that our results are valid and interpretable. Importantly, any violations of key assumptions can lead to biased estimates and incorrect conclusions. Thus, diagnostics are not something that can be ignored. We will visit two major diagnostics that focus on detecting outliers and/or influential cases.

20.1.1 Outliers

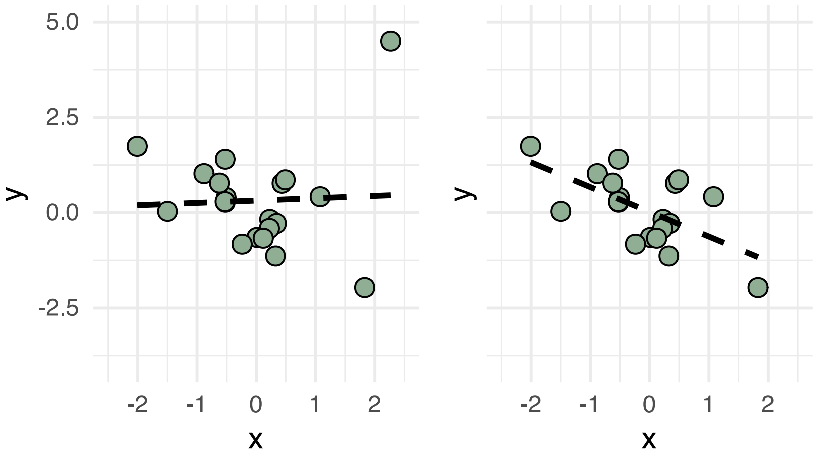
Outliers are observations—such as an individual's score on a variable or multiple variables—that differ substantially from the rest of the data. These values stand out because they are unusually high or low compared to the majority of observations. Outliers can occur for several reasons. First, outliers may be due to **unique characteristics of the individual**. Sometimes, an outlier reflects a genuine difference in one individual in a study. For example, imagine a study measuring depression severity among college students. If one student has a diagnosed mood disorder and experiences severe symptoms, their depression score may be much higher than the rest of the sample. Second, outliers may be due to **measurement or data entry errors**. Here, outliers result from mistakes such as typing errors or incorrect coding of responses. Imagine a participant who is entering their age in an online survey, but puts their year of birth (2005) instead of age in years (20). The mean age of participants may be concernedly high with this error not remediated. Last outliers may results from **sampling issues**. If the sample includes individuals who do not represent the target population, their scores may appear extreme relative to others.

Outliers matter because they can influence statistical analyses. For instance, extreme values can distort the mean of a variable, inflate variance estimates, and affect correlation or regression results. Thus, researchers should examine outliers carefully to decide whether they should be retained, transformed (i.e., value changed), or removed.

Consider the following data that is presented on a scatterplot with a OLS line of best fit:



Do you notice any data points that seem to not fit the trend of the data? The point on the top right seems to be an exceptional case–unlike the others. Let's investigate what happens the OLS LOBF when this case is removed from the analysis. The original figure is presented on the left, the new one on the right:



Hopefully you can see how the removal of this potential outlier has reduced the magnitude of the residuals. In other words, our model has less error. While in this example we can be quite confident of the presence of an outlier, there are more objective ways for us to tests this.

There are many ways to detect outliers. We will focus on one major method: standardized residuals. To understand these, we must familiarize ourselves with residuals. Recall that in regression analyses we have our observed outcome (y_i) and our predicted outcome ($\hat{y}_i$) (i.e., what the regression model predicts the data to be). This difference was the residual/error:

$$e_i = y_i - \hat{y}_i$$

A **standardized residual** is one way to represent how far an observation deviates from the model's prediction. It takes the errors of a regression model and standardized them by dividing by the standard deviation of the residuals. Thus, we can measure a standardized residual with:

$$\text{Standardized Residual} = \frac{e_i}{\sqrt{MSE}}$$

As a general rule of thumb, larger residuals indicate potential outliers. Like any normal distribution, we expect the distribution of errors to have 95% of the distribution within ± 2.58 and 99.9% within ± 3.29. Thus, should a residual fall beyond those values, we can conclude that it is an extremely unlikely value and a potential candidate to remove from our data. In our data, we can estimate the predicted score by filling in the blanks in the regression equation. The regression equal solves to be (using OLS):

$$\hat{y_i} = 0.3189 + 0.06126(x_i)$$

Consider person 10, who's score on the outcome is 4.5 and score on the predictor is around 2.27. Their observed outcome is 4.5 and their predicted outcome is:

$$\widehat{y_{10}} = 0.3189 + 0.06126(2.27) = 0.4579$$

Thus, their residual is:

$$e_{10} = y_{10} - \widehat{y_{10}} = 4.5 - 0.4579 = 4.042$$

When we calculate the regression model we get a mean square error of 1.86767. Thus, the standardized residual is:

$$\text{Standardized Residual} = \frac{4.042}{\sqrt{1.86767}} = \frac{4.042}{1.3666} = 2.958$$

Sometimes your statistical software will calculate **studentized residuals**, which are calculated as:

$$r_i = \frac{e_i}{\widehat{\sigma_e}\sqrt{1 - h_{ii}}}$$

Where $\hat{y}_{i(i)}$ is the fitted value for i from the model excluding case i. Studentized residuals are often considered a more robust indicator of influence on a regression model.

In our data from above (including all cases), the top five studentized residuals, sorted by magnitude, are:

ID	y	x	.fitted	.resid	.std.resid
10	4.5	2.2678	0.4578	4.042	3.5886
2	−1.964	1.8287	0.4309	−2.395	−1.992

ID	y	x	.fitted	.resid	.std.resid
1	1.7416	−2.0087	0.1958	1.546	1.3169
17	−1.1355	0.3241	0.3388	−1.474	−1.11
9	−0.8263	−0.2376	0.3043	−1.131	−0.8502

It seems that participant 10–the top right point–has a high standardized residual of 3.59.

20.1.2 Influential Observations/Cases

Influential observations are data points that have a disproportionate impact on the estimates of a statistical model–such as regression coefficients (β). In other words, these cases can significantly change the results of an analysis simply by being included.

For example, in regression analysis, most observations contribute *modestly* to the overall fit of the model. However, an influential case might pull the regression line toward itself, altering slope estimates and predicted values. This influence is not just about being an outlier in terms of the outcome variable. Instead, it depends on the combination of predictor values and their leverage in the model.

Researchers typically assess influence using diagnostic measures such as Cook's Distance of DFBETAs.

1. Cook's Distance: Evaluates how much the regression coefficients would change if a particular observation were removed. Higher values indicate more influence on the regression coefficients. Often times a distance ≥ 4 is concerning. Cook's distance is calculated using:

$$D_i = \frac{e_i^2}{p\,\hat{\sigma}^2} \cdot \frac{h_{ii}}{(1 - h_{ii})^2}$$

Where:

- p is the number of predictors (including the intercept if counted),
- $\hat{\beta}_{j(i)}$ denotes the coefficient with observation/case i deleted

Luckily for us, most statistical software will provide these for you. For example, the `augment()` function from the `broom` package in r. Here are five highest Cook's distance values for our data:

ID	y	x	.fitted	.cooksd
10	4.5	2.2678	0.4578	3.03966
2	−1.964	1.8287	0.4309	0.57936
1	1.742	−2.0087	0.1958	0.30844
17	−1.136	0.3241	0.3388	0.03622
12	1.402	−0.5243	0.2868	0.02452

Again, Cook's distance for observation 10 appears problematic.

2. DFBETAs: assess the influence of individual observations on the estimated regression coefficients. For each observation and each coefficient in the model, DFBETA measures the difference between the coefficient estimated with all data and the coefficient estimated when that observation is removed. Observations that significantly alter coefficients when included can distort the model's interpretation and predictions.

A large absolute DFBETA value indicates that the observation has a strong influence on that particular coefficient. There are some go to rules of thumb for interpreting DFBETAs. For example, some consider $|\text{DFBETA}| > \frac{2}{\sqrt{n}}$ (where n = number of observations) may be considered influential. Others consider values greater than absolute 1 concerning.

Again, like Cook's distance, most statistical software packages will provide these (or give the option to provide these). The following is from the `dfbeta()` function from the `stats` package in r. We will get one column for each coefficient (here, 2: one for the intercept and one for x). Here are some the higher DFBETAs for our data:

ID	X.Intercept.	x
10	0.29752	0.71022
2	−0.15472	−0.29783
1	0.10478	−0.22155
17	−0.07805	−0.02663

ID	X.Intercept.	x
9	−0.0597	0.01493

Oh observation 10! What will we do with you. All diagnostics are leading towards removing this observation from the data. A recommended practice is to provide the results for two regression analyses: one with and one without the influential cases. Perhaps the results won't be that different after all.

20.2 Assumptions

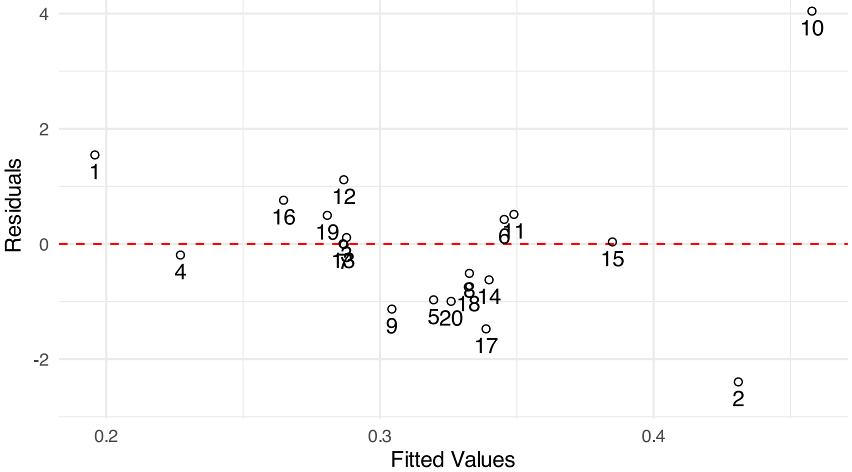
To ensure the validity of a multiple regression model, several key assumptions must be met. **First**, the variables should an appropriate types. Predictors can be continuous or binary categorical. Additionally, categorical predictors with more than two categories must be dummy coded. Furthermore, the dependent variable (DV) should be continuous and unbounded, as ceiling or floor effects can restrict variability.

Second, predictors must exhibit non-zero variance. If all participants score identically on a predictor, it cannot explain any variation in the DV.

Third, there should be no perfect multicollinearity among predictors. While some correlation between predictors is acceptable, extreme multicollinearity can distort results. This can be assessed using the Variance Inflation Factor (VIF), calculated as $\text{VIF} = \frac{1}{1-R^2}$. A VIF greater than 10 indicates severe multicollinearity, while values above 5 warrant caution. Your statistical software can provide VIFs for each predictor. In our example there are no VIFs; with only one predictor, there's not other predictor for it to be correlated with!

Fourth, omitted variable bias should be considered. You should ensure you are not excluding an important predictor that may influence the DV. Having good theory-derived hypotheses can help prevent this.

Fifth, the assumption of homoscedasticity requires that residual variance remains consistent across all predictor values; systematic increases or decreases indicate heteroscedasticity. An easy way to assess this is to plot residuals and fitted values. The residuals should be consistently dispersed across all levels of fitted values.



Our regression has few observations, so it is harder to assess heteroscedasticity. Most points look good, except observation 10.

Sixth, error terms should be independent, meaning residuals are uncorrelated.

Seventh, residuals should follow a normal distribution, which can be checked using the Shapiro-Wilk test or visually with a QQ plot.

Finally, the relationship between predictors and the DV should be linear. This was discussed in a previous chapter. If non-linear patterns exist, transformations such as x^2 or $\frac{1}{x}$ may be necessary.

20.3 Types of Regression

In this last section, we will cover various types of entry methods for regression. When dealing with multiple variables, sometimes researchers do not want to put all of the variables in the model at once. Instead, they may wish to put in subsets in sequentially. The term **block** is often used to describe predictors or a set of predictors that get put into the model. Why would one want to add blocks sequentially in regression? One could add blocks of predictors to determine changes in R^2 and sr^2.

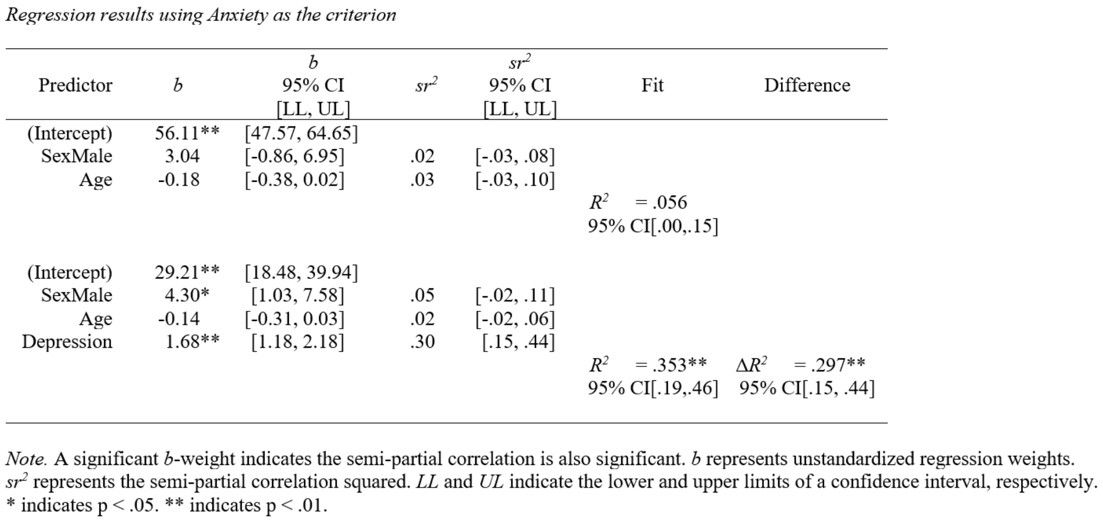
For example, enter one block, get an R^2 and then add a second to assess the changes in R^2. Any significant increase will highlight that the second block of variables are seemingly important predictors, as they explain more variance in the outcome. The following are some common entry methods.

20.3.1 Hierarchical

Hierarchical regression is theory-driven and widely used in psychological research. Predictors are entered in *blocks*, based on conceptual importance or prior evidence. Known predictors are typically entered first, followed by additional variables to assess their incremental contribution to the regression model (i.e., do they explain more variance). This approach allows researchers to test hypotheses about whether new predictors improve model fit after controlling for established variables.

For example, suppose we aim to predict anxiety symptoms. In Block 1, we enter demographic variables (age, sex). In Block 2, we add depressive symptoms to determine whether they explain additional variance in anxiety beyond demographics.

Hierarchical regression results can often be nicely summarized in tables, with each block having it's own section. For example, the following is output from the `apa.reg.table()` function from the `apaTables` package.



Regression results using Anxiety as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	*sr*^{2}	*sr*^{2} 95% CI [LL, UL]	Fit	Difference
(Intercept)	56.11**	[47.57, 64.65]				
SexMale	3.04	[-0.86, 6.95]	.02	[-.03, .08]		
Age	-0.18	[-0.38, 0.02]	.03	[-.03, .10]		
					R^2 = .056 95% CI[.00,.15]	
(Intercept)	29.21**	[18.48, 39.94]				
SexMale	4.30*	[1.03, 7.58]	.05	[-.02, .11]		
Age	-0.14	[-0.31, 0.03]	.02	[-.02, .06]		
Depression	1.68**	[1.18, 2.18]	.30	[.15, .44]		
					R^2 = .353** 95% CI[.19,.46]	ΔR^2 = .297** 95% CI[.15, .44]

Note. A significant *b*-weight indicates the semi-partial correlation is also significant. *b* represents unstandardized regression weights. *sr*^{2} represents the semi-partial correlation squared. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.
* indicates p < .05. ** indicates p < .01.

Figure 18: Results table for hierarchical regression.

20.3.2 Forced Entry

In forced entry, all predictors are entered at once, assuming theoretical justification for their inclusion. Like all multiple regressions, each coefficient is interpreted in the context of all other predictors. This method is straightforward and appropriate when all predictors are equally important. However, it does not allow testing of incremental validity.

20.3.3 Stepwise Regression

Stepwise methods rely on statistical criteria rather than theory. Predictors are added or removed based on *a priori* metrics such as hitting a specific p-values or AIC threshold. There are specific variants of stepise methods.

First is forward Selection. In this method, the regression model starts with no predictors and first adds the most significant one (i.e., lowest p-value). It then continues to test all predictors and add the most significant one. This continues until improvement stops (i.e., no more predictors reach statistical significance). A second method is backward Elimination. In this methods, the regression model starts with all all predictors and removes the least significant one iteratively. A last method is bidirectional selection, which combines forward and backward methods of stepwise regression.

Stepwise regression is generally discouraged in psychological research. Specifically, because it iteratively tests numerous models to find the 'most significant' one, it often inflates type 1 error rates. Furthermore, it generally discounts predetermine theory that should specify which variables are included. Remember, variables in a regression model are always interpreted within the context of the full model. Thus, if theory specifies that x should be in a model, the other predictors results will change with x in the model versus without it (unless x is orthogonal to the other predictors and the outcome variable).

20.3.4 All-Subsets Regression

This method evaluates every possible combination of predictors to identify the best-fitting model. While comprehensive, it is rarely used in practice. Again, using you sample data to iterative identify the 'best' model increase type 1 error rates and likely goes against any grounded theory.

For example, if a research has eight predictors, there will be 255 regression models: 8 models with one predictors, 28 with two predictors, 56 with three predictors, 70 with four predictors, 56 with five predictors, 8 with seven predictors, and 1 with eight predictors. You statistical software would run all possible models and return the one with the best criteria (e.g., R^2_{adj}).

20.4 Conclusion

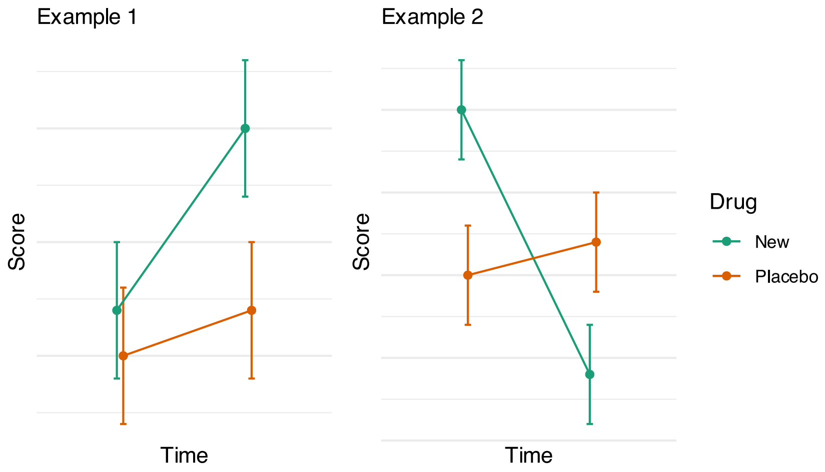
Regression is a powerful statistical analysis that is capable of answering numerous research question and testing hypotheses. However, with great power comes great responsibility. Knowing the assumptions and approaches to regression will help ensure your analyses are valid.

21 Moderation

Psychological phenomenon are complex and multifaceted. That is, there are many causes and influences on any given psychological process. Even seemingly simple cognitions, emotions, and behaviors can be difficult to understand and predict. Thus, developing theories that recognize the complexity and building subsequent hypotheses to test these theories is imperative. Moderation is one way to test complex relationships between multiple variables by uncovering nuanced relationships between variables. As Lerner et al. (2015) aptly states:

> …nature never affects behaviour directly; it always acts in the context of internal and external environments. Environment never directly influences behaviour either; it will show variation in its effects depending on the heredity-related characteristics of the organism on which it acts.
>
> — Lerner et al., 2015

You may not recognize it, but you have much experience with moderation. Quite simply, a moderation is an interaction. For example, we may test the efficacy of a new drug versus a placebo on 'happiness'. We measure scores before taking the drug/placebo, and measure scores after some time of taking the drug/placebo. In this example we have two independent variables (time and drug) and one dependent variable (happiness). *An interaction would indicate that the association between an IV and a DV is dependent on some other IV*. That is, the change in time on happiness depends on whether you got the drug or the placebo. These figure represent potential interactions. Note that the lines are not parallel.



In previous chapters on factorial and mixed ANOVAs you encountered interactions between qualitative or categorical variables with mutually exclusive levels. However, regressions are quite flexible in the types of variables they can include. We can test interactions between different types of variables such as categorical X categorical (what we did in ANOVA), continuous X categorical, and continuous X continuous. Furthermore, we can test (not unlike ANOVAs), 3-, 4-, or more-way interactions.

Knowledge Check

Explain to a classmate what moderation is. Draw a figure to help you explain.

21.1 Some Assumptions

Moderation has the same assumptions as multiple regression. There is one, however that deserves special attention: multicollinearity. As you will learn, moderators/variables in interactions in regression are at risk of being correlated (i.e., multicollinearity). Thus, measures must be put in place to counter this. The most implemented method is to mean-center all continuous predictors. Let's have a brief tangent prior to continuing.

Getting the interaction variable

The interaction variable is not a different variable you measure, per se. We simply multiply scores on the variables that compose the interaction.

Consider the following four variables: one outcome y (continuous) and three predictors x_1 (continuous), x_2 (categorical and dummy coded), and x_3 (continuous). I will show the first five participants (out of 100):

ID	y	x1	x2	x3
1	16	50	1	79
2	18	69	0	102
3	17	69	0	84
4	16	50	0	110
5	21	47	1	99

We want to test the interaction/moderation between some of the variables. Let's multiply the variables of interest (let's say we want an interaction between x_1 and x_3. We can multiple them together:

ID	x1	x3	x1_x3
1	50	79	3950
2	69	102	7038
3	69	84	5796
4	50	110	5500
5	47	99	4653

If we get the correlation between these three predictors, $x1$, $x3$, and $x1_x3$ (the interaction), we notice that they are quite correlated. Here is the correlation matrix:

```
         x1    x3 x1_x3
x1    1.000 0.004 0.750
x3    0.004 1.000 0.654
x1_x3 0.750 0.654 1.000
```

The correlation between the predictors and the interactions terms are high; x_1 and the interaction is $r = .750$ and x_2 and the interaction is $r = .654$. There is a way to bypass this multicollinearity, which violates an assumption of regression models: mean-centering.

Let's go create two new interaction variables using the data in the table above. First, let's model the interaction between x_1 (continuous) and x_2 (categorical). To mean center $x1$, we find the mean and subtract it from every score. The mean of $x1$ is 51.58. We can create a new variable that is mean-centered (x_{1c} may be a good naming convention, where c stands for centered). Each persons score would be on this new variable would be:

Person 1: $50 - 51.58 = -1.58$ Person 2: $69 - 51.58 = 17.42$ … Person 5: $47 - 51.58 = -4.58$

This new variable is what is used in our new regression analyses (as you will learn, in addition to the initial variables). We multiply that with x_2. Because x_2 is categorical, not continuous, we do not mean center that variable. The interaction variable between x_1 and x_2 can be found in the following table:

ID	x1	x2	x1c	x1c_x2
1	50	1	−1.58	−1.58
2	69	0	17.42	0
3	69	0	17.42	0
4	50	0	−1.58	0
5	47	1	−4.58	−4.58

To model an interaction between x_1 and x_3, two continuous variables, we must mean center *both*. The following show the mean centered scores and their interaction:

ID	x1	x3	x1c	x3c	x1c_x3c
1	50	79	−1.58	−19.61	30.984
2	69	102	17.42	3.39	59.054
3	69	84	17.42	−14.61	−254.506

ID	x1	x3	x1c	x3c	x1c_x3c
4	50	110	−1.58	11.39	−17.996
5	47	99	−4.58	0.39	−1.786

Recall above the high correlations between the uncentered predictors and the interaction. Let's calculate the correlation between the mean-centered variables and that interaction:

```
             x1c   x3c x1c_x3c
x1c        1.000 0.004  -0.068
x3c        0.004 1.000   0.093
x1c_x3c   -0.068 0.093   1.000
```

The correlation are extremely small. The correlation between the mean-centered predictors and the interactions terms are low; x_{1c} and the interaction is $r = -.068$ and x_{3c} and the interaction is $r = .093$. Multi-collinearity has been avoided!

Now you know how interaction variables are created. Many statistical software package will automatically create this for you. However, here are some major points to remember:

1. You only mean-center predictors; the outcome is not centered
2. You only mean-center predictors that are continuous
3. Categorical predictors, while not mean-centered, must be dummy coded
4. You must include the both the mean-centered predictor and the interaction in your analysis, even if you only care about the interaction. For example, you must include x_{1c} and x_{2c} if you will be investigating their interaction, x_{intx}

21.2 RUNNNNN! Get to the protein!

We are working with a new health company and been tasked with testing the association between average daily protein intake (DPI; measured in grams) and lean muscle mass (LMM; the amount of muscle tissue measured in pounds) in a group of individuals. The company is also interested in the association between LMM and gym activity (gym goers versus gym abstainers). They believe that protein will be most helpful in creating LMM for those who go the gym.

21.3 Continuous X Categorial Interactions

21.3.1 1. Generating hypotheses

Our conceptual hypotheses can be phrased as:

H1: Individuals who have a higher daily protein intake (DPI) will have more lean muscle mass (LMM) H2: Individuals who are gym goers will have more LMM compared to those who do not H3: Gym activity will moderate the relationship between DPI and LMM. Specifically, the relationship between gym activity and LMM will be stronger for those who have higher DPI

We will use two-sided tests for each regression coefficient. Importantly, interactions are simply regression coefficients. Thus, we can model our hypotheses like we did in previous regression analysis using either our coefficients (sr^2):

$$H_0 : \text{all}\, \beta = 0$$

and

$$H_A : \text{all}\, \beta \neq 0$$

OR our entire model (R^2):

$$H_0 : R^2 = 0$$

$$H_A : R^2 > 0$$

For our protein research, we have three regression coefficients (four including the intercept):

- β_{dpi}: protein intake

- β_{gym}: gym activity
- β_{inx}: interaction between protein intake and gym activity

Our resulting model will be:

$$y_i = b_o + b_{dpi}(x_{i1}) + b_{gym}(x_{i2}) + b_{inx}(x_{3i}) + e_i$$

Where:

- y_i is person i's lean muscle mass
- x_{i1} is person i's daily protein intake
- x_{2i} is person i's gym activity
- x_{3i} is person i's score on the interaction variable (remember, the product x_1 and x_2 which are mean centered, if continuous)

21.3.2 2. Designing a study

Participants: Participants were recruited through local advertisements and social media platforms. Eligible participants were adults (18+) residing in the Corner Brook region. A total of 100 participants were included: 50 regular gym-goers (attending the gym at least three times per week for the past six months) and 50 non-gym-goers (no structured exercise routine in the past six months). Recruitment materials and procedures were approved by the Grenfell Campus Ethics Review Board.

Measures: *Daily Protein Intake (DPI)*: Participants reported their average daily protein consumption (grams per day) using a validated dietary recall questionnaire.

Lean Muscle Mass (LMM): Lean muscle mass was assessed using bioelectrical impedance analysis (BIA), providing an estimate in kilograms.

Gym Activity: A binary variable indicated gym status (1 = gym-goer, 0 = non-gym-goer).

Procedure: Participants provided informed consent and completed the dietary recall questionnaire. Lean muscle mass was measured during an in-person session using standardized BIA procedures. Demographic information (age, sex) was also collected for descriptive purposes.

Design and Analysis: The study employed a cross-sectional design. A multiple regression analysis was conducted to examine whether Daily Protein Intake (continuous) and Gym Activity (binary) predicted Lean Muscle Mass. The regression model included DPI and gym activity as predictors, with LMM as the outcome variable. Interaction effects between DPI and gym activity were explored to determine whether the relationship between protein intake and lean muscle mass differed by gym status.

Assumptions of linear regression (normality, homoscedasticity, multicollinearity) were checked prior to analysis. Effect sizes and 95% confidence intervals were reported. Statistical significance was set at $\alpha = .05$.

All recruitment materials and procedures were approved by the Grenfell Campus Ethics Review Board.

21.3.3 3. Collecting data

You follow through with your research plan and get the following data (only the first 10 participants are shown to show the structure of the data):

ID	Gym	DPI	LMM
67	No Gym	22	80
29	Gym	23	93
42	Gym	28	74
74	No Gym	30	65
96	No Gym	33	59

21.3.4 4. Analyzing data

Data is analyzed analogously to multiple regression. We will analyze are variables in a single block. However, one may want to analyze main effects in one block followed by the interactions in a second block.

Prior to our analysis, we must center our continuous independent variable: DPI. We would model an interaction by creating a new variable (x_3=interaction) that is the product of the other variables ($x_{1c} = Centered_DPI$, $x_2 = gym$) - $x_3 = (x_{1c})(x_2)$

Let's run our model with centered interaction terms:

Observations	100
Dependent variable	LMM
Type	OLS linear regression

F(3,96)	166.84
R²	0.84
Adj. R²	0.83

	Est.	2.5%	97.5%	t val.	p	VIF
(Intercept)	64.27	59.32	69.21	25.80	0.00	NA
DPI_Centered	−0.06	−0.31	0.20	−0.44	0.66	2.25
GymGym	70.27	63.28	77.27	19.95	0.00	1.00
DPI_Centered:GymGym	1.31	0.97	1.66	7.57	0.00	2.24
Standard errors: OLS						

Let's work out the regression equations using this model:

No Gym ($x_2 = 0$)

$$y_{lmm} = 64.27 - 0.06(x_{i1c}) + 70.27(x_{i2}) + 1.31(x_{i1c})(x_{i2}) + e_i$$

$$y_{lmm} = 64.27 - 0.06(x_{i1c}) + 70.27(0) + 1.33(x_{i1c})(0) + e_i$$

$$y_{lmm} = 64.27 - 0.06(x_{i1c}) + e_i$$

Gym ($x_2 = 1$)

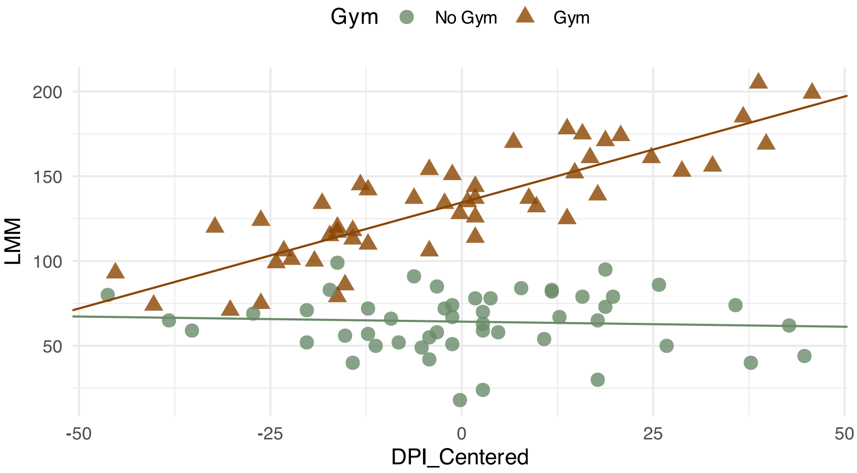
$$y_{lmm} = 64.27 - 0.06(x_{i1c}) + 70.27(x_{i2}) + 1.31(x_{i1c})(x_{i2}) + e_i$$

$$y_{lmm} = 64.27 - 0.06(x_{i1c}) + 70.27(1) + 1.31(x_{i1c})(1) + e_i$$

$$y_{lmm} = (64.27 + 70.27) - (0.06(x_{i1c}) + 1.31(x_{i1c})) + e_i$$

$$y_{lmm} = 134.54 + 1.25(x_{i1c}) + e_i$$

And a new visualization of data:



Notice how the new lines, with each group having their own intercept and slope seem to fit the data quite well. An interaction allows this.

21.4 Continuous X Continuous Interaction

We often will deal with multiple continuous variables that may interact. Fortunately, the process is similar. However, we will now need to center all predictors in the interaction.

Let's stick to a similar example. Assume we want to determine if LMM regresses on DPI. But, we think that average hours in the gym per week will interact with protein intake to predict lean muscle mass. So, our previous categorical predictor of being a gym goer versus not is now a *continuous predictor* of average hours in the gym per week. So, we reach out to people and ask to now consider the average gym hours per week.

Our specific hypotheses are as follow:

21.4.1 1. Generating hypotheses

1. DPI will predict LMM

- $H1: \beta_1 \neq 0$

2. Average gym hours will predict LMM

- $H2: \beta_2 \neq 0$

3. There will be an interaction between DPI and gym hours on LMM. Specifically, the relationship between DPI and LMM will be stronger for those who average more gym hours per week.

- $H3: \beta_3 \neq 0$

21.4.2 2. Designing a study

We conduct a study almost identical to the previous but collect average gym hours per week versus someone being a gym goer or not.

21.4.3 3. Collecting data

We collect data according to the plan and get the following (only first 10 participants shown to demonstrate the structure of the data):

ID	DPI	Gym	LMM
31	27	5	198
38	29	3	165
24	36	4	183
1	39	5	192
39	39	4	157

21.4.4 4. Analyzing data

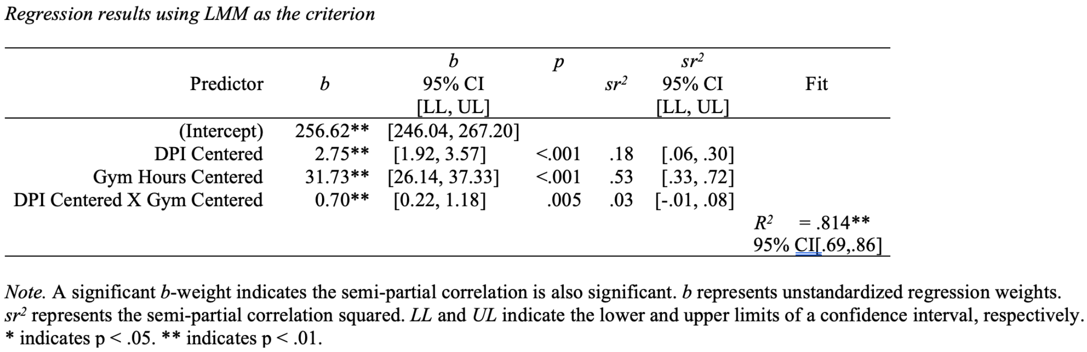
We mean-centered both DPI and average gym time to create the interaction term. Our analysis results in the following:

Observations	50
Dependent variable	LMM
Type	OLS linear regression

F(3,46)	67.304
R²	0.814
Adj. R²	0.802

	Est.	2.5%	97.5%	t val.	p	VIF
(Intercept)	256.624	246.044	267.204	48.824	0.000	NA
DPI_C	2.747	1.921	3.573	6.694	0.000	1.011
Gym_C	31.734	26.138	37.330	11.414	0.000	1.037
DPI_C:Gym_C	0.700	0.218	1.181	2.927	0.005	1.047
Standard errors: OLS						

Or, as another output from the `apaTables()` package in r.:



Regression results using LMM as the criterion

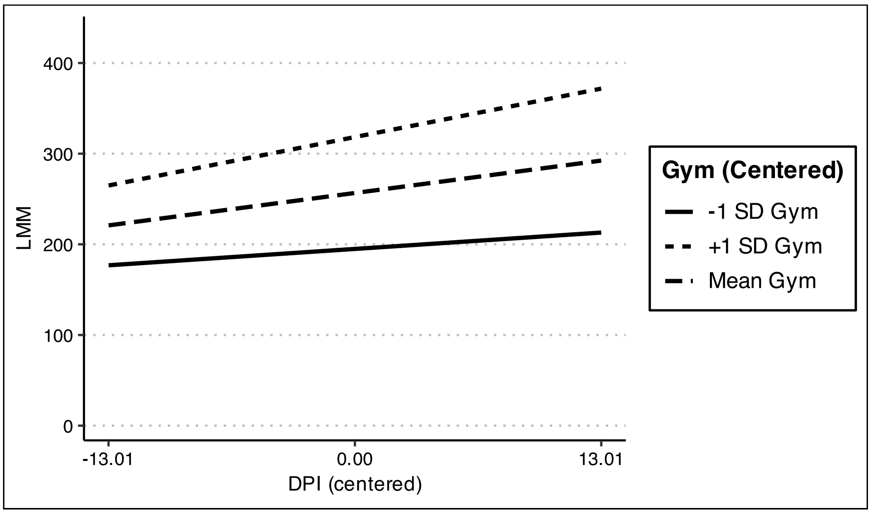
Predictor	*b*	*b* 95% CI [LL, UL]	*p*	*sr*²	*sr*² 95% CI [LL, UL]	Fit
(Intercept)	256.62**	[246.04, 267.20]				
DPI Centered	2.75**	[1.92, 3.57]	<.001	.18	[.06, .30]	
Gym Hours Centered	31.73**	[26.14, 37.33]	<.001	.53	[.33, .72]	
DPI Centered X Gym Centered	0.70**	[0.22, 1.18]	.005	.03	[-.01, .08]	
						*R*² = .814** 95% CI[.69,.86]

Note. A significant *b*-weight indicates the semi-partial correlation is also significant. *b* represents unstandardized regression weights. *sr*² represents the semi-partial correlation squared. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively. * indicates p < .05. ** indicates p < .01.

Figure 19: apaTables() output.

Remember, the intercept here is for when all other variables are 0. We have centered our variables, so 0 carries a specific meaning; a score of 0 on a mean-centered variable is equal to the mean. So, the intercept in these results reflect the expected LMM score for an individual with an average DPI and Gym hours. We could expect someone who consumes the mean amount of protein and who goes to the gym an average amount of time to have 256.62lbs of lean muscle mass.

Let's visualize the new interaction.



Here we can see that the relationships between DPI and LMM is less strong for those who do not go to the gym often (i.e., −1SD). However, as gym time increase, the strength of the relationship increases. In other words, the slope of the lines increases when we move from −1SD, to 0SD, to +1SD. A formal simple slopes analysis can tell us the exact slope for these lines.

21.5 Simple Slopes

Simple slopes analysis tests whether the slope (coefficient) of one predictor (i.e., one IV) differs from 0 at given levels or values of the moderator (i.e., another IV). Although we can you any level of value on the moderator, the typically convention if to test at −1SD, 0, and +1SD on the moderator. Thus, if a variable has a mean of 20 and SD of 10, then our simple slopes analysis will test if the slope for the IV and DV is statistically significant for the values 10, 20, and 30 on the moderator.

These analyses can be used to provide additional information about a potential interaction. Let's use the data from our last example and run a simple slopes analysis.

```
SIMPLE SLOPES ANALYSIS
```

```
Slope of DPI when Gym = 2.71578 (- 1 SD):

  Est.   S.E.   t val.      p
------ ------ -------- ------
  1.39   0.59     2.36   0.02

Slope of DPI when Gym = 4.66000 (Mean):

  Est.   S.E.   t val.      p
------ ------ -------- ------
  2.75   0.41     6.69   0.00

Slope of DPI when Gym = 6.60422 (+ 1 SD):

  Est.   S.E.   t val.      p
------ ------ -------- ------
  4.11   0.65     6.31   0.00
```

As is seen from the output, we are given a separate analysis for each value of the simple slopes analysis. In this specific example, the relationship between DPI and LMM was statistically significant at all tested levels (-1SD, mean, and +1SD) of gym hours.

21.5.1 5. Write your results/conclusions

Let's write up the results of this last model that used two continuous predictors and the interaction.

We regressed individual's lean muscle mass (LMM) onto their daily protein intake (DPI). Additionally, we believed that average gym hours per week would moderate this relationship (i.e., an interaction). The results suggest that DPI was a statistically significant predictor and accounted for 18% of the variance in LMM, $b = 2.75, p < .001, sr^2 = .18, 95\% \, CI[.06, .30]$. Gym hours was a statistically significant predictor of and accounted for an addition 53% of the variance in LMM, $b = 31.73, p < .001, sr^2 = .53, 95\% \, CI[.33, .72]$. Finally, the interaction between DPI and Gym hours was statistically significant and accounted for an addition 3% of the variance in LMM, $b = 0.70, p = .01, sr^2 = .03, 95\% \, CI[-.01, .08]$.

21.6 Conclusion

Moderation is a powerful tool to model the relationships between variables and how they may depend on a different variable–a moderator. Up next is mediation, where we will formally draw on using separate blocks in regression analysis.

22 Mediation

Mediation is a commonly implemented analysis. You already have a wonderful background that can easily be applied to mediation analysis by being familiar with multiple regression. Quite simply, we are seeking to understand whether the association between two variables is **mediated** by a third variable. That is, mediation can help us understand whether the association between x_1 and y is only indirectly through x_2. We can visualize this as:

$$x_1 \rightarrow x_2 \rightarrow y$$

22.1 A Quick Caveat

Much mediation research seeks to answer research questions about **causality**–and sometimes frames them that way. That is, one thing (x_1) *causes* changes in another thing (x_2), which causes changes in a third thing (y). However, running a mediation analysis is no different than running a multiple regression. Causality requires the appropriate research design; mediation cannot achieve this on its own. There is an abundance of literature discussing the complexities of mediation and causality (Pearl, 2014, Zhao et al. (2010)).

Despite this, mediation is useful for explain mechanisms of association. We will use the famous (maybe infamous) steps by Baron & Kenny (1986), which are often used (and often misinterpreted).

22.2 Mediation Steps

We will focus on basic mediation where there is:

- A predictor (x)
- A mediator (m)
- An outcome/criterion (y)

We believe that x is related to y THROUGH m

$$x \rightarrow m \rightarrow y$$

There are three major steps to our mediation approach.

22.2.1 Step 1: Is the predictor associated with the criterion?

We first run a regression model with y regressed on x

$$y_i = b_o + b_1 x_i + e_i$$

Observations	100
Dependent variable	y
Type	OLS linear regression

F(1,98)	4.67
R²	0.05
Adj. R²	0.04

	Est.	2.5%	97.5%	t val.	p
(Intercept)	44.75	28.38	61.12	5.42	0.00
x	0.17	0.01	0.33	2.16	0.03
Standard errors: OLS					

This tells use that x is a statistically significant predictor of y. Indeed, 5% of variation in y is accounted for by x. This is analogous to simple regression.

22.2.2 Step 2: Is the predictor associated with the mediator?

Second, we run a new regression model with m regressed on x. Notice how the coefficient is b_2 because b_1 was used in step 1. These are different coefficients and based on the results of the analyses. The results of this step are:

$$m_i = b_o + b_2(x_i) + e_i$$

Observations	100
Dependent variable	m
Type	OLS linear regression

F(1,98)	22.04
R²	0.18
Adj. R²	0.18

	Est.	2.5%	97.5%	t val.	p
(Intercept)	31.44	14.83	48.05	3.76	0.00
x	0.38	0.22	0.55	4.70	0.00
Standard errors: OLS					

This tells use that x is a statistically significant predictor of m. Indeed, 18% of the variance in m is accounted for by x. This is also analogous to simple regression.

22.2.3 Step 3: Is the predictor still associated with the criterion after the mediator is included?

Our last step is *slightly* more complicated. We will run a regression model with:

$$y_i = b_o + b_3(x_i) + b_4(m_i) + e_i$$

Please note the coefficients above are different from steps 1 and 2 because they will be different when all three variables are in the model. The results of this model are:

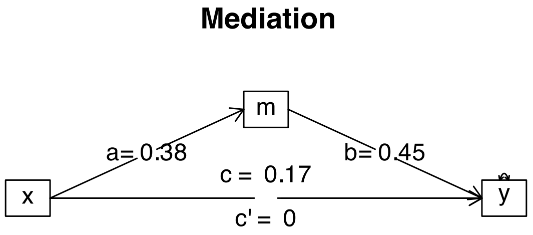
Observations	100
Dependent variable	y
Type	OLS linear regression

F(2,97)	15.83
R²	0.25
Adj. R²	0.23

	Est.	2.5%	97.5%	t val.	p
(Intercept)	30.55	14.90	46.19	3.87	0.00
x	0.00	−0.16	0.16	0.01	0.99
m	0.45	0.28	0.63	5.08	0.00
Standard errors: OLS					

From these results we can understand that although x was a statistically significant predictor or y in step 1, it is no longer statistically significant when m is included in the model. And since x predicts m, and m predicts y, it would seem that any association between x and y is *through* m. Please note that I have not used causal language.

A basic mediation (as we have done) figure typically includes a few different subscripts. It will typically look like:



In this figure:

- c is the coefficient from Step 1
- c' ("c *prime*") is the coefficient from Step 3
  - If it becomes 0, m **fully mediates** the relationship between x and y
  - If it gets smaller (vague, I know), m **partially mediates** the relationship between x and y

22.3 Writing up mediation results

We conducted a mediation analysis to determine the extent to which the association between x and y is mediated by m (see Figure 1). Means, SD, and correlations between variables are presented in Table 1.

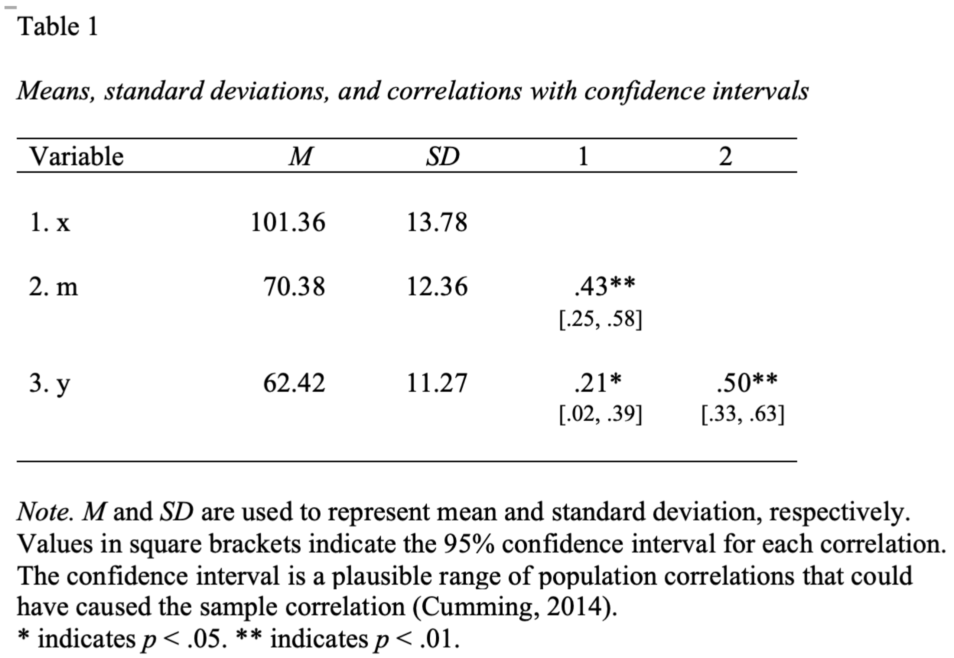


Table 1

Means, standard deviations, and correlations with confidence intervals

Variable	*M*	*SD*	1	2
1. x	101.36	13.78		
2. m	70.38	12.36	.43** [.25, .58]	
3. y	62.42	11.27	.21* [.02, .39]	.50** [.33, .63]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).
* indicates *p* < .05. ** indicates *p* < .01.

Figure 20: apaTables() output.

We used Baron and Kenny's recommended steps to determine a potential mediation. In step 1, we used x as a predictor of y. Here, x was a statistically significant predictor of y, $b = .17, p = .003, sr^2 = .045, 95\% CI[.00, .015]$. Thus, x accounted for 4.5% of the unique variance in y.

Second, we ran a separate regression model to determine if x was a suitable predictor of m. The results suggest that x is a statistically significant predictor of m, $b = .38, p < .001, sr^2 = .18, 95\% CI[.06, .31]$.

Third, and subsequently, we ran a second block on our first regression model, adding m as a predictor of y. The results indicate that m was a statistically significant predictor of y, $b = .45, p < .001, sr^2 = .20, 95\% CI[.06, .34]$. Importantly, x was no longer a significant predictor of y after m was included in the model, $b < .00, p = .99, sr^2 = .00, 95\% CI[-.00, .00]$. The full model of x and m predicting y accounted for 24.6% of the variance in y, $R^2 = .246$. Thus, m appears to fully mediate the association between x and y (see Figure 1 and Table 2).

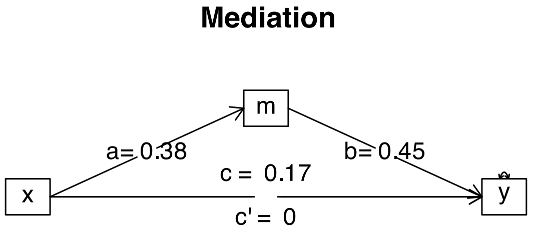


Table 2

Regression results using y as the criterion

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	*sr*²	*sr*² 95% CI [LL, UL]	*r*	Fit	Difference
(Intercept)	44.75**	[28.38, 61.12]							
x	0.17*	[0.01, 0.33]	0.21	[0.02, 0.41]	.05	[.00, .15]	.21*		
								*R*² = .045* 95% CI[.00,.15]	
(Intercept)	30.55**	[14.90, 46.19]							
x	0.00	[-0.16, 0.16]	0.00	[-0.19, 0.19]	.00	[-.00, .00]	.21*		
m	0.45**	[0.28, 0.63]	0.50	[0.30, 0.69]	.20	[.06, .34]	.50**		
								*R*² = .246** 95% CI[.10,.37]	Δ*R*² = .201** 95% CI [.06, .34]

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. *sr*² represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.
* indicates *p* < .05. ** indicates *p* < .01.

Figure 21: apaTables() output.

22.4 Conclusion

Mediation is an important tool to add to your statistical toolbox. It is commonly used and allows us to understand whether the association between two variables is **mediated** by a third variable. Using direct and indirect effects can help us test more complex psychological theories.

23 Chi-square

This chapter will cover the chi-square test, a statistical method used to examine the relationship between categorical variables. Unlike regression (and ANOVA), which focuses on continuous dependent variables, the chi-square test assesses whether there is an association between two categorical variables. But why do researchers need to examine associations between categorical variables?

Understanding relationships between categorical variables is essential in many fields of research. Real-world behaviors, traits, and classifications are often categorical—such as gender, education level, voting preferences, or disease status. The chi-square test allows researchers to determine whether observed frequencies in different categories differ significantly from what would be expected under a null distribution (i.e., no association). By doing so, we can identify patterns and relationships that might not be immediately apparent.

In short, it's another tool to add to your statistical toolbox.

Think about it

Note that the chi-square test can be applying to more than two categorical variables. However, in this chapter we will primarily involve examples with two variables.

23.1 Some Additional Details

The chi-square test is particularly useful when researchers want to examine whether two categorical variables are independent or related. For example, a researcher might investigate whether gender is associated with voting preference or whether treatment group membership affects recovery rates.

The general form of the chi-square test statistic is:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where:

- O_{ij} represents the observed frequency for cell ij (actual counts in each category),
- E_{ij} represents the expected frequency for cell ij (counts that would occur under the assumption of independence),
- χ^2 is the chi-square test statistic, which follows a chi-square distribution.

23.2 Key Assumptions

Like all of our analyses thus far, a chi-square test is valid under the certain assumptions. Some of which we have already explored:

1. Independence of Observations
Each observation should belong to only one category, and observations should not be related to one another.

2. Expected Frequency Rule
Expected counts in each category should generally be 5 or more for the chi-square approximation to be valid. When expected counts are low, alternative methods (e.g., Fisher's Exact Test) may be needed.

3. Large Sample Size
The chi-square test performs best with a sufficiently large sample, as small sample sizes may produce unreliable results. A typical rule of thumb is to avoid expected cell counts less than 5 (more to come).

4. Categorical Data
Both variables should be measured at the categorical level (e.g., nominal or ordinal scales) rather than continuous.

23.3 Contingency Tables and Expected Frequencies

Before conducting a chi-square test, it is important to organize the data into a contingency table. A contingency table, also known as a cross-tabulation or crosstab, displays the frequencies of observations of the two categorical variables. This table allows researchers to compare observed frequencies with expected frequencies under the assumption of independence.

A simple contingency table for two categorical variables (e.g., Gender and Voting Preference) might look like this:

	Candidate A ($j = 1$)	Candidate B ($j = 2$)	
Male ($i = 1$)	40	60	*Row total: 100*
Female ($i = 2$)	50	50	*Row total: 100*
	Column total: 90	*Column total: 110*	*Total sample size: 200*

While a contingency table may only display the actual frequencies in each cell (block), it is helpful to also write the row, column, and grand total, like the above table. It is also helpful to think of each row (i) as and column (j) as having a number. Combining values of row and columns, we can determine a cell of interest. For example, $n_{i=1,j=1}$, refers to the frequency in row 1, column 1; this is the cell of the table representing males who voted for candidate A: $n = 40$.

Think about it

What is the cell frequency for the cell in the second row, first column: $n_{2,1}$?

Continuing, to determine whether the variables are independent, we need to calculate the expected frequency for *each cell* using the formula:

$$E_{i,j} = \frac{(\text{Row Total}_i) \times (\text{Column Total}_j)}{\text{Grand Total}}$$

For example, the expected frequency for Male/Candidate A would be:

$$E_{1,1} = \frac{100 \times 90}{200} = 45$$

The expected frequencies allow us to use row and columns totals to determine what data would look like if there were no association. However, you could, in theory, choose any values for the expected frequencies that align with your theory. Regardless, we need to calculate the expected frequency for each cell in our contingency table. Doing so, we would get the following. This first table represents the observed frequencies:

	Candidate A ($j = 1$)	Candidate B ($j = 2$)
Male ($i = 1$)	40	60
Female ($i = 2$)	50	50

This second table represents the expected frequencies:

	Candidate A ($j = 1$)	Candidate B ($j = 2$)
Male ($i = 1$)	45	55
Female ($i = 2$)	45	55

You may find it easy to view discrepancies in observed versus expected frequencies–and to do any potential calculations– by combining both tables into one. Here, expected frequencies are in parentheses following the observed frequencies:

	Candidate A ($j = 1$)	Candidate B ($j = 2$)
Male ($i = 1$)	40 (45)	60 (55)
Female ($i = 2$)	50 (45)	50 (55)

Comparing these expected frequencies with the observed counts allows us to determine whether any differences are statistically significant.

The next step is to compute the chi-square test statistic and assess its significance using the chi-square distribution.

23.4 Chi-square Statistics

After obtaining the observed and expected frequencies, we compute the chi-square test statistic using the formula:

$$\chi^2 = \sum \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

For our example, the chi-square test statistic is calculated as follows:

$$\chi^2 = \frac{(40-45)^2}{45} + \frac{(60-55)^2}{55} + \frac{(50-45)^2}{45} + \frac{(50-55)^2}{55}$$

Computing each term:

$$\chi^2 = \frac{(-5)^2}{45} + \frac{(5)^2}{55} + \frac{(5)^2}{45} + \frac{(-5)^2}{55}$$

$$\chi^2 = \frac{25}{45} + \frac{25}{55} + \frac{25}{45} + \frac{25}{55}$$

$$\chi^2 \approx 0.56 + 0.45 + 0.56 + 0.45 = 2.02$$

Once we calculate the chi-square test statistic, we compare it to the critical value from the chi-square distribution table, or we compute a p-value. The degrees of freedom (df) for a chi-square test are calculated as:

$$\text{df} = (\text{Number of Rows} - 1) \times (\text{Number of Columns} - 1)$$

For our example:

$$\text{df} = (2 - 1) \times (2 - 1) = 1$$

Using a chi-square table or statistical software, we determine the critical value for our chosen significance level (e.g., $\alpha = .05$). If our calculated chi-square statistic exceeds the critical value, we reject the null hypothesis, suggesting that the association between the variables in unlikely given a true null hypothesis.

You can find critical chi-square tables online. Additionally, there are websites that can calculate an exact p-value for a given χ^2 and df–such as here. However, most statistical software packages will provide exact p-values, residuals, and effect sizes.

23.5 Effect Size

It's important to assess the strength of the association between the variables. One common measure of effect size for chi-square tests is **Cramer's V**. Cramer's V provides a standardized measure of association and is calculated as:

$$V = \sqrt{\frac{\chi^2}{n \times (\min(r-1, c-1))}}$$

Where:

- χ^2 is the chi-square statistic,
- n is the total sample size,
- r is the number of rows in the contingency table,
- c is the number of columns in the contingency table.

For example, for our 2×2 table, the effect size can be computed as follows:

$$V = \sqrt{\frac{2.02}{200 \times (1)}} = \sqrt{\frac{2.02}{200}} \approx 0.101$$

There are some bench mark values to help with the interpretation of Cramer's V:

- Small effect: $0.1 \leq V < 0.3$
- Medium effect: $0.3 \leq V < 0.5$
- Large effect: $0.5 \leq V$

In this case, the effect size of $V = .101$ suggests a small association between the variables.

23.6 Post-hoc Analyses: Residuals

Residuals in a chi-square test help us understand the magnitude of discrepancies between observed and expected frequencies. They are calculated as:

$$\text{Residual} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

The residuals give us an indication of how much each observed frequency deviates from its expected frequency in terms of standard deviations. For each cell, a large residual indicates a large difference between observed and expected frequencies, which could be important for identifying patterns in the data.

For our example:

For Male/Candidate A:

$$\frac{40 - 45}{\sqrt{45}} = \frac{-5}{6.708} \approx -0.745$$

For Male/Candidate B:

$$\frac{60 - 55}{\sqrt{55}} = \frac{5}{7.416} \approx 0.674$$

For Female/Candidate A:

$$\frac{50-45}{\sqrt{45}}=\frac{5}{6.708}\approx 0.745$$

For Female/Candidate B:

$$\frac{50-55}{\sqrt{55}}=\frac{-5}{7.416}\approx -0.674$$

These residuals can help us determine which specific categories contribute to the overall chi-square statistic.

Our formal test would result in:

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|            Std Residual |
|-------------------------|

Total Observations in Table:  200

             |
             |     [,1]  |     [,2]  | Row Total |
-------------|-----------|-----------|-----------|
        [1,] |       40  |       60  |     100   |
             |   45.000  |   55.000  |           |
             |    0.556  |    0.455  |           |
             |   -0.745  |    0.674  |           |
-------------|-----------|-----------|-----------|
        [2,] |       50  |       50  |     100   |
             |   45.000  |   55.000  |           |
             |    0.556  |    0.455  |           |
             |    0.745  |   -0.674  |           |
-------------|-----------|-----------|-----------|
Column Total |       90  |      110  |     200   |
-------------|-----------|-----------|-----------|

Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  2.0202     d.f. =  1     p =  0.155218

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  1.63636     d.f. =  1     p =  0.200825


       Minimum expected frequency: 45
```

and for Cramer’s V:

```

    Two-sided 95% chi-squared confidence interval for the
population
    Cramer's V

Sample estimate: 0.100504
Confidence interval:
   2.5%   97.5%
0.00000 0.24933
```

Let’s now explore a full example relevant to the study of psychology.

23.7 Predominantly effective?: Another Example

You want to investigate whether teenagers with different ADHD subtypes will prefer various forms of treatment. You have reason to believe, based on a review of the literature, that individuals may prefer psychosocial treatments as opposed to medication treatments; however, results are mixed (Schatz et al., 2015). You decide to formally investigate the topic.

23.8 1. Generating hypotheses

The main null and alternative hypotheses for this chi-square test can be stated as follows:

- **Null Hypothesis** ($H_0 : O_{ij} = Eij$):
  - The observed frequencies are equal to the expected frequencies
  - Therapy preference is independent of ADHD sub-type.
  - In other words, there is no relationship between ADHD sub-type and therapy preference.
- **Alternative Hypothesis** ($H_A : O_{ij} \neq Eij$):
  - The observed frequencies are not equal to the expected frequencies
  - Therapy preference is dependent on ADHD sub-type.
  - That is, different ADHD sub-types are associated with different therapy preferences.

Any post-hoc analyses will used standardized residuals ≥ 2 to determine particularly influential cells.

23.9 2. Designing a study

You and your team plan a research study. The method follows:

Participants: A power analysis using an effect size of $\phi = .2828$ (derived from the literature) was used to determine the needed sample to achieve a power of $1 - \beta = .8$. The results of the power analysis suggested a required sample size of $n = 300$.

Power Analysis

Power analysis can be completed in R. The `pwr.chisq.test()` function from the `pwr` package is a sound method. It does, however, require Cohen's W and not Cramer's ϕ. This is an easy calculated. Per Cohen (1988), W is:

$$W = \sqrt{\sum_{i=1}^{m} \frac{(P_{Ai} - P_{0i})^2}{P_{0i}}}$$

Where: P_{0i} is the proportion in cell i as indicated by the null hypothesis H_0; P_{Ai} is the proportion in cell i posited by the alternate hypothesis H_A; m = the number of cells.

A major difference in this and the typical analyses we have been doing is that these are *proportions*, not frequencies.

This may seem taxing, particularly because you don't have proportions. Well, we can approximate W using:

$$W \approx \frac{\phi}{\sqrt{k-1}}$$

Where k is the smallest number of rows or columns. So for our power analysis:

$$W \approx \frac{\phi}{\sqrt{k-1}} = .\frac{2828}{\sqrt{2}} = 0.20$$

We can then use R to compute our power analysis:

```

     Chi squared power calculation 

              w = 0.2
              N = 298.382
             df = 4
      sig.level = 0.05
          power = 0.8

 NOTE: N is the number of observations
```

Which suggests a sample of of $n \approx 298.38$, which we would round up to 300.

Participants were recruited from local ADHD support groups and clinical settings. Flyers and online advertisements were used to reach individuals diagnosed with ADHD. Eligible participants were required to have a confirmed ADHD diagnosis of one of the three sub-types: **Predominantly Inattentive (PI), Predominantly Hyperactive-Impulsive (PHI), or Combined Type (CT).** A total of 200 participants were surveyed.

Materials: A structured questionnaire was used to collect self-reported therapy preferences. Participants selected their preferred treatment from three options: **Cognitive Behavioral Therapy (CBT)**, **Behavioral Therapy**, or **Medication**

Procedure: Participants completed an online survey that collected demographic information, ADHD sub-type (based on a clinical diagnosis), and their preferred therapy type. Informed consent was obtained before participation. The ethics review board at Grenfell Campus reviewed and approved the study.

23.10 3. Collecting data

The study was completed as described, and a total of **300** participants provided data. The responses were summarized in the following contingency table:

ADHD Sub-type	CBT	Behavioral Therapy	Medication	**Total**
PI	50	30	20	100
PHI	30	50	70	150
CT	20	40	40	100
Total	100	120	130	300

23.11 4. Analyzing data

A chi-square test of independence was conducted to determine whether there was a significant relationship between ADHD sub-type and therapy preference. The results are as follows:

```
   Cell Contents
|-------------------------|
|                   Count |
| Chi-square contribution |
|            Std Residual |
|-------------------------|

Total Observations in Table:  350

             |
             |      CBT  |        BT  |       Med  | Row Total |
-------------|-----------|-----------|-----------|-----------|
          PI |       50  |       30  |       20  |      100  |
             |   16.071  |    0.536  |    7.912  |           |
             |    4.009  |   -0.732  |   -2.813  |           |
-------------|-----------|-----------|-----------|-----------|
         PHI |       30  |       50  |       70  |      150  |
             |    3.857  |    0.040  |    3.663  |           |
             |   -1.964  |   -0.199  |    1.914  |           |
-------------|-----------|-----------|-----------|-----------|
          CT |       20  |       40  |       40  |      100  |
             |    2.571  |    0.952  |    0.220  |           |
             |   -1.604  |    0.976  |    0.469  |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |      100  |      120  |      130  |      350  |
-------------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  35.8226     d.f. =  4     p =  3.14726e-07
```

```
       Minimum expected frequency: 28.5714
```

```
    Two-sided 95% chi-squared confidence interval for the
population
    Cramer's V

Sample estimate: 0.226219
Confidence interval:
    2.5%    97.5%
0.159234 0.301578
```

Our overall Chi-square was statistically significant, indicating that the observed data are unlikely given our expected data. We can further explore which cells seem to be driving our results by inspecting the standardized residuals. In our results, there are two cells that seem to be particularly influential: individuals with predominantly inattentive type (PI) seem to prefer CBT much more than expected, and prefer medication much less than expected.

23.12 5. Write your results/conclusions

A chi-square test of independence was conducted to examine the relationship between ADHD sub-type (PI, PHI, CT) and therapy type (CBT, Behavioral Therapy, Medication). The results of the chi-square test were statistically significant, $\chi^2(4) = 35.82,\ p < .001,\ V = .226$, $95\% CI\ [.159, .302]$, indicating that the distribution of therapy types differs significantly across ADHD sub-types.

To further explore these results, we examined the standardized residuals for each cell. The standardized residuals indicated that individuals with predominantly inattentive type (PI) were more likely to prefer CBT (standardized residual $= 4.009$) and less likely to prefer medication (standardized residual = -2.813) than expected. The remaining cells showed minor deviations from expected frequencies, with standardized residuals less than 2.

These findings suggest a strong preference for CBT among individuals with PI. Further research may be necessary to explore the underlying factors contributing to these preferences.

23.13 Conclusion

The chi-square test is a powerful tool for analyzing relationships between categorical variables. By comparing observed and expected frequencies, we can determine whether a meaningful association exists. While straightforward to compute, the test has key assumptions that must be met for valid results. Understanding and applying the chi-square test correctly is an essential skill for researchers working with categorical data.

23.14 Chi-square in R

I have found that the best function in R for Chi-square is `CrossTable()` from the `gmodels` package. It is comprehensive.

TO calculate Cramer's V, you can use the `cramersv()` function from the `confintr` package.

24 Other non-parametric tests

Not all data behave the way we want them to. Real-world data are messy. People don't always fit neatly on a bell curves. Measurements might be imprecise or better captured with rankings rather than precise numbers. And often, especially in psychological and social research, we want to explore patterns in **counts**, **ranks**, or **yes/no responses**—things that aren't continuous or normally distributed. Thus far we have explored tests primarily used with continuous and normally distributed data. Non-parametric tests, such as the chi-square, will expand our toolbox and allow us to answer more research questions and test the associated hypotheses.

In many research situations we find that our data do not meet the assumptions required for parametric tests—assumptions like normality, homogeneity of variance, or interval-level measurement. This is where **nonparametric statistics** come in.

Nonparametric methods are a flexible set of statistical tools that make fewer assumptions about the underlying distributions of the data. They're particularly useful when working with **categorical data**, **ordinal data**, or **small sample sizes**, or when the assumptions of normality or homoscedasticity are clearly violated. In short, nonparametric tests allow us to analyze data that might otherwise fall outside the scope of traditional parametric techniques like *t*-tests and ANOVA.

One of the most widely used nonparametric methods is the **chi-square test**, which assesses whether there is a relationship between two categorical variables. We have covered a full chapter on this test. In this chapter, we will cover other, less common non-parametric tests. Specifically, we will cover:

1. The Rank Sum Test (Wilcoxon-Mann–Whitney Test)

Use this when you want to compare two independent groups (e.g., treatment vs. control) on an ordinal or skewed continuous variable. Think of this as the nonparametric sibling of independent samples t-test.

2. Wilcoxon's Matched-Pairs Signed-Ranks Test

Use this when you want to compare two dependent groups (e.g., pre- and post-treatment) on an ordinal or skewed continuous variable. Think of this as the nonparametric sibling of repeated-measures t-test.

3. The Kruskal–Wallis Test

A generalization of the Rank Sum Test for comparing **three or more independent groups**. Think of this as the nonparametric sibling of one-way ANOVA.

4. The Friedman Test

Used for **repeated-measures** or **matched-subjects** designs involving ordinal data. Think of this as the nonparametric sibling of repeated-measures ANOVA.

> Think about it
>
> These tests don't require your data to be normally distributed, but they do assume that your observations can be ranked and that ranks are meaningful.

Rather than comparing means, these tests compare **distributions of ranks** across groups or conditions. The idea is simple: if the distributions are similar in multiple groups, ranks should be evenly spread across groups (e.g., the most depressed person should have equal chance of being in one group over the other). If not, we'll see systematic differences in the ranks, which can inform us that one condition may outperform (or under-perform) the others.

24.1 Nonparametric ≠ inferior

Because we encounter non-parametric tests less often, it's easy to assume they are inferior or 'less' valid or reliable than parametric tests. In fact, these tests are **robust, flexible, and widely used** in real-world research (including psychology!), where ordinal scales and non-normal data are common.

24.2 Rank Sum Test

To begin understanding nonparametric tests, we must first understand the idea of rankings. Data can be ranked in many ways. First, data may initially be ranked. For example, we know the places individuals finished in a political race. There was a 1st, 2nd, and 3rd place. Otherwise, we may manually rank the data. For example, if we know that politician 1 had 3,992 votes, while politician 2 had 3,027 votes, we can assign them ranks of 1 and 2, respectively. Or, imagine we get students final grades; we could also assign ranks here:

Student	FinalGrade	Rank
Erica	77	5
Chelsea	59	3
Nikki	52	2
Steven	51	1
Aaron	63	4

So why assign ranks? Well, recall that sometimes we fail to meet the assumptions of our parametric tests. In these cases–or when the data are ranked by nature–you should implement these tests.

Now that we understand ranks, let's try another example. I want to get six people together to run a race. Racers will fast (i.e., not eat) 12-hours before the race. However, an hour before the race, I will give three of these individuals Gatorade. The the others will get nothing. Consider only one group of individuals (either Gatorade or nothing): *what are the* *possible ranks they can get?* Our three runners in the Gatorade group could get: a) 1st, 2nd, 3rd; b) 1st, 2nd, 4th; c) 1st, 2nd, 5th, etc. In fact, there are 20 possible combinations of ranks for this group of three among six runners. Let's sum their ranks as well (assume Person 1, 2, and 3 are our Gatorade group):

Scenario	Person1	Person2	Person3	Sum
1	1	2	3	6
2	1	2	4	7
3	1	2	5	8
4	1	2	6	9
5	1	3	4	8
6	1	3	5	9
7	1	3	6	10
8	1	4	5	10
9	1	4	6	11
10	1	5	6	12
11	2	3	4	9
12	2	3	5	10
13	2	3	6	11
14	2	4	5	11
15	2	4	6	12
16	2	5	6	13
17	3	4	5	12
18	3	4	6	13
19	3	5	6	14
20	4	5	6	15

You may notice that some of the sums of ranks repeat. For example, obtaining a sum of ranks = 8 occurs two times. It can be helpful to create a cumulative frequency table to highlight this:

Sum	Count	Cumulative_Frequency	Cumulative_Percentage
6	1	1	5

Sum	Count	Cumulative_Frequency	Cumulative_Percentage
7	1	2	10
8	2	4	20
9	3	7	35
10	3	10	50
11	3	13	65
12	3	16	80
13	2	18	90
14	1	19	95
15	1	20	100

In this case, if the Gatorade has no impact, than the sum of ranks should be similar to the group that received nothing. If the Gatorade had an impact, our three Gatorade drinkers should have a relatively lower sum of ranks compared to the non-drinkers, indicating they finished quicker than the non-drinkers. This is the rationale behind the Rank Sum Test.

The Rank Sum Test is like a t-test for ranked data. We can use ranked data or manually create ranks from other data. In the case of ties, we take the mean of the ranks. For example, if two people scored the same and would have been ranks 4 and 5, we would take the mean 4.5.

24.2.1 When to use

There are a a few use cases for the Rank Sum Test.

1. Data violates the assumptions of the independent t-test
2. Super small sample size (Tyler's S^4)
3. Data are ordinal

24.3 Cognitive Race

You are part of a lab that is testing the impacts of test instructions on performance. In this study, participants completed a cognitive task designed to resemble a race. Participants were randomly assigned to either a control group or an experimental group. The control group was informed that the task was a race and that they would be ranked based on their performance. In contrast, the experimental group was explicitly told that the task was not a race and that rankings would not be emphasized. Despite the framing, all participants completed the task individually and in isolation. The study aimed to examine how competitive framing influences cognitive performance and perceived pressure.

The results of the study are as follows:

ID	Group	Rank
1	Experimental	2
2	Experimental	11
3	Experimental	6
4	Experimental	1
5	Experimental	4
6	Experimental	13
7	Experimental	12
8	Experimental	3
9	Experimental	17
10	Experimental	9
11	Control	15
12	Control	7
13	Control	19
14	Control	18
15	Control	14
16	Control	8
17	Control	5
18	Control	20
19	Control	16
20	Control	10

To complete our rank sum test, we need to calculate a few things. First, we will need W_s, which is define as:

$$W_s = \sum rank$$

And, U, which is defined as:

$$U = W_s - \frac{n(n+1)}{2}$$

Here, U represents the sum of ranks that is adjusted fo the number of observations in each group. This means that we can compare the U statistic regardless of sample size. Because we calculate a U for each group, we select the U value that is smaller.

The sum of ranks, W_s for each group is:

Group	Sum of Ranks
Control	132
Experimental	78

And, thus, our U values are:

Control Group

$$U = 132 - \frac{10(10+1)}{2} = 77$$

Experimental Group

$$U = 78 - \frac{10(10+1)}{2} = 23$$

Recall the table above outlining the cumulative frequencies? We can easily find in this table the extremely unlikely values. J.K. Wilcox did just this for various sample sizes, etc. We compare our resulting statistic to these values.

Using tables like these.

As can be seen from this table, when comparing two groups of $n = 10$ individuals:

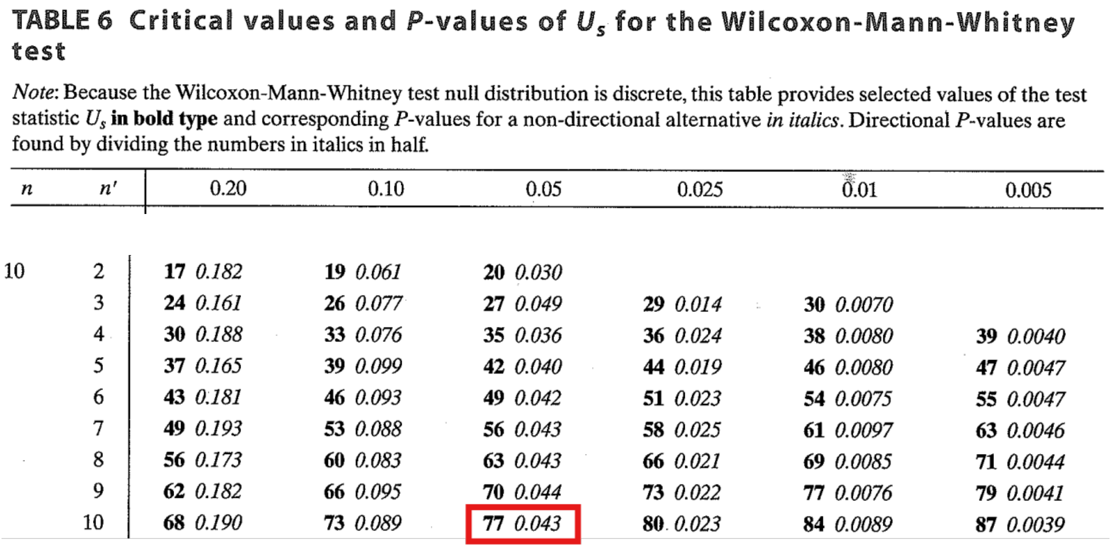


TABLE 6 Critical values and *P*-values of U_s for the Wilcoxon-Mann-Whitney test

Note: Because the Wilcoxon-Mann-Whitney test null distribution is discrete, this table provides selected values of the test statistic U_s **in bold type** and corresponding *P*-values for a non-directional alternative *in italics*. Directional *P*-values are found by dividing the numbers in italics in half.

n	*n′*	0.20	0.10	0.05	0.025	0.01	0.005
10	2	**17** *0.182*	**19** *0.061*	**20** *0.030*			
	3	**24** *0.161*	**26** *0.077*	**27** *0.049*	**29** *0.014*	**30** *0.0070*	
	4	**30** *0.188*	**33** *0.076*	**35** *0.036*	**36** *0.024*	**38** *0.0080*	**39** *0.0040*
	5	**37** *0.165*	**39** *0.099*	**42** *0.040*	**44** *0.019*	**46** *0.0080*	**47** *0.0047*
	6	**43** *0.181*	**46** *0.093*	**49** *0.042*	**51** *0.023*	**54** *0.0075*	**55** *0.0047*
	7	**49** *0.193*	**53** *0.088*	**56** *0.043*	**58** *0.025*	**61** *0.0097*	**63** *0.0046*
	8	**56** *0.173*	**60** *0.083*	**63** *0.043*	**66** *0.021*	**69** *0.0085*	**71** *0.0044*
	9	**62** *0.182*	**66** *0.095*	**70** *0.044*	**73** *0.022*	**77** *0.0076*	**79** *0.0041*
	10	**68** *0.190*	**73** *0.089*	**77** *0.043*	**80** *0.023*	**84** *0.0089*	**87** *0.0039*

The critical value for would be greater than greater than 77, with an exact p-value of $p = .043$ for $U = 77$. A formal statistical test in our software would reveal the same thing.

Rank Sum Test in R

In R, we can complete a Rank Sum Test using the `stats` package. Specifically, the function:

```
wilcox.test(DV ~ IV, data=your_data)
```

Where:

- DV is your DV, which can be ranked or unranked (R will rank for you)
- IV if your grouping variable
- your_data is replaced with your data.frame name

For us, the results would be:

```

    Wilcoxon rank sum exact test

data:  Rank by Group
W = 77, p-value = 0.0433
alternative hypothesis: true location shift is not equal to
0
```

24.3.1 Rank Sum Test - Write-up

Results

To examine the effect of competitive framing on cognitive performance, a Wilcoxon rank-sum test (Mann-Whitney U test) was conducted to compare task performance ranks between participants who received competitive instructions (control group) and those who received non-competitive instructions (experimental group). The results revealed a statistically significant difference in performance ranks between the two groups, $U = 77$, $p = .043$. This suggests that the way the task was framed—as a race versus not a race—had a measurable impact on how participants performed. Specifically, participants in the non-competitive condition tended to perform worse than those in the competitive condition, despite completing the task in isolation. These findings support the hypothesis that competitive framing can influence cognitive task outcomes.

24.4 Wilcoxon's Matched-Pairs Signed-Ranks Test

Sometimes we violate the assumptions of parametric tests like the repeated measures/paired t-test—such as the assumption of normality. In these cases, or when dealing with **ordinal** data, we can turn to nonparametric alternative: the **Wilcoxon's Matched-Pairs Signed-Ranks Test**.

This test is used when each participant is measured twice (e.g., before and after a treatment), and we want to determine if there's a statistically significant change in the direction and magnitude of their scores.

24.4.1 Paired Well-being: Pre- and post-COVID

Suppose we are studying how COVID-19 impacted high school students' mental well-being. We assess the same group of students before the pandemic and again after restrictions were lifted. Mental well-being is measured using the Global Assessment of Functioning (GAF) scale, which was adjusted so that lower numbers reflect better functioning and higher numbers reflect worse functioning. We could only recruit 20 people, which was means our sample size is limiting and we have small statistical power (hypothetical power analysis suggests $n = 112$). In this example:

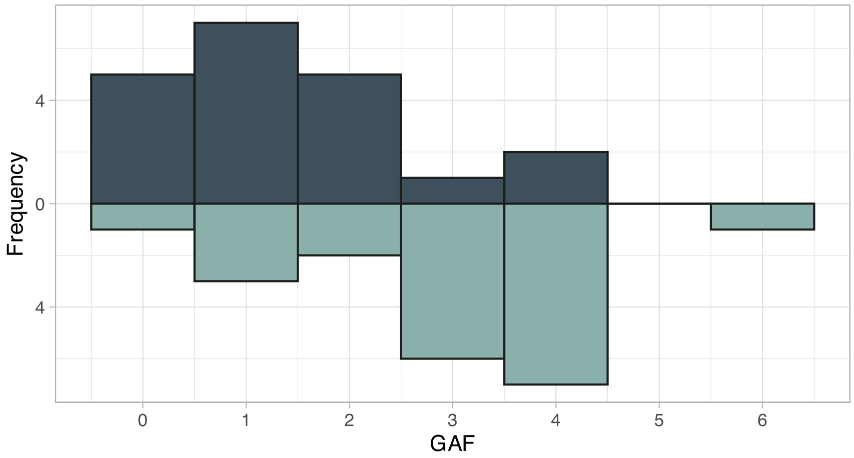
- Independent Variable (IV): Time (Pre-COVID vs. Post-COVID)
- Dependent Variable (DV): Mental well-being (GAF scores)
- Design: Repeated measures (paired observations)

Here are the results:

ID	Pre	Post
1	1	3
2	4	1
3	1	0
4	1	4
5	2	1
6	1	2
7	2	3
8	2	3
9	0	4
10	3	3
11	1	1
12	0	6
13	0	4
14	1	4
15	1	3
16	2	4
17	2	2
18	4	4
19	0	3
20	0	4

After inspecting the data visually, we can tell the data are positively skewed and may not meet the assumptions of the paired samples t-test.

A Shapiro-Wilks test revealed that our data are non-normally distributed (all $ps < .05$).



Pre is above x-axis. Post is below x-axis.

The signed rank test is similar to the rank sum test, with some minor adjustments. In short, the test is completed by:

1. Calculate difference scores for each participant.
2. Ignore zero differences.
3. Rank the absolute value of the differences.
4. Apply the sign of the original difference to the ranks.
5. Compute the sum of positive ranks and the sum of negative ranks.
6. The Wilcoxon test statistic, W (or V), is the smaller of these two sums.

- Note: some software will just use the value for the positive change ranks

We can summarize steps 1, 2, 3, and 4 in the following table:

ID	Pre	Post	Difference	AbsoliteDifference	Sign	Rank	SignedRank	Change
1	1	3	−2	2	−1	7	−7	Increase
2	4	1	3	3	1	10.5	10.5	Decrease
3	1	0	1	1	1	3	3	Decrease
4	1	4	−3	3	−1	10.5	−10.5	Increase
5	2	1	1	1	1	3	3	Decrease
6	1	2	−1	1	−1	3	−3	Increase
7	2	3	−1	1	−1	3	−3	Increase

ID	Pre	Post	Difference	AbsoliteDifference	Sign	Rank	SignedRank	Change
8	2	3	−1	1	−1	3	−3	Increase
9	0	4	−4	4	−1	14	−14	Increase
12	0	6	−6	6	−1	16	−16	Increase
13	0	4	−4	4	−1	14	−14	Increase
14	1	4	−3	3	−1	10.5	−10.5	Increase
15	1	3	−2	2	−1	7	−7	Increase
16	2	4	−2	2	−1	7	−7	Increase
19	0	3	−3	3	−1	10.5	−10.5	Increase
20	0	4	−4	4	−1	14	−14	Increase

We can then sum the ranks for the positive and negative ranks separately. Here, we get:

Change	Sum
Decrease	16.5
Increase	119.5

24.4.2 Critical Value

To determine significance, we can compare our observed W (or V) to critical values in a Wilcoxon Signed-Ranks Table. For small samples (e.g., $n = 15$), these tables provide exact values. Alternatively, software can compute exact p-values or use a normal approximation. Looking up our critical value here, we can see that for a sample size of 20, the $p < .05$ when $V < 52$ (our smallest value is 16.5).

Signed Rank Test in R

You can perform the test in R using the same `wilcox.test()` function as above, but you need to specify that `paired = TRUE`. For example:

```
wilcox.test.(DV ~ IV, data = your_data, paired = TRUE,
alternative = "two.sided")
```

For our data it would result in (and give an exact p-value):

```

    Wilcoxon rank sum test with continuity correction

data:  GAF by Time
W = 315, p-value = 0.00157
alternative hypothesis: true location shift is not equal to
0
```

24.4.3 When to Use

Use Wilcoxon's Matched-Pairs Signed-Ranks Test when:

1. You have **paired samples** or **repeated measures**.
2. The assumption of normality is violated.
3. Your data are **ordinal** or you prefer a robust nonparametric method.

24.4.4 Signed Rank Test - Write-Up

Results

To examine the effect of time (pre- vs. post-COVID) on adolescent mental well-being, a Wilcoxon's Matched-Pairs Signed-Ranks Test was conducted to compare Global Assessment of Functioning (GAF) scores before and after the onset of the COVID-19 pandemic. The results revealed a statistically significant difference in well-being scores across time points, $V = 119.5$, $p = .008$. This suggests that students' mental well-being changed significantly from pre- to post-COVID. Specifically, scores were higher after the onset of COVID-19, indicating an worsening in functioning. These findings suggest that some students may have experienced psychological impairments in the face of pandemic-related changes.

24.5 Kruskal-Wallis Test

There may be research scenarios where we have three or more groups and want to test differences on some ranked outcome. Here, the Kruskal-Wallis (*KW*) test is a suitable method to analyse the data. Conceptually, this test is similar to our previous non-parametric tests. Let's work through an example.

24.5.1 Gatorade, Milk, or H20?

You are hired by the Canadian Olympic Committee to test the impacts of types of drinks prior to competition. They want to know if certain drinks increase performance during short races. You decide to help them by conducting a brief study. In this study, you aim to investigate whether the type of drink consumed prior to a 2km race impacts performance. You manage to recruit 15 lucky participants. Participants were randomly assigned to one of three conditions, which required them to drink 500mL of one of three liquids: Gatorade, milk, or water. After consuming their assigned beverage, all participants completed the race individually. Performance was measured by finishing place (i.e., rank), with lower ranks indicating faster times (i.e., first place ran the fastest). You hypothesize that individuals who drink Gatorade will perform better.

In this scenario, a KW test is suitable. It can be used to determine whether race ranks differed significantly across the three drink conditions (i.e., people who drink Gatorade typically rank higher).

In short, the Kruskal-Wallis test is a nonparametric alternative to one-way ANOVA, used when comparing **three or more independent groups**. The test is completed by:

1. Combine all scores from all groups into a single list.

2. Rank all values from lowest to highest, regardless of group.
3. Sum the ranks within each group.
4. Compute the Kruskal-Wallis H statistic, which evaluates how different the group rank sums are from what would be expected under the null hypothesis.
5. Compare the H statistic to a chi-square distribution with $k-1$ degrees of freedom (where k is the number of groups).
6. If the result is statistically significant, it suggests that at least one group differs in median from the others.

After collecting data, we assign ranks (steps 1 and 2). We obtain the following data:

ID	Group	Place
1	Milk	12
2	Milk	9
3	Milk	14
4	Milk	15
5	Milk	13
6	Water	5
7	Water	1
8	Water	3
9	Water	8
10	Water	7
11	Gatorade	2
12	Gatorade	6
13	Gatorade	4
14	Gatorade	10
15	Gatorade	11

Here we have our ranks. For other contexts, you may need to manually assign ranks to the data.

For step 3, we will not calculate the sum of ranks for each group.

Group	Sum
Gatorade	33
Milk	63
Water	24

Next we calculate the H statistic, which is defined as:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(N+1)$$

Where:

- N is our total sample size;
- n_i is the sample size for group i;
- T_i is the sum of ranks for group i

So, for our data, the H statistics can be calculated as:

$$H = \frac{12}{15(16+1)} \times \left(\frac{33^2}{5} + \frac{63^2}{5} + \frac{24^2}{5} \right) - 3(15+1)$$

$$= 0.047 \times (217.8 + 793.8 + 115.2) - 48$$

$$= 56.34 - 48$$

$$= 8.34$$

Importantly, the H statistics is distributed as a χ^2 with $df = k - 1$ (k is the number of groups). Thus, we can compare to a critical χ^2 value, as can be found on many websites), or calculate an exact p-value using statistical software. For example, in R:

```
pchisq(q = 8.34, df = 2, lower.tail = F)
```

Which results in:

```
[1] 0.0154523
```

KW Test in R

The KW test can be done using the `rstatix` package. The `kruskal_test()` function is similar to our Wilcox function:

```
kruskal_test(DV ~ Group, data=your_data)
```

The results from our example would be:

```
# A tibble: 1 × 6
  .y.        n statistic    df       p method
* <chr> <int>     <dbl> <int>   <dbl> <chr>
1 Place    15      8.34     2 0.0155 Kruskal-Wallis
```

24.5.2 KW Post-hoc Analyses

Like an ANOVA, the KW test is an omnibus test. It can tell us whether there is an overall difference between the groups, but does not let us know where it is. As a result, we need to conduct post-hoc analyses, which will compare the rankings of our various groups. This is primarily done through **Dunn's Test of Multiple Comparisons**.

Dunn's test provides a z-score for each group comparison. By now, you have a sound understanding of z-scores and their interpretations, including what may be considered an unlikely value. We can calculate Dunn's test using the following:

$$z_i = \frac{y_i}{\sigma_i}$$

where:

- z_i is the resulting z-score for comparison i
- y_i is the difference in mean of the sum of the for two groups, $\bar{W}_A - \bar{W}_B$
- σ_i is the standard error of the differences, which is given by:

$$\sigma_i = \sqrt{\frac{N(N+1)}{12} - \frac{\sum_{s=1}^{r} \tau_s^3 - \tau_s}{12(N-1)} \left(\frac{1}{n_i} + \frac{1}{n_j}\right) \cdot C}$$

Wow. That's a handful. Don't worry, we will use our statistical software to calculate our resulting z-scores.

Dunn Test in R

The `rstatix` package also has a `dunn_test()` function. The structure is the same as we have encountered. I like to adjust using the Bonferroni correction. Also, setting `detailed = T` provides some other useful output.

```r
dunn_test(DV ~ IV, data = your_data, p.adjust.method =
'bonferroni', detailed = T)
```

In the following, which is the results from our Gatorade, milk, and water example, 'statistic' is the z-score.

.y.	group1	group2	n1	n2	statistic	p	p.adj
Place	Gatorade	Milk	5	5	2.1213	0.033895	0.10168
Place	Gatorade	Water	5	5	−0.6364	0.524518	1
Place	Milk	Water	5	5	−2.7577	0.005821	0.01746

Here we want to focus on the adjusted p-value, which has the Bonferroni correction. As we can see, the only pairwise difference that is unlikely given the null is between the Milk and Water groups. Specifically, the water group finished faster than the milk group. The other comparisons were not unexpected under the null.

24.5.3 KW Test - Write-up

Results

The results of the Kruskal-Wallis test suggests that the ranks of the three groups unexpected given a true null hypothesis, $\chi^2(2) = 8.34, p = .016, \eta^2_H = .528$.

Bonferroni-corrected post-hoc test were conducted to determine additional group differences. Specifically, post-hoc tests indicate that the milk group had lower rankings than the water group, $z = -2.76, p = .017$.

However, the Gatorade group did not differ from the milk group, $z = -2.12, p = .102$, nor the water group, $z = -0.636, p > .999$.

24.6 Friedman Test

There may be research scenarios where we have three or more **repeated-measures** groups and want to test differences on some ranked outcome. Here, the Friedman test is a suitable method to analyse the data. Here, we have the same experimental units (e.g., people) measured multiple times with some ranked data.

24.6.1 Teaching Methods and Student Outcomes

You're interested in evaluating how different teaching methods affect students' performance. You recruit 10 undergraduate students, and each student receives instruction in three different teaching formats over the course of three weeks:

1. Traditional Lecture
2. Problem-Solving Workshop
3. Online Learning Module

After each session, students complete a standardized test designed to assess their understanding.

In short, the **Friedman test** is a nonparametric alternative to repeated-measures ANOVA, used when comparing **three or more related (paired) groups**. The test is completed by:

1. Organize the data so that each row represents a subject (or matched set), and each column represents a treatment or condition.
2. Rank the scores across each row (i.e., within each subject) from lowest to highest. Tied values receive average ranks.
3. Sum the ranks for each treatment condition (i.e., column-wise).
4. Compute the Friedman test statistic (Q or χ^2_F), which evaluates whether the rank sums differ more than expected by chance under the null hypothesis.

5. Compare the statistic to a chi-square distribution with $(k-1)$ degrees of freedom (where k is the number of conditions).
6. If the result is statistically significant, it suggests that at least one condition differs in its effect compared to the others.

We obtain the following data:

ID	Lecture_Score	Problem_Score	Online_Score
1	14.2	17.8	14.4
2	14.2	21.6	9
3	15.4	20.9	14
4	14.7	16.2	8.1
5	16.9	20.3	17.7
6	16.3	19.7	12.6
7	16.6	16.4	15.6
8	14.8	18.9	8.1
9	18.4	19.4	9.4
10	12.4	18.8	8.7

Unfortunately, due to our small sample size and non-normal data, we must rank the data. For step 1 and 2, we will rank each row's (i.e., unit/person) data. Consider person 1–their highest score was problem based, followed by online, followed by lecture. Thus, they would receive rankings accordingly. Completing this for each row would result in (note that a rank of 1 indicates a 'highest' score).

ID	Lecture_Score	Problem_Score	Online_Score
1	14.2	17.8	14.4
2	14.2	21.6	9
3	15.4	20.9	14
4	14.7	16.2	8.1
5	16.9	20.3	17.7
6	16.3	19.7	12.6
7	16.6	16.4	15.6
8	14.8	18.9	8.1

ID	Lecture_Score	Problem_Score	Online_Score
9	18.4	19.4	9.4
10	12.4	18.8	8.7

ID	Lecture_Score	Problem_Score	Online_Score
1	3	1	2
2	2	1	3
3	2	1	3
4	2	1	3
5	3	1	2
6	2	1	3
7	1	2	3
8	2	1	3
9	2	1	3
10	2	1	3

Next we will sum the ranks of each treatment condition. Intuitively, this makes sense. If one method is superior, then the ranks should be higher (or lower, depending on how you coded) for that group. In our example, if Problem-Solving workshops are superior, people's ranks should be more '1' than '2' or '3'. Thus, we we add the sum of ranks, if should be lower than the other groups.

When we add the sum of ranks for each group, we get:

Teaching_Method	Sum
Lecture_Score	21
Online_Score	28
Problem_Score	11

The Q (χ^2_F) statistic (with $df = k - 1$) is defined as:

$$\chi^2_F = \frac{12}{Nk(k+1)} \times \left(\sum R_i^2 - 3N(k+1)\right)$$

Where:

- R_i^2 = squared sum of ranks for condition i
- N is the number of participants
- K is number of conditions/groups

For our example:

$$\chi_F^2 = \frac{12}{10(3)(4)} \times (21^2 + 28^2 + 11^2) - 3(10)(4)$$

$$= 0.1 \times 1346 - 120$$

$$= 14.6$$

We can then compare this statistic to a critical value table or, more likely, get the exact p-value from our statistical software. For example:

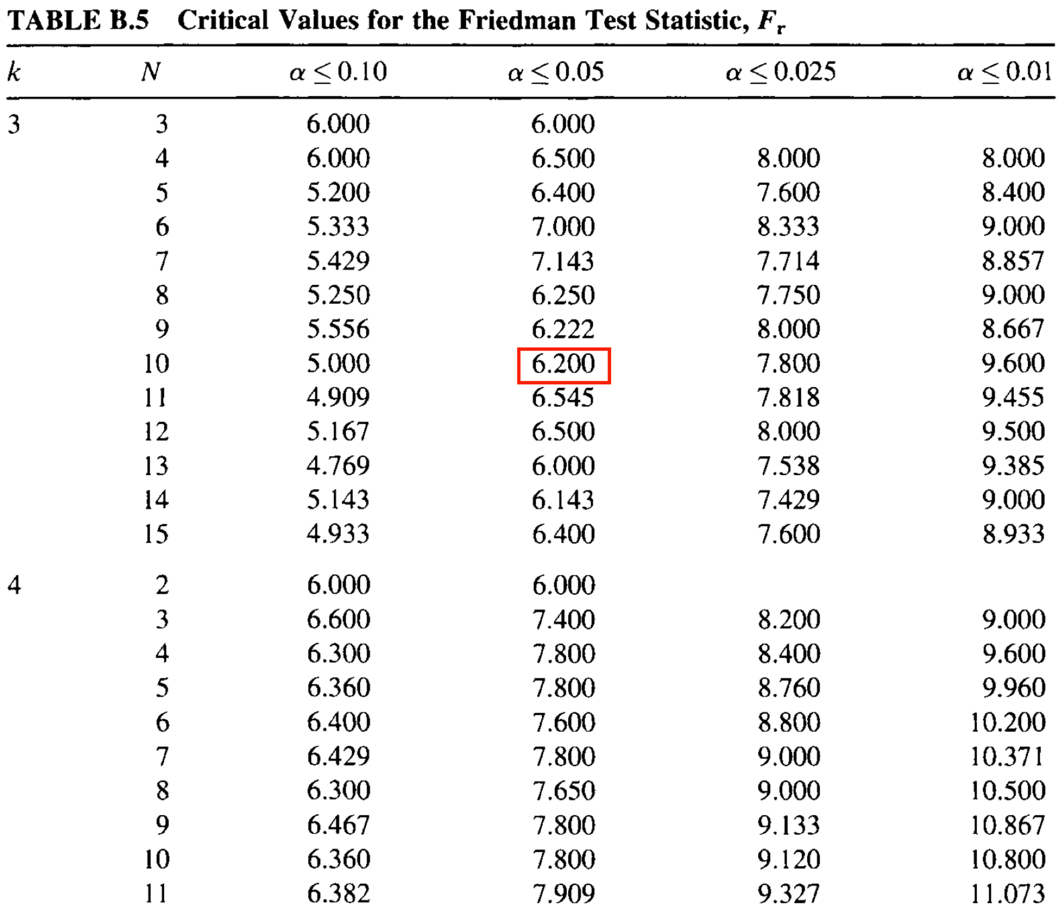


TABLE B.5 Critical Values for the Friedman Test Statistic, F_r

k	N	$\alpha \leq 0.10$	$\alpha \leq 0.05$	$\alpha \leq 0.025$	$\alpha \leq 0.01$
3	3	6.000	6.000		
	4	6.000	6.500	8.000	8.000
	5	5.200	6.400	7.600	8.400
	6	5.333	7.000	8.333	9.000
	7	5.429	7.143	7.714	8.857
	8	5.250	6.250	7.750	9.000
	9	5.556	6.222	8.000	8.667
	10	5.000	6.200	7.800	9.600
	11	4.909	6.545	7.818	9.455
	12	5.167	6.500	8.000	9.500
	13	4.769	6.000	7.538	9.385
	14	5.143	6.143	7.429	9.000
	15	4.933	6.400	7.600	8.933
4	2	6.000	6.000		
	3	6.600	7.400	8.200	9.000
	4	6.300	7.800	8.400	9.600
	5	6.360	7.800	8.760	9.960
	6	6.400	7.600	8.800	10.200
	7	6.429	7.800	9.000	10.371
	8	6.300	7.650	9.000	10.500
	9	6.467	7.800	9.133	10.867
	10	6.360	7.800	9.120	10.800
	11	6.382	7.909	9.327	11.073

In the above image, we would look to k (number of groups) and n (total sample size). For us, the critical value at $\alpha = .05$ is $\chi_F^2 = 6.20$.

Friedman Test in R

It's similar to our last tests–shocker! We would use the `rstatix` package and the `friedman_test()` function. The major difference is we need to specify our ID variable. We should have a column specifying the repeated treatment/group, a column for outcome, and a column for ID/participant specifier. Our function is specified as:

```r
friedman_test(DV ~ IV | ID, data = your_data)
```

Alternatively, you could specify three columns, each representing the DV, IV, and ID for participants. For example:

```r
friedman.test(your_data$DV, your_data$IV, your_data$ID)
```

Which, for our example, would result in:

```
# A tibble: 1 × 6
  .y.            n statistic    df        p method
* <chr>      <int>     <dbl> <dbl>    <dbl> <chr>
1 Test_Score    10      14.6     2 0.000676 Friedman test
```

24.6.2 Friedman Test - Effect Size

The typical effect size for the Friedman test is Kendall's W, which is defined as:

$$W = \frac{\chi^2_f}{N(K-1)}$$

Kendall's W has possible values of 0-1, with higher values indicating a higher effect size. For our data, the effect size is:

$$W = \frac{14.6}{10(3-1)}$$

$$W = .73$$

24.6.3 Friedman Post-hoc

Much like our KW test, we will need to conduct post-hoc comparisons to determine where the group differences lie. We can do this using Bonferroni-adjusted signed rank tests.

```
# A tibble: 3 × 8
  .y.         group1         group2            n1    n2 statistic
p p.adj
  <chr>       <chr>          <chr>          <int> <int>     <dbl>
<dbl> <dbl>
1 Test_Score Lecture_Score Online_Score      10    10        52
0.014 0.043
2 Test_Score Lecture_Score Problem_Score    10    10         1
0.004 0.012
3 Test_Score Online_Score  Problem_Score    10    10         0
0.002 0.006
```

24.6.4 Friedman - Write-up

We conducted a Friedman test to determine whether tests scores were associated with teaching method. The results suggest that test score ranking were statistically significantly associated with teaching method, $\chi^2_F = 14.6, p < .001, W = .73$.

Post-hoc analysis were conducted using Bonferroni-adjusted signed rank tests. These results suggest that, first, students scored significantly higher following the **problem-solving method** compared to the **lecture method**, $V = 1$, $p_{\text{adj}} = .012$. Second, the **problem-solving method** also significantly outperformed the **online method**, $V = 0$, $p_{\text{adj}} = .006$. Last, scores following the **lecture method** were significantly higher than those following the **online method**, $V = 52$,$p_{\text{adj}} = .043$.

Taken together, these results suggest that all three teaching methods yielded significantly different student outcomes, with the **problem-solving approach consistently associated with better performance**.

24.7 Conclusion

Non-parametric tests are great options for data that violates assumptions of the more common parametric methods. These tools will help you answer important research questions when the data requires. Keep in mind that these tests are not inferior and, instead, have appropriate use cases–just like t-tests, ANOVAS, and regression.

24.8 Practice Questions

1. A researcher wants to compare **three independent groups** using ordinal data. Which nonparametric test should they use?
2. You conduct a **Friedman test** on students' scores across three different learning conditions and obtain a significant result. What post hoc test would be appropriate, and why?
3. What is the primary assumption difference between the **Mann-Whitney U test** and the **Wilcoxon signed-rank test**?
4. In the context of the **Kruskal-Wallis test**, why do we apply a tie correction to the standard error formula in post hoc tests?
5. Suppose you conduct multiple pairwise comparisons after a Friedman test. What correction method can you use to control for Type I error?
6. You find that the **Wilcoxon signed-rank test statistic (V) = 0**. What does this imply about your data?
7. Describe one situation where you would use the **sign test** instead of the **Wilcoxon signed-rank test**.
8. What are the **null and alternative hypotheses** for a Kruskal-Wallis test?

24.9 Answers

1. The **Kruskal-Wallis test** is appropriate for comparing three or more independent groups using ordinal data.
2. The **Dunn-Bonferroni** test is appropriate for pairwise comparisons following a significant Friedman test because it controls for Type I error in multiple comparisons using ranked data.
3. The Mann-Whitney U test compares **independent samples**, while the Wilcoxon signed-rank test is for **paired or related samples**.
4. A tie correction is applied because **tied ranks reduce the variability** in the rank distribution, which affects the accuracy of the standard error and p-value.
5. The **Bonferroni correction** (or Holm-Bonferroni) is commonly used to adjust p-values when conducting multiple pairwise comparisons.
6. A (V) value of 0 suggests that **all participants scored higher (or lower) in one condition than the other**, indicating a strong directional difference.
7. Use the **sign test** when the **magnitude of differences isn't meaningful or reliable**, such as when data violate the assumptions of the Wilcoxon test (e.g., non-symmetric distributions).
8. **Null hypothesis**: All groups come from the same population (i.e., have the same median ranks).
   Alternative hypothesis: At least one group differs in its distribution (i.e., in median ranks) from the others.

24.10 Practice Questions 2

A health psychologist is studying whether different **relaxation techniques** affect self-reported **anxiety levels**. Each of **10 participants** tries **three different methods** on separate days:

1. **Deep breathing**

2. **Progressive muscle relaxation (PMR)**
3. **Mindfulness meditation**

After each session, participants rate their **anxiety level** on a scale from 1 (no anxiety) to 10 (very anxious). The psychologist uses a **Friedman test** to determine whether there are statistically significant differences in anxiety ratings across the three techniques.

Participant Anxiety Ratings

Participant	Breathing	PMR	Mindfulness
1	5	4	3
2	6	5	4
3	7	6	4
4	4	3	2
5	5	4	2
6	6	5	3
7	7	6	5
8	6	5	3
9	5	4	3
10	4	3	2

Use the data above to perform a **Friedman test** to determine whether anxiety levels differ across the three relaxation techniques.

- Step 1: Rank the anxiety ratings **within each participant**
- Step 2: Sum the ranks for each technique
- Step 3: Use the Friedman formula or statistical software to calculate the test statistic
- Step 4: Interpret the result

Step 1: Rank within each participant (lower anxiety = better)

Participant	Breathing	PMR	Mindfulness	R_Breathe	R_PMR	R_Mind
1	5	4	3	3	2	1
2	6	5	4	3	2	1
3	7	6	4	3	2	1

Participant	Breathing	PMR	Mindfulness	R_Breathe	R_PMR	R_Mind
4	4	3	2	3	2	1
5	5	4	2	3	2	1
6	6	5	3	3	2	1
7	7	6	5	3	2	1
8	6	5	3	3	2	1
9	5	4	3	3	2	1
10	4	3	2	3	2	1

Step 2: Sum ranks for each condition

- Breathing: $3 \times 10 = 30$
- PMR: $2 \times 10 = 20$
- Mindfulness: $1 \times 10 = 10$

Step 3: Friedman Test Formula

$$\chi_F^2 = \frac{12}{nk(k+1)} \sum R_j^2 - 3n(k+1)$$

$$= \frac{12}{10 \cdot 3 \cdot 4}(30^2 + 20^2 + 10^2) - 3 \cdot 10 \cdot 4$$

$$= \frac{12}{120}(900 + 400 + 100) - 120 = \frac{12}{120}(1400) - 120 = 140 - 120 = 20$$

Step 4: Conclusion

With $\chi^2(2) = 20.00$, $p < .001$, the Friedman test is significant. Anxiety levels **significantly differed** depending on the relaxation technique used. Post hoc comparisons would likely show that **mindfulness** resulted in **lower anxiety ratings** than either breathing or PMR.

25 Concluding Remarks

The purpose of this chapter is to briefly review and consolidate some of the knowledge and skills covered in this book. But first, I want to congratulate you on getting this far. Research and statistics can feel intimidating–both before beginning and while in the midst of your analysis. You have powered through the course (or just book), and that deserves recognition of perseverance and courage. Great job!

25.1 Psychological Science

Becoming a scientist requires you to have a thorough understanding of the scientific process and some key concepts relevant to psychology and other fields. Here are some of my recommendations on becoming a scientist. First, recall that sound science is the interweaving of inductive and deductive methods, which allow scientists to generate ideas and theories and then test them through well-designed research methods. The cyclical nature of generating theories, deriving hypotheses, and conducting research–which subsequently fine-tunes our theories–is believed to make science progressive. My recommendation to budding scholars is to become familiar with psychological theories of interest, as they will be the foundation of your research.

Second, being a psychological scientist requires being aware of, acknowledging, and working to remedy the major issues with current psychological science culture and practice. For example, concerns with reproducibility and lack of appreciation for replications are major concerns in our science. Working to remedy these through open science or other ways to promote honesty and transparency in our research methods is everyone's business.

Third, this book focuses solely on quantitative analyses. I have neglected a major portion of research methods that draw on either qualitative research or the blending of qualitative and quantitative research (i.e., mixed-methods designs). I may integrate qualitative methods into a later edition, but for now, please refer to such introductory texts as Banister et al. (2011) or Smith (2024).

In line with this, it is common for individuals to perceive qualitative methods as inferior or less important than quantitative methods. For example, the misconception that qualitative work is not valuable because it's not generalizable exists (Povee & Roberts (2014)). I strongly recommend you alter this mindset. Each research method has a time and place; the research question and hypotheses determine the methods. Thus, certain questions require a quantitative approach and others a qualitative approach. Regardless, just because you encounter something less often—qualitative research—does not mean it is not useful or important. Unfortunately, this is the mindset of many psychological scientists. Avoid this and expand your research toolbox as large as possible. Mixed-methods research is valuable and informative.

Last, while your journey in this book is complete, I invite you to approach psychological science as a lifelong learner. There are always new concepts, analyses, or other concepts to learn, which can improve the way you "science." As such, keep an open mind to learn and try new developments in research methods and statistics.

25.2 Avoiding Common Pitfalls

Here is a brief list of recommended practices to avoid common pitfalls in psychological science:

1.Do not put too much weight in p-values. Understand what they mean ($p(DATA \mid H_0)$) and don't mean ($p(H_0 \mid DATA)$). Small p-values do not necessarily mean large or relevant effects.

2. Ensure you test all required assumptions of an analysis. The test statistic distributions or use cases are built on assumptions. Violating the assumptions means the statistics do not operate the way they should. When violated, alter the approach or implement a recommended change (e.g., use a non-parametric alternative).

3. Report effect sizes and confidence intervals. Effect sizes are more informative than p-values. Including confidence intervals, in most cases, can tell you just as much as a p-value. For example, reporting indicates that the test was statistically significant for that alpha level ($p < .05$ for $95\% \, CI$). It also tells us that the most likely population effect is , but that any value between and are plausible.

4. Use APA formatting. Although I often question some of APA's recommendations on reporting and I typically adopt a more flexible approach, be sure to at least LEARN APA formatting. Many institutions or academic journals require this formatting.

5. When unsure, ask. You likely have resources, colleagues, supervisors, or former professors who you can draw on to help you with your analyses. Ask for help when you need it.

I would avoid using generative artificial intelligence (GAI) for your analyses right now. I have performed some preliminary testing and sometimes questioned GAI's approach, which sometimes returns results that are incorrect. I hope to publish some results in this area in the future. Instead, use the connections you have built over your degree.

25.3 Final Thoughts

This book was a lot of work. I'm tired.

References

Banister, P., Bunn, G., Burman, E., Daniels, J., Duckett, P., Goodley, D., Lawthom, R., Parker, I., Runswick-Cole, K., Sixsmith, J., & others. (2011). *Qualitative methods in psychology: A research guide*. McGraw-Hill Education (UK).

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173.

Beck, A. T., Steer, R. A., Brown, G. K., & others. (1996). *Beck depression inventory*.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Chambers, C. (2019). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*.

Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist*, *49*(12), 997.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

Cowles, M., & Davis, C. (1982). On the origins of the. 05 level of statistical significance. *American Psychologist*, *37*(5), 553.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*(4), 532–574.

Dahiru, T. (2008). P-value, a true test of statistical significance? A cautionary note. *Annals of Ibadan Postgraduate Medicine*, *6*(1), 21–26.

Docherty, A. R., Shabalin, A. A., DiBlasi, E., Monson, E., Mullins, N., Adkins, D. E., Bacanu, S.-A., Bakian, A. V., Crowell, S., Chen, D., & others. (2020). Genome-wide association study of suicide death and polygenic prediction of clinical antecedents. *American Journal of Psychiatry*, *177*(10), 917–927.

Hales, A. H. (2023). One-tailed tests: Let's do this (responsibly). *Psychological Methods*.

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 70–85.

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin*, *54*(4), 351.

Krause, N., & Borawski-Clark, E. (1995). Social class differences in social support among older adults. *The Gerontologist*, *35*(4), 498–508.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863.

Lerner, R. M., Lerner, J. V., P. Bowers, E., & John Geldhof, G. (2015). Positive youth development and relational-developmental-systems. *Handbook of Child Psychology and Developmental Science*, 1–45.

The MTA Cooperative Group. (1999). A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *56*(12), 1073–1086.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. McGraw-Hill Companies,Incorporated.

O'Connor, K., Aardema, F., Robillard, S., Guay, S., Pelissier, M.-C., Todorov, C., Borgeat, F., Leblanc, V., Grenier, S., & Doucet, P. (2006). Cognitive behaviour therapy and medication in the treatment of obsessive–compulsive disorder. *Acta Psychiatrica Scandinavica*, *113*(5), 408–419.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, *19*(4), 459.

Popper, K. (1959). *The logic of scientific discovery*. Routledge.

Povee, K., & Roberts, L. D. (2014). Qualitative research in psychology: Attitudes of psychology students and academic staff. *Australian Journal of Psychology*, *66*(1), 28–37.

Schatz, N. K., Fabiano, G. A., Cunningham, C. E., dosReis, S., Waschbusch, D. A., Jerome, S., Lupas, K., & Morris, K. L. (2015). Systematic review of patients' and parents' preferences for ADHD treatment options and processes of care. *The Patient-Patient-Centered Outcomes Research*, *8*(6), 483–497.

Schneidman, E. S. (1998). Perspectives on suicidology: Further reflections on suicide and psychache. *Suicide and Life-Threatening Behavior*, *28*(3), 245.

Smith, J. A. (2024). *Qualitative psychology: A practical guide to research methods.*

Spence, J. R., & Stanley, D. J. (2018). Concise, simple, and not wrong: In search of a short-hand interpretation of statistical significance. *Frontiers in Psychology*, *9*, 2185.

Stevens, S. S. (1951). *Mathematics, measurement, and psychophysics* (pp. 1–49). Wiley.

Strickland, B., & Suben, A. (2012). Experimenter Philosophy: the Problem of Experimenter Bias in Experimental Philosophy. *Review of Philosophy and Psychology*, *3*(3), 457–467. https://doi.org/10.1007/s13164-012-0100-9

Van Orden, K. A., Witte, T. K., Cukrowicz, K. C., Braithwaite, S., Selby, E. A., & Joiner, T. E. (2010). The Interpersonal Theory of Suicide. *Psychological Review*, *117*(2), 575–600. https://doi.org/10.1037/a0018697

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*(2), 158.

Wu, S., Wu, F., Ding, Y., Hou, J., Bi, J., & Zhang, Z. (2017). Advanced parental age and autism risk in children: a systematic review and meta-analysis. *Acta Psychiatrica Scandinavica*, *135*(1), 29–41.

Zhao, X., Lynch Jr, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, *37*(2), 197–206.