

GRADUATE-LEVEL STATISTICS IN PSYCHOLOGY



Peter Ji
Adler University



Graduate-Level Statistics in Psychology

By Dr. Peter Ji
Adler University

Acknowledgements

Contributors

This textbook would not have been possible without the assistance and support that I received from my colleagues on this project. I'm grateful to Dorinne Banks for keeping me on track with the many tasks associated with creating this textbook, and for stepping in to complete the tasks that gave me trouble. I appreciate the collaborative work of Matt Barrington--for his editing and being my "right-hand man" for many years with statistics projects. I'm grateful for the editing prowess of Amin Ghorbaniasad and for his assistance with reviewing the accuracy of my writing. I'm grateful for Ariel Orlov and her review of my writing for conceptual clarity.

Thank you all for making this project a reality!

~Peter Ji

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

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

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



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

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

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

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

This text was compiled on 04/19/2026

Licensing

CC-BY 4.0 "Graduate-Level Statistics in Psychology" by Dr. Peter Ji

TABLE OF CONTENTS

Acknowledgements

LibreTexts Information

Licensing

0: Introduction

- 0.1: A Perspective on Statistics
- 0.2: Psychology Profession
- 0.3: Evidence-Based Practice
- 0.4: Scientist-Practitioner Model; Scholar-Practitioner Model
- 0.5: Social Justice
- 0.6: Description of Intended Audience
- 0.7: Approach to Learning Statistics

1: What is a Number?

- 1.1: How to Start with Statistics
- 1.2: A "Geek Out" Explanation of Variation
- 1.3: Variations Lead to Numbers
- 1.4: The Battle of Words Versus Numbers
- 1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words
- 1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift
- 1.7: Statistics Do Something with Numbers
- 1.8: Numbers Don't Lie, but Numbers Don't Tell the Full Story
- 1.9: Seeing Numbers – What I Want You to Learn
- 1.10: Connection Between Understanding How Things Vary and What You are Trying to Do
- 1.11: Discussion Questions

2: What are Statistics?

- 2.1: Statistics Involves Comparison and Context
- 2.2: Statistics is About Finding Patterns
- 2.3: Statistics as Estimates, or Making Guesses
- 2.4: Two Broad Categories of the Process of Using Statistics
- 2.5: Discussion Questions

3: Variable Types

- 3.1: Variables - Not Always Obvious
- 3.2: How to Organize Variation With Variables
- 3.3: Categorical Variables- Variables That Vary by Type
- 3.4: Continuous Variables- Variables That Vary by Level
- 3.5: Nuances in Determining if a Variable Should be a Certain Type
- 3.6: Making Decisions About the Variables in Your Research Study
- 3.7: Lessons Learned About Variables
- 3.8: Discussion Questions

4: Distributions

- 4.1: Distributions – A Picture of Variation
- 4.2: Categorical Distributions
 - 4.2.3: Evaluating the Plot
- 4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables
- 4.4: Continuous Distributions
- 4.5: Evaluating the Quality of Frequency Distributions
- 4.6: Normal Distribution for Continuous Variables
- 4.7: Skewed Distribution- The Opposite of Normal Distribution
- 4.8: Evaluating a Continuous Distribution
- 4.9: Discussion Questions

5: Measures of Central Tendency

- 5.1: What is Central Tendency? Why Do We Use It?
- 5.2: Three Types of Central Tendency and Why We Need Them
- 5.3: The Mean, Median, and Mode in Normal and Skewed Distributions
- 5.4: Discussion Questions

6: Measures of Variability

- 6.1: What is Variance?
- 6.2: How Do You Determine Variance?
- 6.3: Partitioning of the Variance
- 6.4: Standardized Values of Reporting Scores
- 6.5: Discussion Questions

7: Sampling, Standard Error, Central Limit Theorem

- 7.1: The Purpose of Sampling
- 7.2: Sampling from the Population
- 7.3: Getting a Good Sample - Get a Random Sample
- 7.4: Sampling Bias or Error
- 7.5: Central Limit Theorem (CLT) and Its Implications
- 7.6: Summary of Reasons Why You Need a Large Sample Size
- 7.7: Discussion Questions

8: Probability, Hypothesis Testing, Type I and Type II Errors

- 8.1: The Role of Probability
- 8.2: Statistical Decision - Is a Result Significant or Not Significant?
- 8.3: Hypotheses for Statistical Inference
- 8.4: How to Determine if a Result is Significant
- 8.5: Interpreting a Statistical Test Result
- 8.6: Testing a Hypothesis
- 8.7: How Do We Get the P Value for the Statistical Test?
- 8.8: Confidence Intervals
- 8.9: Type I and Type II Errors
- 8.10: The Definition of Type I and Type II Errors
- 8.11: Discussion Questions

9: Effect Sizes and Power Analysis

- 9.1: Significant (P Value) Versus Meaningful (Effect Size) Results
- 9.2: What Are Effect Sizes?
- 9.3: Statistical Power
- 9.4: How to Conduct a Power Analysis
- 9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together
- 9.6: Discussion Questions

10: Introduction to Inferential Statistics

- 10.1: The Meaning of Inferential
- 10.2: Patterns
- 10.3: General Linear Model
- 10.4: Discussion Questions

11: T-tests

- 11.1: When to Use the Independent T-test
- 11.2: Design Set Up
- 11.3: Commentary
- 11.4: T-test Assumptions
- 11.5: Reading T-tests
- 11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology (EPPP) Licensing Exam
- 11.7: Reading Articles That Involve a T-test
- 11.8: Reading T-tests from Journal Articles
- 11.9: Summary
- 11.10: Discussion Questions

12: Analysis of Variance – ANOVA

- 12.1: When to Use the ANOVA
- 12.2: The Concept Behind Analysis of Variance
- 12.3: - Interpreting the F-test and Post-Hoc Comparisons
- 12.4: Commentary
- 12.5: Reading Articles – ANOVA
- 12.6: Summary
- 12.7: Discussion Questions

Index

Appendices

Detailed Licensing

CHAPTER OVERVIEW

0: Introduction

- 0.1: A Perspective on Statistics
- 0.2: Psychology Profession
- 0.3: Evidence-Based Practice
- 0.4: Scientist-Practitioner Model; Scholar-Practitioner Model
- 0.5: Social Justice
- 0.6: Description of Intended Audience
- 0.7: Approach to Learning Statistics

This page titled [0: Introduction](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

0.1: A Perspective on Statistics

In the end, there's not a damn thing I can do to make you enjoy statistics.

If you want to do something else other than statistics, then you do. If you prefer doing a qualitative analysis versus a quantitative analysis in your dissertation, then you do. I could “scare” you into learning statistics by stating there are future situations where you will be using statistics. The psychology profession espouses the importance of learning statistics as part of a scientist-practitioner or scholar-practitioner model. You might find yourself as part of a group that sees a need for developing measures, evaluating services, or investigating something. That is the professional pressure for learning statistics. For most of us, we would rather obtain a statistical consultant. After all, if you do not know how to fix your light bulb problem, hire an electrician to help you fix your light bulb because no one is saying that you alone must fix your light bulb. If learning statistics is out of your comfort zone, out of your orbit, if you feel external pressures to learn statistics, if you would rather be somewhere else, then there is not a damn thing I can do to help you feel better about learning statistics.

But so what?

If you are reading this textbook, then we are here, in the present moment, and we will make the best of it. Whatever you gain out of this textbook, I hope it is useful to you. It is up to you, though, how you engage in the process of getting through a statistics course. You can go through the process by doing the minimum to pass or engaging at a deeper level to learn it. I will do what I can to increase your knowledge of statistics with this textbook, but in the end, it is all about your process and your goal.

This page titled [0.1: A Perspective on Statistics](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

0.2: Psychology Profession

Statistics is thought of as an advanced, sophisticated discipline that only people who are good with numbers can understand. But that can be said for any profession. There are many ways for people to be involved in statistics.

This statistics book is written from the perspective of the psychology profession. Statistics cannot be applied uniformly across all disciplines and definitely not across all research questions. Statistics, as applied to academic disciplines such as economics or political science, medical fields, or sciences such as chemistry, forestry, or natural sciences, will have their own nuances and expectations. Statistics among different psychology topics will have nuances and expectations. Statistics, as it applies to testing cognitive-behavioral treatments for alleviating anxiety, will look different compared to understanding alcoholism among military veterans struggling with PTSD and compared to examining trauma severity for undocumented immigrant populations. The implication is that if a person blindly tells you what you should do when you apply statistics to your project without considering the context of your project, the person is an idiot. A statistics consultant should first listen to your description of the context of your project and adjust their application of statistics.

This page titled [0.2: Psychology Profession](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

0.3: Evidence-Based Practice

Psychology makes a valid and emphatic point about practitioners being good consumers of statistics as part of the process of understanding evidence-based research. We cannot conduct our psychology practice without evidence to support our claims. We cannot advertise to clients that we use trauma-informed care without providing evidence that trauma-informed care is better than usual care. We cannot advertise to society that harm reduction programs, such as clean needle sharing programs, will reduce deaths due to overdose without providing evidence to support this claim, especially when the idea of handing out clean needles to drug users is counterintuitive to those who reasonably claim that providing clean needles could increase drug use.

Statistics is integral to the process of providing evidence-based research. After collecting data, we need to process it. As psychologists, we collect clinical data about our clients to assess their functioning and prognosis. The same process applies to statistics. We collect our data about an issue and use statistics to process the data into a message about the validity of our data, which is evidence, to support our claim about the effectiveness of our psychology practice.

This page titled [0.3: Evidence-Based Practice](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

0.4: Scientist-Practitioner Model; Scholar-Practitioner Model

Acquiring updated knowledge of the field is essential for the psychology profession. Learning statistics, reviewing results, and understanding how new research advances the field is a skill for psychologists to develop. Psychologists espouse two roles for engaging in this aspiration: the scientist-practitioner model and the scholar-practitioner model.

Statistics as part of the scientist-practitioner model could be considered as using statistics for data analysis. The practitioner part of the model wants science to inform practice, and the scientist part of the model wants practice to inform science. Statistics plays a role by analyzing the data to produce conclusions about how practice should proceed. Statistics plays a role by asking questions about practice and determining how statistics could be used to address them. The scientist-practitioner uses statistics, such as by conducting the statistical analysis and interpreting the results, to achieve the goal of informing and discovery.

Statistics as part of the scholar-practitioner model could be considered as reading statistics as part of a scholarly engagement of an issue. The practitioner part of the model wants the scholarly engagement to inform practice, and the scholarly part wants the practitioner part to understand how scholarly ideas can be implemented. Statistics plays a role as the practitioner and scholar read and interpret statistical presentations. The role of a scholar-practitioner leans more on the consumer role. As a consumer, just like the phrase “let the buyer beware,” the scholar-practitioner reads the statistics, produces the conclusions of a study, and determines if the statistics results are valid and the statistical analysis decisions are valid. Evaluating the statistical validity of the study is part of the process for the scholar-practitioner to decide if the study should inform their clinical practice.

Either model involves statistics in the process. The art of conducting statistical analysis and the art of evaluating the validity of the statistical results are part of the process of deciding if science and scholarly ideas inform clinical practice. This process is evidence-based treatment and is the bedrock of psychology as a profession. We cannot do whatever we want to do simply because we believe in it. The evidence legitimizes our profession, and either the scientist-practitioner model or the scholar-practitioner model will involve statistics in the process.

This page titled [0.4: Scientist-Practitioner Model; Scholar-Practitioner Model](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

0.5: Social Justice

Social justice needs research and, hence, statistics to understand the complexity of using psychology to further the social justice mission. There are many underserved, overlooked, complex populations and messy issues. We lack a framework to organize our perspectives on these issues. Statistics, as applied to social justice issues, can be used to make sense of these issues. However, statistics cannot be bludgeoned into a social justice issue. When using statistics to understand social justice, the nuances and idiosyncrasies of social justice issues must be considered. For example, in statistics, sample size is a key consideration for evaluating the stability of the findings. In brief, larger sample sizes are better. In social justice issues, large sample sizes are not possible. People affected by social justice issues are often hard to find, recruit, and participate. Stigma, discrimination, and reluctance interfere with a researcher's goal of recruiting enough of a sample size to power a statistical test. Considerations when using statistics need modification because the usual rules of conducting statistics do not apply when using statistics to investigate social justice issues.

Statistics should not be a barrier to investigating social justice issues. Just because you cannot obtain a sample size of 30 participants does not mean that your study and statistical analysis should be abandoned. The important perspective is that you are trying to investigate the social justice issue. Yes, be mindful of the statistical framework, but just because you cannot satisfy the statistical framework does not mean you abandon your investigation. There are other ways to conduct statistics, but this book only discusses one of the ways, which is the general linear model. There are other possibilities to consider when investigating social justice issues that will not lend themselves to satisfying the requirements of conducting a statistical analysis.

This page titled [0.5: Social Justice](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

0.6: Description of Intended Audience

The intended audience is students in psychology graduate programs that do not have research and statistics as their primary learning objective. Psychology programs that teach students to become researchers, obtain grants, and eventually become career professors will have a statistics and research curriculum to train students in statistics. Other psychology programs do not emphasize this research training. This book is intended for those students who want to be familiar with statistics and to feel they have some command of statistics and its application.

Any student or faculty member can use this book for whatever purpose. Anybody who explores a topic will use all resources to learn something. This book intends to supplement any training you have to achieve your goal for yourself regarding statistics.

This page titled [0.6: Description of Intended Audience](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

0.7: Approach to Learning Statistics

Taking courses and reading textbooks is not enough to compel you to learn these skills. If you want to master statistics and are not enrolled in a research-oriented graduate program, quite frankly, you are in the wrong course and program.

You will encounter individuals with different opinions about the right thing to do in statistics. Different individuals will want more advanced statistics, they will want you to collect more data, they will ask you to check this or that possibility, they will ask you to test everything to cover all bases, to be thorough, they will say you are doing it wrong. Each person has their own approach they think is correct. Other professions have similar experiences. Different home improvement contractors will have different opinions of what must be done, and they will chastise others and the quality of their work. Different chefs have their own way of making the same dish. Different approaches and conclusions use the same context, situation, and data. It's hard to know what the "right way" is of going about learning statistics.

What does help you learn these skills?

Learning anything is a multi-faceted process. People who master statistics take several courses, attend several workshops, engage in research projects to apply statistics, join research groups, and are immersed in this process. Mentoring helps. Every person who becomes proficient in their field has had a mentor who guided their development, allowed questions and curiosities, and provided opportunities to practice applying their skills.

Most of us want a "working knowledge" of statistics. I propose that "working knowledge" means we want to read, review, understand, and discuss statistics in the context of understanding the craft of psychology. Instead, consider the following position - follow your curiosity. Move statistics out of the realm of overwhelming math, tense, bewildering discussions, and advanced intellectual discourse. Move your statistics into the realm of satisfying your own curiosity. To satisfy your own curiosity, start with your emotions. Tune into your own emotions about what makes you engaged and excited about your interest in your psychology topic. Approach statistics by understanding how statistics will help you understand your psychology topic. It is better to learn statistics in the context of understanding your psychology topic than understanding statistics by itself in the form of math equations and numbers. Use your own interest in psychology to guide your use of statistics.

I offer the following guides to serve your curiosity best. Start by thinking conceptually about statistics. Conceptual thinking involves the "why" of statistics. Why is it important to have a normal distribution? Conceptual thinking involves how this stuff works together. How do we understand the p value and effect size in determining significance and meaning?

Conceptual thinking is better than following the rules. Statistics is not a cookbook; it is not an IKEA manual; it is not "Statistics for Dummies." It is not a step-by-step process. It does require you to attend to the context and the scope of how you are applying statistics to answer your curiosity. Just because you take a course in statistics does not mean you are ready. There is an art to practicing statistics, and learning this new skill takes discussion, mentoring, and the freedom to ask questions. But it does take your effort, and you must begin somewhere. Start with what you know, your experience, your curiosity. If you are starting from a place of vulnerability where you feel scared, unprepared, bewildered, lost, avoidant, or indifferent, do not start there. Only you can change your starting point from a place of weakness, fear, or apathy to a place that feels engaged and positive. Start with your curiosity, start with what feels good about what you find interesting about psychology and social justice, and then figure out how to use statistics to further your curiosity.

What does help if you want to learn statistics? There seems to be universal techniques, mindsets, and practices for anyone to learn anything and adopt new habits. I would contend that psychologists espouse these same habits for their clients when they want to recover from their mental health distress.

Let us start here - Set aside the time to learn statistics. Everything we learn, develop, and nurture takes more time than we think. If you want to work out, you gotta make the time. You want to publish a creative writing piece, you gotta make the time. If you want to learn a hobby, you gotta make the time. Same with statistics. The "times four rule" applies here. Whatever you think about how much time it takes to do something, take that time and multiply it by four, and that is how long it really takes. The "times four rule" means that whenever we start something, there is always an anomaly, an oddity, a setback, or a mistake, which means you have to start over. Nothing goes linearly from point A to point B. Things happen that derail us. That is part of the gig. But that also means if you want to learn statistics and do a project, be aware that it will be a struggle, and it is common to spend more time than you think on it. That doesn't mean that you are not effective or don't know what you are doing; the struggle bus is part of the gig.

Know when to quit and seek help. We get frustrated when our statistical effort does not work. Know your limits and seek help if you hit a wall. I am reminded of my workout instructor – “If it hurts, don’t do it.” Your body has limits, and if you are in pain, stop. Seek help and see if you need to adjust or something. The internet is full of listservs, message boards and so forth that might help you with whatever you encounter. You are always welcome to contact me using my contact information. Just ask; it never hurts to ask.

Have a plan. Learning any endeavor means having a structure to learn. Practicing music requires warmups with scales, practicing your breathing using long tones, practicing your fluid playing using legato exercises, practicing your articulation with staccato studies, and practicing your expression by practicing the pieces you will perform. Working out requires warmups, then focusing on which body part to build, then on aerobics, then on building your core, then cool down and stretches. Learning statistics follows similar routines. Start with reading, start with demonstrations, start with datasets, take courses, and take workshops.

Get a mentor. As an extension of having a plan, it helps to have a mentor. Practicing music requires a private teacher; working out requires a coach. Even Neil Peart, the best drummer on the planet, took lessons from Freddie Gruber to learn more about his craft. In statistics, you are on your way if you can find a trusted mentor. If you need one, ask me.

Change your mindset. There is no “secret sauce,” as my colleague likes to say, no gift to give you to transform you into a statistician. The mindset of “I just want to get through this” will not help you absorb these lessons. If you have a problem where statistics blocks your thinking, sit with that block and ask yourself if you want to let that block define you. Then, get help with your mindset block because you need someone else to find your potential for learning statistics or anything else for that matter.

Stress is an opportunity, not a signal of impending failure. Related to the previous point, stress does not mean you failed; stress means that something needs addressing. When you are stressed about statistics, it is a learning opportunity, not a sign of failure.

Put yourself out of your comfort zone. Related to the previous point, venturing out of your comfort zone will yield dividends. Joining statistical teams and being part of research projects are ways to get yourself involved in the process. You might not feel ready, but no one is ever ready.

Being constant. Like everything, keep at it. Show up. Statistics is hard to learn, but by continuing to apply what you know, improvement will happen. Progress is slow. According to workout coaches, it takes four weeks of workout to gain muscle and two weeks of atrophy to lose it. Slowly build your competency. And try not to let it slack.

What doesn't work when learning statistics?

Students can get tripped up by focusing on surface definitions of statistics terms. Knowing the definition of the mean, median, and mode does nothing for you without understanding why we have those different terms in the first place and why they have different uses in psychology research. Students get tripped up or fear math equations. The goal is to think conceptually about how those math equations work. Why do we use standard error in computing a statistic? Students get tripped up in trying to connect all statistical terms together. Students ask questions such as “But what about this, what about that?” Unless there is a conceptual reason to make sense of all statistics terms all at once, it is best to learn statistics one term at a time. Otherwise, students run the risk of spiraling down a rabbit hole and trying to connect everything together when there is no need to.

Do not try to impress with over-the-top jargon and try to make yourself more advanced than you really are. Related, do not be with people who intimate you by giving you the impression that they know more than they do and that you don’t know anything. Don’t be with people who criticize you and your use of statistics. Intimidation never works anywhere, and definitely not in the field of statistics. We are human beings; let us relate to each other as human beings.

Finally, if you are reading this textbook and I am your instructor, these are the norms for my class. One: Dumb questions are allowed because there are things that we take for granted that we really should not. We deserve the right to question everything. Two: Asking to repeat an answer is allowed because I never got it right the first time, and neither will you. Three: Rambling, incoherent questions are allowed because it is probably a lot harder to ask a good question than to simply ask a question rambling away and eventually arrive at a question. Questions are an interactive process; you must ask the question; as your instructor, it is up to me to help you navigate and collaborate to understand your question. Four: Tangents are allowed. We all have experiences connected to statistics, research, and psychology that we obtained from others. We deserve the right to make sense of what we learn here and how it connects to whatever else we learn. Overall, these norms exist for the following reason: We are learning new awkward skills and taking a closer look at what we think we know.

If you are in my class and anyone questions your knowledge level, you say the following to them: "You're an idiot," "I don't trust anybody," "I don't give a rat's ass," "Bite me." We are here to learn, and no one should make you feel less than anyone else. At least you tried to do something, which is more than others who just sit there and think it is okay to criticize you.

The number one rule is the following – call me whenever you have a problem. I follow this quote – "The Purpose of Life Is to Discover Your Gift. The Meaning of Life Is to Give Your Gift Away" - I cannot tell you how many times people gave me free advice, help, or anything when I had a question. Even outside of statistics, I cannot tell you how many times people offered to help me with a home improvement project for free, gave me a ride, or gave me medical or computer advice for free. Each of us has a gift, and someone needs it. So...You can call me and ask questions whenever you'd like.

Even more finally, I use analogies, metaphors, and visuals to aid the learning process. Examples are good places to make sense of statistics. I also use examples to take the edge off learning statistics by injecting humor, lightness, and irreverence to relax the process. I use examples such as cooking, sports, and dating to make my points.

That said, I take strong positions about what I love and hate when I use examples. So, you will read about how the Chicago White Sox are the best baseball team, the Cubs are a minor league team, kittens are better than puppies, homemade brownies are better than box mix brownies, and Taylor Swift belongs to me and not to Travis Kelce.

You've been warned.

This page titled [0.7: Approach to Learning Statistics](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

CHAPTER OVERVIEW

1: What is a Number?

Learning Objectives

By the end of this chapter, you will be able to answer the following questions:

1. What do numbers represent?
2. What are the pros and cons of using numbers versus words?
3. What are the advantages of quantifying a construct?

Key Terms

- Variation
- Validity

Statistics is the study of variation. We notice variation as part of our curiosity. Variations are observations of differences among entities. In the psychology profession, we ask, “what makes this person different from another person?” We use research to create a systematic procedure to gather our observations of how something varies. We use statistics to analyze these variations. To understand statistics, we need first to understand how we generate the numbers that are eventually used by statistics.

[1.1: How to Start with Statistics](#)

[1.2: A "Geek Out" Explanation of Variation](#)

[1.3: Variations Lead to Numbers](#)

[1.4: The Battle of Words Versus Numbers](#)

[1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words](#)

[1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift](#)

[1.7: Statistics Do Something with Numbers](#)

[1.8: Numbers Don't Lie, but Numbers Don't Tell the Full Story](#)

[1.9: Seeing Numbers – What I Want You to Learn](#)

[1.10: Connection Between Understanding How Things Vary and What You are Trying to Do](#)

[1.11: Discussion Questions](#)

This page titled [1: What is a Number?](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

1.1: How to Start with Statistics

Let's start with you. By you, we mean how you observe "it."

We know something is different by observing it. This "it" is another way of saying "construct." A construct is something that we experience as varying. Obvious examples are height, weight, and age. The reason I say those are "obvious" examples is because we can readily observe the physical indicators of the variations in height, weight, and age. Somewhat obvious examples are intelligence and personality. The reason I say those are "somewhat obvious" is because while we readily experience people as more or less intelligent than others, or people having different personalities, there is no agreed upon method of definitively stating how people have different intelligence levels or personality traits. These experiences are called "constructs" because we constructed them. We observe one person as being able to solve a problem better than another person, and we call that experience as "intelligence." We observe one person as having more friends than another person and we call that experience as "extroversion personality." The term "it" throughout this textbook is shorthand for "construct."

It is not always easy to describe how "it," or the construct varies. We first must define what "it" is. When we define something, we do have to agree on what it is that we are observing. Height, weight, and speed seem obvious enough to define. That something has to seem higher than others, heavier than others, or faster than others. Other constructs are not that easy to define, such as depression, attachment, racism, post-traumatic stress disorder, post-traumatic growth. This discussion is beyond this chapter, but suffice to say, we can observe "it" by observing how it varies.

We know something is different based on comparison. All differences, all variations, are based on comparison. In fact, statistics are based on comparison. You are always comparing something to something else. We see that some people are taller than others, some things are heavier than others, and some cars are faster than others. Statistics are based on comparing one person to another person, or one thing to another. Fast forward – when we say something is "statistically significant," we always need to think, "significant, compared to what?"

This process becomes interesting when there are many ways to describe how "it" varies. Let us say we want to quantify the number of beers you drank over a weekend. The number of beers you drank becomes an interesting observation because there are so many ways to define this variation. How many glasses are the same size? Is drinking cheap beer (insert your beer brand here) the same as drinking craft beer (insert your beer brand here)? Is the alcohol content the same across each beer? Whatever we observe, we notice that it varies, and we use statistics as a framework to organize how we observe this variation.

This page titled [1.1: How to Start with Statistics](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

1.2: A "Geek Out" Explanation of Variation

When something varies, we observe that the “it,” or the construct, is different across contexts. What are the ways that the “it” is different? There are only two ways things vary: by type and amount. By type, that means that whatever this “it” is one thing versus another thing. If we identify an “it” as gender (assuming the binary for the moment, and we will acknowledge the continuum of gender identity), we have two types: female and male (and in this world, females are first, most definitely they should be first, 😊). The number of “types” for it does vary according to one’s perspective. Race varies by type (assuming the U.S. government’s definition for the moment), and we have five types: White, Black, Latinx, Asian, and Middle Eastern.

For now, we will assume that “type” means sorting entities into one or another group, and there is no overlap between the groups. Honestly, the number of true “types,” or one thing is exclusively one type versus another, is mainly in the eye of the beholder. Yes, race can be subdivided further into additional types. There are different types of “Whites,” “Blacks,” “Latinos, Latina,” and so forth. Gender is now rightfully thought of as more than two types. Does it matter how many types there are? Yes, and it depends on how you conceptualize your research or curiosity. Save this placeholder for future discussions when we discuss variables.

First, things vary by type; second, they vary by amount. The amount is thought of in two ways. First, the amount can vary by counting or frequency, which ranges from more to less. A given classroom has more males than females. Someone is taller than another person. Someone has more money than another person. Someone drank more beer than another person. What is interesting about counting is that, ostensibly, counting is something that can be readily observed in the physical sense. You can observe more males than females; you can observe that someone is taller than the other; you can observe that someone has more money in their bank account than someone else; and you can observe that someone is drinking more glasses of beer than another person. For now, amounts that vary by counting are based on observing the physical characteristics of something.

And yes, things get interesting when you are trying to count something that is not readily observable or something that not everyone agrees should be counted. Is going one mile over the speed limit an observed count of speeding that warrants a ticket? Is going one degree over your body temperature an observed count that warrants you staying home from class and calling in sick? It gets interesting, so let’s take that ponderance for now.

The second way that an amount can vary is by intensity. What is interesting about intensity is that it is a perceived experience. Sometimes, you observe it directly; other times, you observe a manifestation of it. Take temperature. You can’t readily observe the temperature. However, you can feel it. One day it is hot, the other day it is not so hot. One day, you can observe a gentle, romantic rain; the other day, it is a raging lightning thunderstorm. No one would doubt that there is a difference in the temperature and in the rain because the experience of these “it’s”, or constructs, or phenomenon, varies.

Yes, we should acknowledge here that these issues all get interesting when applied to psychology topics. Take anxiety. Are there different types of anxiety? Generalized anxiety, social anxiety, and performance anxiety. Are there different amounts of anxiety that vary by amount in terms of counts? Number of anxious days in a week, increased heart rate when experiencing an anxious event, number of minutes per panic attack? Are there different amounts of anxiety that vary by amount in terms of intensity? Feeling scared to the point of avoidance, a feeling of anxiety that could be considered common, such as feeling anxious when giving a speech in front of a group, or feeling anxious about your safety at night? What about racism? Are there different types of racism, different amounts of racism that occur, different intensities of racism per event? Statistics and research can sort through these questions, but importantly, the theme of learning statistics and research rests on this premise. Your conceptualization of the issue you are curious about will drive the answer to these questions. The sorting and calibrating of the variation depends on whether the sorting and calibration system serves your purpose. This is another issue we will put in our pockets for now.

This page titled [1.2: A "Geek Out" Explanation of Variation](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

1.3: Variations Lead to Numbers

So now what? What you want to look for in these next sections is to understand the following. Numbers represent variation. Statistics analyzes numbers as a way to analyze variation. So, when you see numbers, see them as data, and statistics analyzes these data. Eventually, statistics is the art of analyzing data to see if there are any patterns in how the data varies. These patterns are going to be useful in understanding the issue that we are curious about.

Why do you give a rat's ass about variation? First of all, that's how we know things matter because of how they vary. Why do some people get cancer and others do not? Why do some people recover from alcoholism and some do not? Why are there less forests in certain parts of the world compared to other parts? Why are some people taller than others? (Curiously, a podcast devoted to that question states that the lack of access to quality health care and healthy practices contributes to people getting shorter, source: [Why are Americans getting shorter? | On Point \(wbur.org\)](#)). Why are some White Sox fans and some are Cubs fans (A: White Sox fans are intelligent, and Cubs fans are well.... less so, boom!). Understanding differences in terms of variations is what the endeavor is all about in research and statistics.

This page titled [1.3: Variations Lead to Numbers](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

1.4: The Battle of Words Versus Numbers

We could just talk about the variation as part of prose and discourse. After all, we readily discuss variation. By type, we say, this person is a male, and that person is a female (again, acknowledging the binary for only this example); this person is in psychology, that person is in medical school; this person is an extrovert, this person is an introvert; this person is a social drinker, that person has a problem with alcohol. By amount, we say this person is taller than that person, this person is faster than that person, that person has more money than that person, and that person drinks more than that person. In most cases, simply stating that one something is more or less than another is sufficient for communicating differences in amount.

You might already anticipate that the preceding discourse will have problems. All discussions of variation depend on perspective, and perspectives are biased. Not that bias is a bad thing, it is part of our flawed human nature. However, we must check our biases, and we don't always do that. As an innocuous example, when one person from Chicago says it is a hot day, another person from the South might say there is nothing to whine about because they have a different experience of that same temperature. Yes, it does become a problem when one person says this child has a behavioral problem, but another person says the child is displaying behaviors that reflect struggles with trauma.

The bias inherent when we use words is that it is difficult to extend the meaning of the conversation beyond those who are having the conversation. When a teacher says to another teacher that one of their students has a behavioral problem, those words are not likely going to be shared by the parent of that child. A behavioral problem to a teacher has a different meaning compared to the parent's meaning of a behavioral problem. Having a number that describes the extent of the behavioral problem is the beginning of some common frame of reference for discussion.

The numbers can guard against when we generalize something. When we say someone is depressed, the word alone doesn't give you the full description of the depression. The devils are in the details. To what extent is the person depressed, how depressed, in what ways does depression show, and is it bad or bad enough that action needs to be taken? Words can describe these details, but we need in-depth explanations. With a set of numbers, we can efficiently summarize these details about the extent of the depression.

This page titled [1.4: The Battle of Words Versus Numbers](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words

In a quantitative study, we are trying to quantify a construct.

"Quantifying" means we are trying to use numbers to describe how a construct varies. When something varies, we mean that we see two people or things possessing two different levels of something. Person A is more depressed than person B. This type of abuse is different than that type of abuse. Variation implies a comparison between at least two entities.

Why Numbers are Better than Words

We could use words to describe how something varies on a continuum. This person is more depressed than that person. This abuse is physical, while that abuse is emotional.

But words can shift in meaning depending on the person – "What do you mean by that?"; "I've seen this self-harm before, it's not that big of a deal vs. No, this self-harm is pretty severe"; "Your impression of racism is more severe than my impression."

Instead of using words, numbers help to communicate the variation. With numbers, we can do so much more. We can communicate by how much more, how many more, how much more frequent, and how much more intense. Instead of simply saying "it" is more than "that," we can quantify the amount of the "more" – "This person is how much more depressed than that person?"; "This person is how much more isolated than that person?"

Using Numbers to be on the Same Page

Communication involves standardization and validity.

Using a standard set of numbers, we are "on the same page," i.e., we agree on what the numbers mean, not just if two things/people are different, but by how much.

This is called standardization – we agree that these numbers tell us what we want to know.

The number represents the variation as we experience it. That number has to represent one entity's level of variation in comparison to another entity's level of variation. We observe a variation, and that number represents that variation, and we agree that the number represents that variation.

Validity means that these numbers are helpful. It means that we agree that these numbers help us predict something else.

Instead of using biased discussion, we use numbers to (as best we can) have a common discussion framework. One person from Chicago says it is a hot day; the person from the South says it is not that hot. The purpose of a number is to represent the level of "it," in this case, temperature. Hence, we are on the same page about the level of "it." Then, we can say separately that the experience of how comfortable the temperature is is independent and varies depending on the geographical region of the person. So, if we say the day's temperature is 85 degrees, we can agree that number represents the temperature. Our experience of the comfort level of the temperature varies according to a separate "it," which is the region of the country for each person.

The previous paragraphs use numbers to represent how "it" varies by level, either by low or high amount or low or high intensity. "It" also varies by type. We also use numbers to represent variation by type, but the numbers are simply codes. Numbers as codes do not carry any inherent value of low to high. Gender (again assuming binary) is an "it" that varies by type: male or female. Females receive a code of "1," and males receive a code of "2." There is no inherent value of more or less. However, statistics processors cannot use words to compute their statistics, so we must use a number to represent the type. So, when you see numbers representing type, it is just code and nothing more. We will return to this discussion when we discuss variable types.

This page titled [1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift

The purpose of the number is to aid in communicating our experience of how “it” varies. We are trying to communicate variation in terms of the amount or level of “it” or the type of “it.” The purpose of a number is to organize how everything varies and give it a common frame of reference to aid in communicating our experience of variation to others.

We love numbers because we need them to give us an understanding of an issue. If you are sick, you go to the doctor and provide three numbers: your temperature, blood pressure, and pulse. These three numbers are the beginning of understanding your health and why you are at the doctor. You go on a date and need three numbers: how much money they make, do they like the White Sox (yes or no), do they like Taylor Swift (yes or no). Those numbers give you a quick impression of your dating experience for the next 10 minutes on that date.

Notice in the preceding examples how parsimonious the set of numbers is. For the doctor, you only need three numbers. For a date, you only need three numbers. One overlooked art of statistics is parsimony, which distills a complex phenomenon into a succinct set of numbers that gives you a complete understanding of the phenomenon. In statistics and research, parsimony is what you are striving for. This goal starkly contrasts most research and statistical situations I have encountered where researchers want to make variables and analyses of everything. This situation is the “everything, including the kitchen sink” approach, which is not good. It is grasping for straws, and it wastes time. Always strive to keep it simple.

This page titled [1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

1.7: Statistics Do Something with Numbers

Part of the problem is that too many numbers are too overwhelming. The other issue is that numbers aren't exactly bias-free, and we will discuss this in the next section.

There is such a thing as too many numbers. We need to organize our experiences. It follows that we need to organize our numbers. In statistics and life, every time we get overwhelmed with too much "stuff," we seek to simplify.

We want to understand the temperature of a region, so we use numbers to gather our observations of the temperature, which are based on our experience of "it." But there are so many numbers generated from different contexts. We take temperature readings in the morning, then in the afternoon, and then in the evening. Then, we take temperature readings when it is raining, snowing, or windy. The temperature readings vary by context. And then there are more contexts. In Chicago, the temperature in the north will differ from the south and in the west, and if you are genuinely from Chicago, you know that temperatures by Lake Michigan are different compared to inland.

All those temperature readings are gathered, and we call them data and put them into a dataset. A dataset is nothing more than a series of numbers based on our observations of variation. When we say "gather the data," we really mean gather our observations of the "it" in as many ways as possible. By the way, those "ways" are considered the research method, which basically refers to "how you are gathering the data." As you gathered, we will discuss that issue in more detail.

So, what do we do with this data? We use statistics to do something with the data to achieve a goal. What is that something? There are several somethings: organization, summarize, estimate, predict, and infer. What is that goal? The goal is always this— to understand something.

First, we organize the data, otherwise known as our observations of the variations of "it." Organization means making sense of something, and using a framework helps. The easiest and simplest framework is to sort something from low to high. We can organize our temperature data from low to high. Sorting from low to high is an intuitive way to manage our experience because we naturally experience "it" from low to high.

Second, we want to take inventory of our "types." We have temperature as one variable. But if you notice in our temperature example, we have other variables. We have regions: north, south, and west (notice there is no "east side" in Chicago). We have context: rain, snow, wind. We have a unique context: inland or near Lake Michigan in Chicago," (Google "The Lake and temperature in Chicago" and you will understand the reference). We want to inventory our variables and see how they vary by amount or type.

Third, we want to summarize. No one can process the meaning and the usefulness of a hundred temperature readings. It is efficient to summarize the readings into just one number. We can do much more with just one number. What do we mean by "do much more?"

Fourth, we mean we want to use the numbers for something. That something is usually to make predictions and to understand something.

These four uses for numbers are why we use statistics. These uses are our investigations that form the cornerstone of building our knowledge about what is meaningful.

This page titled [1.7: Statistics Do Something with Numbers](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

1.8: Numbers Don't Lie, but Numbers Don't Tell the Full Story

Let us summarize. Numbers have problems. We think that numbers are as is. Numbers aren't clean. Numbers are not immutable. A statistician's axiom is "data are messy."

Numbers are "man-made" (assuming humankind for now, not male dominance). They were created. Even the number "0" had to be created because we realized that we had to have something to represent the absence of value.

Because numbers are man-made, we need to acknowledge that the *generation* of numbers is based on human perspectives that are subject to bias and error. The carpenter's credo is – "measure twice, cut once." There is an error in how the measurement is made based on the starting point and the variations in how the line for cutting is placed. I attended a lecture by an economist who bewilderedly stated that historical records of state budgets are modified, and only after 50 years are they considered set. Numbers change depending on the point of view.

So, numbers can change. In statistics terms, we call that error. We will discuss the concept of error more later, but for now, we must appreciate how people think we have data that we are set to go. We still must examine the data and determine if we agree on what it can do for us.

Numbers do not tell the full story. It is hard to reduce all our experiences into one number. Everything is multi-dimensional; everything can be subdivided and parceled out into different experiences. Someone who is a chef knows there are many dimensions to being a chef: baking, grilling, and plating. Everyone has different experiences about what they think "it" is. Racism means different things to each racial group, and within that racial group, there are variations in what racism means. One number cannot contain all the information about "it" that we need to match our experience of "it."

Qualitative research is meant to explore people's experience of "it." That is one way to approach understanding the meaning behind the numbers. You can use qualitative research instead of quantitative statistics to understand something. Suffice to say that qualitative research and quantitative research are not opposites of each other. They do not show limitations that are countered by the advantages of the other. It is a mistake to pit the two approaches against each other. The two approaches are just two separate ways of exploring experiences.

This page titled [1.8: Numbers Don't Lie, but Numbers Don't Tell the Full Story](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

1.9: Seeing Numbers – What I Want You to Learn

I want you to “see” numbers. What are numbers?

1. Numbers represent variation.
 2. We understand the world by understanding variation.
 3. The variation comes from our observation of “it.”
 4. When observing, we are observing how “it” varies.
 5. Variation is simply “this is different than that.”
 6. What is interesting about variation is it involves comparison. We know something is different by comparing “it” to another “it.”
 7. Everything varies in two ways.
 - a. By type and by amount.
 - b. Amount can be as a frequency count.
 - c. Amount can be as an intensity.
 - d. Both frequency count and intensity are thought of as from low to high.
 8. Our goals are to understand the variation by
 - a. Describing the variation.
 - b. Using the variation to predict something about why “it” varies.
 9. Statistics is how we achieve the goal of understanding variation.
-

This page titled [1.9: Seeing Numbers – What I Want You to Learn](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

1.10: Connection Between Understanding How Things Vary and What You are Trying to Do

Trying to Do

Why is it important to understand that numbers come from our observation of variations of “it” —whatever “it” is you are trying to do?

Because when we design a quantitative research study, we must be aware of what we are studying, how it varies, and what the variations mean as being high, low, or in the middle. When we are trying to understand a quantitative study or statistics, understanding variation is our framework for understanding how to read statistics and the conclusions drawn from statistical analysis.

Quantifying an Issue

There are many ways to start designing a research study. One possibility is to understand the process of quantifying an issue. Where to start? Start with your words and how you want to describe your experience of observing “it.”

- Pain – rate your headache pain. Use words. Then, describe it on a scale from 1 to 10.
- Depression. Describe the depression. Use words. Then, describe it on a scale from 1 to 10.

The ongoing challenge is to “translate” an issue into something quantifiable that communicates how the issue varies and how to organize (i.e. scale) that variation using numbers. How do you quantify these issues: racism, historical trauma, internalized homophobia, and attachment?

An issue varies by **type** and **level**.

Let us start with “type.” Are there different types of this issue?

Next, within each type, there is a level. Someone must experience more frequency or intensity (severity) of the issue than another person. What does that look like?

Third, let us consider ensuring that these levels communicate differences, which means we are comparing something to something else. Someone has more of “it” than another person.

The Axioms

So, we use the following axiom: Numbers = variation → Establish Comparison of the differences in “it.” This is the schematic process of going from observations to data:

Observation → variation → number → collection of numbers → data

Further Reading

To know more about this issue, here is an additional reading: Abelson, R. P. (1995). *Statistics as a principled argument*. Lawrence Erlbaum Associates, Inc.

This page titled [1.10: Connection Between Understanding How Things Vary and What You are Trying to Do](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

1.11: Discussion Questions

Discussion Questions

1. Answer the questions in the learning objectives.
 2. If someone asked you, “Should I use numbers or words to assess a construct, such as depression or racism,” what would you tell them?
 3. Write a response to this question, “How do I quantify racism?”
-

This page titled [1.11: Discussion Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

CHAPTER OVERVIEW

2: What are Statistics?

Learning Objectives

By the end of this chapter, you will be able to answer the following questions:

1. What does it mean when we say, “statistics uses comparisons?”
2. Explain “estimation” in statistics.

Key Terms

- Descriptive statistics
- Inferential statistics
- Population parameter
- Sample statistic

Statistics is the study of variation. But what processes are involved when using statistics? The process of studying variation involves the following: organizing the variation and describing it, comparing the variation, looking for patterns in the variation, and making estimates, otherwise known as “guesses.”

We have already briefly examined how to organize variation and describe it. We determined what the “its” are, otherwise known as variables, and understood the variation by examining the “its” by type and by amount.

[2.1: Statistics Involves Comparison and Context](#)

[2.2: Statistics is About Finding Patterns](#)

[2.3: Statistics as Estimates, or Making Guesses](#)

[2.4: Two Broad Categories of the Process of Using Statistics](#)

[2.5: Discussion Questions](#)

This page titled [2: What are Statistics?](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

2.1: Statistics Involves Comparison and Context

Statistics help us to understand variation by using comparison. Variation helps us to set up a comparison – “This is different than that.” Comparison is basically this question: “This is different compared to what?” We use variation all the time – “This person is taller than this person”; “This person is more depressed than that person”; “This person has severe ADHD compared to that person.” To show that comparison, we need a context. The context, in this case, is comparing something to something.

For example, suppose we want to study eating disorders among different ethnic groups. We might find that Asian American women struggle more with eating disorders compared to Caucasians. So, the variable is eating disorders; the level is more struggle or less struggle with an eating disorder. We find that Asian American women struggle more, but more compared to what? That context is the racial group. Asian American women struggle more compared to Caucasians. What is the basis for that context? Alternatively, what is the conceptual basis for this comparison? Suppose we say the difference we are noticing between Asian Americans and Caucasians is due to cultural values. In Asian American culture, there is more cultural emphasis on thinness and more family criticism of weight gain compared to Caucasian culture.

Why am I making a big deal about this? Because to understand statistics, you must understand the numbers by understanding the context in which those numbers are generated. This context consists of the comparison, which is comparing what to what, and the conceptual basis for that comparison, which is why we are comparing this to that. Students think that the number exists by itself. That is not true. You cannot interpret a number and its accompanying statistical values without understanding the context of the number. The context always involves a comparison and the conceptualization basis for why these numbers vary.

A slight fast forward – the context is necessary to show statistical significance. This finding is significant but compared to what and under what context. Consider these statements -

“This is significantly different from that because the numbers show that this issue is more severe than usual.”

“This person’s depression is more severe than what we typically encounter in this counseling center. May need to hospitalize this person.”

Without the numbers, we must rely on verbal descriptions of what entities are different from each other.

We will discuss statistical significance in greater detail. But for now, when you think about statistics, always think in terms of this statement: “What is being compared to what?”

Knowing that statistics involve comparison is important because all statistics involve ratios. Go back to your math classes and recall that a ratio involves a numerator over a denominator. That ratio involves a comparison – a numerator compared to the denominator.

In statistics, these ratios have many forms. The most common form is the following: Actual effects vs. random effects. Actual effect is what happened, or your result, your outcome, versus a random effect, which is a random chance.

Another example consists of changes in scores versus error shifts. An actual change in score can be a depression score before treatment starts compared to a depression score at the end of treatment. This change is compared to error shifts, such as random fluctuations in depression scores that occur over time.

Another example is this group versus that group. A treatment group’s mean score on a depression scale is compared to a control group’s mean score on a depression scale.

In all of these examples, we are comparing variations that are due to what actually happened to variations that are due to random chance. In most cases, we are hoping to find that there are true variations rather than chance variations. We want to see that a decrease in depression from before treatment to after treatment is due to depression actually improving. We want to see that the patients in the treatment group are doing better than the control group because the treatment actually worked, and random chance has nothing to do with the variations in the treatment effects. We are hoping to find true variations rather than chance variations.

This page titled [2.1: Statistics Involves Comparison and Context](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

2.2: Statistics is About Finding Patterns

Data can be presented as an array of numbers that looks like a mess. We organize it and describe it to find some order within those numbers. The next step is to find any patterns within those numbers. Patterns help us figure out how to make sense of the variations present in the data. With patterns, we can determine what is going on with the variations we see in the data. The opposite of patterns is random. Random means we cannot tell what is going on with the variations we see in the data.

Statistics help us find patterns in the data. These patterns are between the variables. In statistics, you want your actual results to be a clear pattern, compared to random, error, chance, or no pattern.

What is a "pattern?" Statistical tests produce an outcome, a value that tells you if something is going on. That's called a pattern. A pattern is something discernable and predictable that occurs between two variables. Something is going on systematically between the two variables. By systematic, we mean that there is a relationship between two variables that we can easily describe. Systemic variation is often referred to as a model or a prediction. When you hear the terms "the model fits" or "we want to model this," what we are saying is that the variables we selected decide, or predict, a relationship, or a pattern, among the variables.

We need patterns; we need predictions because they give us a framework. If we know that the more times a person attends therapy sessions, the more likely they will alleviate their mental health symptoms, then we can say something about the importance of attending therapy regularly and its impact on symptom reduction. Without a framework, we are useless and ineffective because we can't affect the outcome.

What is the opposite of a pattern? The opposite pattern is random. Random means you can't tell what is going on; you can't predict anything. Whatever happened occurred by luck. Randomness stinks. If we study a lot and get a good grade, then study a lot, then not get a good grade, then not study a lot and get a good grade, and then not study a lot and not get a good grade, then there is randomness, and we can't tell what is going on, at least between these two variables. The implication in this example is that the number of hours studying does not make a difference in the outcome. This result is instructive, though. It means that studying a lot is not the key to obtaining good grades. Something else is going on that leads to good grades.

Another example is alcohol treatment. Pattern: "We know if you are motivated, plus have good friendship support, you will get sober faster, and stay sober longer, compared to someone who doesn't have those qualities. If you don't have those qualities, we need to alter our treatment approach." Random: "Good luck getting sober; we have no idea what will make you sober or not."

When you have a pattern, you are looking for patterns between at least two variables. There are two types of patterns: relationships and group comparison. More on this later on in the chapters that discuss correlations, t-tests, and F-tests. But for now, the relationship pattern is this form: increases in Variable X are associated with increases or decreases in Variable Y. Group patterns take on this form: Group A is lower or higher than Group B.

The statistical value is a value that indicates if a pattern is present or not present. All statistical tests generate a value and that value is a ratio. Ratios compare something to something else. In this case, the statistical value is a ratio that represents the pattern among the variables versus randomness among the variables. Ratios are numerators over denominators. The numerator is the actual difference we observe; the denominator is error variance. If the statistics value is 0, it means that there is no pattern because the numerator is much lower than the denominator, which means there are no actual differences and there is only error variance. If the statistics value is 1, that is still not great because it means that there is no pattern because the numerator is the same as the denominator, which means the actual differences are the same as the error variance. The likelihood of a pattern in the relationships among the variables is equal to the likelihood that there is randomness, or no patterns, among the variables. The higher the statistics value, the greater the numerator is than the denominator, which means that the likelihood of a pattern in the relationships among the variables is greater than the likelihood of randomness among the variables. Consequently, the lower the statistical value, the lower the numerator is than the denominator, and there is more randomness than pattern.

There is only one decision with a statistic test – am I seeing a pattern vs. am I seeing randomness? More on why that is the only decision you make with a statistics test in the discussion of the concept of significance in [Chapter Eight](#).

This page titled [2.2: Statistics is About Finding Patterns](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

2.3: Statistics as Estimates, or Making Guesses

A hallmark of statistics is estimation. Estimations are guesses. What are we guessing at? We are guessing the truth.

The goal of statistics is about how we get to the truth. How do we know that these conclusions about patterns of variation among our variables are the truth? When we say something is true, in statistics and in research, we are guessing that what we found is true for everyone, everywhere.

When we say that studying more leads to good grades, we want that pattern to be true for everyone and across all contexts. When we say that attending more therapy sessions leads to better outcomes, we want that pattern to be true for everyone and across all contexts.

The art of making a guess involves understanding the relationship between Sample Statistics and Population Parameters. And these are the first two “official” statistics terms that you want to know (insert your celebration emoji here).

Your first official statistical term is the population parameter. The population parameter is the “truth” or “This is what’s really going on.” Describing the parameter is our goal.

The problem is that we can’t know the truth. It is impossible to know what is really going on. There are logistical reasons for this. The truth means everything, everywhere, everyone. But we can’t take measures of every variable for everyone and everywhere. We do not have large-scale testing of most phenomena for the population. We do not know who is included in the population of interest. For example, we can do nationwide testing of standardized test scores for all students because, in a sense, we have a roster of all students registered in public high schools. However, we don’t have a “roster” of all high school students who have an eating disorder. There is no such thing as a “roster” for this issue, and the logistics are difficult for determining who has an eating disorder and who does not. We do not have standardized measures of psychological phenomena. Although our measurement of eating disorders has improved, debates and discussions still ensue about how to accurately diagnose eating disorders. So, we cannot be sure we have the correct level of eating disorders among high school students. We do not know if we have the correct prevalence rate and level of severity of this issue among high school students. We cannot simply ask, “Who has an eating disorder?” and expect anyone with an eating disorder to come forward. The stigma accompanying eating disorders means students will not readily come forward and identify themselves as having an eating disorder. Therefore, we cannot obtain a population estimate because it is likely that our sample estimate is low or that there is an underreporting of the prevalence of this issue. For issues such as racial identity, career outlook, and risk of dropping out of school, these issues are conceptualized in many ways, and, rightfully so, there is no standard for how these issues should be assessed. With logistical concerns, determining a roster census, and no agreed-upon standard of assessment of a construct, it is difficult to know the “truth” of the nature of an issue.

So, if we cannot know the truth of an issue, the best we can do is to make a guess at the “truth.” This procedure is considered to be an estimate in statistics. When you hear people say we are estimating a parameter, what they are basically saying is we are making a “guess” about the truth about something. This estimation involves using sample statistics, which is your second official statistical term. Sample statistics is the term that describes the process of making guesses about what we find in our corner of the world, which is considered the sample, and this guess is true about the population, which is everyone or everything else in the world. The reason it is called “sample” is that all we can do is take a piece of the puzzle, or a “sample,” and hope that what we found in the sample is true for everyone else. That hope is an estimation or a “guess,” and it is our estimate that what we found is true for everyone.

The trick is making the best guess possible by using a good research design that convinces others to ascertain that we did everything we could to make the best guess possible. This issue will be revisited throughout this textbook.

Both population parameter and sample statistics are conceptual terms. They do not refer to statistical tests or terminology. They are terms to describe the relationship between what we found as a guess about what the truth is.

So, to recap, this is the process of getting to the truth:

1. The concepts you want to know are population parameters and sample statistic
2. Start here: Parameter – a number that represents the value (level) of a variable that describes what is going on with this issue in the population. Think: Parameter is theoretical: We think this issue exists out there. This is the “truth” about an issue. This is what we want to know.

3. Statistic – a number that represents the value (level) of a variable that we use to estimate (or guess) the level of the variable in the population. Think: locally, right here, right now, my corner.
4. The process: I think my results are the truth that could apply to everyone or the population.

This is the process of statistics.

This page titled [2.3: Statistics as Estimates, or Making Guesses](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

2.4: Two Broad Categories of the Process of Using Statistics

Using statistics involves two broad categories of statistics. The first is descriptive statistics. The second is inferential statistics. These two categories are conducted in the order of descriptive statistics first and inferential statistics second.

Descriptive Statistics – Start Here

First, use descriptive statistics. Our first step is to simply describe what we got in our data. Think of this first step as the first step of cooking. You take stock of your ingredients and how much you have of each ingredient. In statistics, you want to take stock of what variables you have. For each variable, take inventory of how that variable varies by type or by amount. Then, you decide what you know about how much of each variable you have. By how much, we mean the level, the scores or the numbers that represent the level of each variable.

The descriptive statistics usually consist of three values. They are a) your frequency counts, b) your measures of central tendency, otherwise known as the mean, and c) your measures of the spread of score, otherwise known as the standard deviation. These descriptive statistics help you organize your data into your variable sets and describe how much, or the level, you have of each variable.

Conducting descriptive statistics is ALWAYS the FIRST STEP in conducting statistical analysis. Resist the temptation to dive right into your statistical analysis to answer your hypotheses. You first want to take stock of what you have in your variables.

Inferential Statistics – Next Step

Second, use inferential statistics. I am going to infer (estimate, guess) that if I look at those two variables, I can see a relationship between self-esteem and depression. We infer the truth about everything by making a guess about what we have in front of us. We infer a parameter by estimating a relationship based on our sample.

Inferential statistics are your classic statistics analysis. These analyses are your t-tests, ANOVA, correlation, regression, and Chi-square. These are the common statistical analyses that answer most questions we have in statistics.

This page titled [2.4: Two Broad Categories of the Process of Using Statistics](#) is shared under a [CC BY](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

2.5: Discussion Questions

Discussion Questions

Consider your psychology research topic:

1. Explain this concept – statistics involves comparisons and conceptualization.
2. Explain the idea of a ratio in statistics.
3. What patterns do you hope to find when you conduct your research study?
4. What does it mean to estimate something in statistics?
5. How do we estimate something in statistics?

This page titled [2.5: Discussion Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

CHAPTER OVERVIEW

3: Variable Types

Learning Objectives

- Distinguish the difference between the terms variable and variance
- Understand how to determine if a variable is dependent or independent
- Understand the process for selecting a statistical test

Key Terms

- Variance
- Categorical Variables
- Continuous Variables

Statistics is the process of studying variation. What does that process look like? It starts with variables. Variables are terms we use to represent the different types of variation we observe.

[3.1: Variables - Not Always Obvious](#)

[3.2: How to Organize Variation With Variables](#)

[3.3: Categorical Variables- Variables That Vary by Type](#)

[3.4: Continuous Variables- Variables That Vary by Level](#)

[3.5: Nuances in Determining if a Variable Should be a Certain Type](#)

[3.6: Making Decisions About the Variables in Your Research Study](#)

[3.7: Lessons Learned About Variables](#)

[3.8: Discussion Questions](#)

This page titled [3: Variable Types](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

3.1: Variables - Not Always Obvious

When we want to study variation, we must organize it. We use variables as the mechanism to organize the variation in the construct we are observing. If someone dumped a pile of ingredients on the table and asked you to make something out of it, it would be impossible because the pile is a mess. Suppose the dumped pile consisted of sugar, flour, powdered sugar, brown sugar, confectioner's sugar, cake flour, wheat flour, milk, cream, bread flour, sour cream, eggs, salt, baking powder, baking soda, red food coloring, and it was all mixed up. In that case, it is obvious that no one would dare touch that pile of ingredients. Among the apparent reasons this scenario is impossible to work with is because you have no idea which ingredient is which and how each ingredient will be used to create the final product, whatever that could turn out to be.

Variables are like containers that sort ingredients. We have sugar and flour; we figure out what type of sugar and flour we have, then sort the sugar and flour into containers. If we have powdered sugar, regular sugar, and brown sugar, then we need three containers, and if we are meticulously organized, we put those three containers on a shelf away from the flour. If we have regular flour, bread flour, and wheat flour, then we need three containers, and then put those three containers on separate shelves, away from the sugar. This process of sorting and keeping the ingredients separate from each other is the process of sorting variables by type. The initial process of sorting by type starts with your definition of each ingredient. We define flour and sugar as different, then sort them according to their differences.

It is important to note that the first step is sorting. You cannot mix up the ingredients. To state the obvious, each ingredient serves a different purpose in creating the outcome. No one would argue that sugar and flour are not the same ingredients and cannot be put in the same container. However, you might notice that the powdered sugar and the confectioner's sugar look the same and put them into the same container. Is that wrong to do? I will address this question when we discuss variable types.

Once we sort the ingredients, we decide how much sugar and flour we have. We need to know how much each ingredient is in each container. This is in contrast to the first step, which is sorting by type or by definition. The second step in this process is measurement. We want to know how much of each ingredient is in each container.

So, creating a variable involves two processes. First, observe a pile of stuff to sort; next, sort that stuff into types, which is defined by how we define each type, and then measure how much of that type we have. Using this process, we find we have four cups of regular flour, two cups of sugar, one cup of milk, one egg, and one teaspoon each of salt and baking powder, which is enough to make a waffle, at least the way I make it.

The example sounds silly, but it becomes interesting when you think about psychological variables that overlap in content and are hard to measure. It is not always obvious what variable best represents a concept. Gender used to be considered a male/female-only variable. Gender was thought of as a binary concept. But society evolves and understanding of everything evolves. We now know that gender is not quite a binary concept, especially when you consider the continuum of biological variation, and the expression of gender identity. Focusing on gender identity for now, we expand our idea of gender from a binary male-female to male, female, transgender male to female, transgender female to male. You can rightfully argue if these are the only four categories to represent the continuum of gender identity and whether someone only fits into just one of those categories. This example illustrates the difficulty of selecting a variable to present our observations of a concept, such as gender.

The difficulty compounds when sorting through a pile of observations and determining how many variables should be used to represent the pile of observations when observations overlap in content. Self-harm and suicide behaviors are considered two separate issues, but sorting through the items to determine which behaviors indicate self-harm or suicide can be difficult. We then need to decide if there should be two or one variables for this pile of observations. Abuse types become a challenge when you think about distinct types of abuse and the overlap between them. Emotional abuse, physical abuse, and sexual abuse likely overlap in terms of the survivor's experience. Some researchers separate these issues; some collapse them into a single score. We then need to decide if there should be three variables, or an overall variable. Most psychological assessments have a total score, followed by subscale scores. So, you might have one variable or several variables, but then those variables could overlap in terms of shared items or associations. It is not always obvious what variables we should use to sort through our observations of a construct.

Does this problem matter? Possibly. It matters if you think the observations should be merged or separated because it will adversely affect the outcome or is unlikely to affect the outcome. Imagine if you mixed up your types of flour. You would not know if you had regular or bread flour in your measuring cup. To most, that might not matter, especially if you are me and making waffles the

way I make waffles. The end product will be a waffle, and I highly doubt that my son will notice the difference in taste and texture. However, it definitely matters for an experienced chef because mixing those two flours together affects the outcome. The professional chef will know what to do with those two types of flour and how they affect the outcome, which is the waffle. Will my son notice it, Will having a delicious waffle or my waffle affect the outcome of his day? Let us hope not.

This issue occurs in psychology research. We want to be clear about which variable is affecting what outcome. We will revisit this issue later. But for now, it should be noted that while some variables seem evident in terms of type and measurement, such as gender, age, or depression, it is not always obvious what variables we have on hand for our statistical analysis and how these variables will affect our understanding of the outcome we want.

3.1.1: Variance – a Fancy Term

Each variable has a variance. Variance is the term we use to describe variation. Let us take a moment to connect the concepts of variation and numbers. Statistics and its numbers can seem so confusing and bewildering because we do not know how to interpret statistics. Statistics is nothing more than numbers. There are many ways we use numbers in statistics. For now, to describe variance, all a number does is represent our perception of the variation of “it,” or the construct. A variable is simply a way to organize the variation of a construct. Recall that all constructs vary. We use numbers to describe this variation. A variable will have an array of numbers and all that array does is organize the numbers. Person A will have a value of “10,” and is greater than person B with a value of “9.” When the construct varies, the numbers that represent our perception of the variation vary. In this case, person A is greater than person B on whatever construct that the variable represents. In the spirit of de-mystifying statistical datasets, at the end of the day, a variable is simply a collection of numbers that represent which persons have numbers that represent greater or lesser amounts of the construct.

The opposite of a variable is a constant. Generally, true constants are rare. It bears mentioning that if a variable does not vary much at all, it functions as a constant. Constants really do not help us understand variation. We need variation, so you need variables that do vary.

The scenario of having variables that are essentially constants can happen. Suppose you want to examine male and female mental health experiences in the police force. There are more males than females in a given police district. But not just more males; it is likely that there are two or three females compared to 40 males within a given district. With only three females, it might as well be treated as a constant because there is limited variation among the females, thereby making the comparison between males and females in the police department moot. What to do? It depends on your conceptualization of the problem, but you will likely need to extend your recruitment of females if having a male-female comparison is important for your research study.

Back to variance. Variance is a way to communicate how much that variable varies and what the quality of the variation is. In general, you want a good range of variation or more variance. More variance means more variation in the variable. More variation means we can detect differences among entities. We can see how someone is more depressed, less depressed, clinically depressed, or mildly depressed. We want to know what type of variation we are examining. For example, is the depression a chronic type, or is the depression more as an expected response to a situation, such as a death in the family? We want to know the quality of that variation. For a given group of, say, clients attending a counselling center, is the variation all clumped together with everyone having low depression and only a few clients with elevated levels of depression? We will address how to answer these questions when we examine central tendency and variability measures, including means, standard deviations, and distributions.

When using statistics, you will encounter the idea of partitioning the variance. What that means is that we want to divide the variability into types. For now, we are usually interested in two kinds of variability – true variance and error variance. True variance is a variation that does vary because of actual changes in the variable. Quite simply, age changes because there are differences in age among people. Error variance is anything that varies that is not due to the actual changes in the variable. Notice that error variance is not always just an error. There could be something else going on. Things can vary simply by accident. A person can mistype their age into a form. Or someone can round up their age; instead of saying they are 20 years and 11 months, they may round up their age to 21.

For now, we introduce the concept of variance to describe variation so we can turn to how we use variables to describe the variance.

3.1.2: Variables – Way Too Much to Know about Variables

All research is based on variables. Variables represent the variation that we see. Any issue, phenomenon, or construct that varies is a variable. We relate variables to each other so that we can say how phenomena are related to each other. For example, we want to know if the time variable is associated with the variable emotional development. A research question could be: As youth mature over time, do their emotions become sophisticated? Another example is that we want to know if the number of sessions attended is associated with successful therapy outcomes. A research question could be: As clients attend more sessions, does their mental health improve?

A variable is a way to describe a phenomenon's characteristics, which have different values. These values are essentially numbers, and we assign numbers to represent our observations of the level by amount, intensity, or type of a phenomenon.

Any issue, construct, or phenomenon can be represented with variables. Depression is a variable. The number of hours spent studying is a variable. Typically, we encounter issues that have more than one variable—for example, drug use. We can think of variations in drug use by drug type, how often it is used, and how long a person has used it. Or we think of an issue and think of various kinds within that issue. For example, attachment has distinctive styles: Secure, Anxious, Ambivalent, and Undefined. Or we can take an issue and subdivide it. For example, intelligence can be considered a general construct, or we can subdivide it into related types, such as Verbal, Math, and Reasoning. We think there are distinct types of intelligence, such as emotional or streetwise intelligence.

3.1.3: Variables – Why you Give a Rat's Ass....

Knowing how a variable varies helps you organize your understanding of “it”, or the construct, and the variation of the construct. Knowing this information enables you to understand the following:

- Deciding what statistical analysis to use.
- Knowing how to organize and conceptualize your own research question and design in advance.
- Knowing how the variables can be associated with each other. Statistical analysis aims to understand the variation by learning how “it” varies and how other “it’s” vary in association with it. Put differently, we want to know in what ways people are different from each other and if something can help us understand why people are different from each other.

This page titled [3.1: Variables - Not Always Obvious](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

3.2: How to Organize Variation With Variables

There are two broad decisions you need to make when considering your variables for a research study or when understanding the variables you have for your statistical analysis:

1. Identify your independent variables and dependent variables
2. How do we scale the variables: categorical or continuous?
 - a. If continuous: ordinal, interval, or ratio.
 - b. If ordinal or interval, decide if the variable is scaled by amount or intensity.

First Decision: Identify your Independent and Dependent Variables

Statistical analyses are based on independent and dependent variables. Let us start with the independent variable (IV). A basic definition of the IV is the variable you change, or one that changes independently. Other terms are used for the independent variable, including input variable and predictor variables. Some examples of IVs are treatment conditions with two types: control and treatment. Gender has two kinds: male and female (assuming a binary). Time can be variable, with a range of more or less time. Self-esteem can be a variable that ranges from low to high self-esteem.

Dependent variables (DV) vary based on the variations in the independent variable. Other terms for the dependent variable include the outcome variable and criterion variable. Some examples are treatment conditions with two types: substance use recovery, two types no relapse, relapse. PTSD level, none too severe.

It is not always clear what the independent variable and the dependent variable are. IVs and DVs can sometimes be interchangeable. What leads to what? How can you tell the difference between IV and DV?

The answer is always to think conceptually. Your conceptualization addresses this question - how are these variables put together according to your intentions? Conceptualization is synonymous with theory, models, frameworks, or orientation. Briefly speaking, conceptualization is how you make sense of all the information about how the world works. We have so much information to put together, and we need a guide on how to accomplish that task. There are many conceptualizations. In psychology, we have theoretical orientations. If we use a cognitive-behavioral orientation, we focus on cognitive thoughts and behaviors, and we use interventions to amend those cognitive thoughts and change those behaviors. If we use psychodynamic orientation, we focus on past experiences and see how those experiences shape our present experiences. For statistics, use your conceptualization about how the variable relates to the other variable and use your knowledge of the temporal order of the variables, or which variable change occurs first, and which variable change occurs afterwards.

A classic example is the research design, where there is a treatment and an outcome. Patients are put into a treatment group, consisting of medication, versus a control group with no treatment, and the result is less stress. In this example, the IV is the treatment condition; the DV is stress. The treatment precedes the outcome because logically and in temporal order, the treatment precedes the stress reduction. IVs and DVs are usually easier to spot in experimental designs.

Descriptive designs can make variable identification interesting. In descriptive designs, the intent is to describe relationships among variables. There is nothing manipulated as in an experimental design. Usually, these variables occur concurrently or at the same time, and it is not entirely clear which one precedes the other. For example, depression and substance abuse are associated. But which one is the IV, and which is the DV? Depression could be the IV because when people get depressed, they abuse substances, which is the outcome and hence the DV. Substance abuse could be the IV because when people abuse substances, the substance makes them depressed, or they feel depressed because of abusing substances. Substance abuse is the IV, and depression is the outcome, hence the DV. Both scenarios are correct. So, how do we decide which is the IV and which is the DV?

The answer is to use your conceptualization of the issue. In this case, we will use the word “model.” Using a risk factor model, you might be interested in what risk factors are related to your DV. So, depression might be the IV because you want to know if depression is a risk factor that leads to people using inappropriate coping skills, such as alcohol. Using a drug reaction model, you might be interested in the effects of alcohol. So, depression might be the DV because you want to know if alcohol drinking, the type and amount of alcohol drinking, leads to depression as an effect of drinking. The guiding principle here is that variables do not just present themselves automatically as IVs and DVs. Once you have conceptualized the issue, you can decide which “it” is the IV or the DV.

Guidelines help you decide which variables are the IV and the DV. Temporal order is one way to consider which variable is the IV or DV. Temporal order means which variable occurs first and which occurs second. The variable that comes first can be viewed as the IV, and the variable that comes second is the DV. An obvious example is treatment. Treatment comes first, then the outcome. So, treatment is the IV, and the outcome is the DV. Students' study time comes first. Hence, the IV and grades are the outcome, hence the DV.

Anything that precedes something is usually an IV. For example, personality is usually a trait that occurs before the outcome. Personality usually precedes career choice. Extroverted personalities tend to choose careers involving social interactions. In psychology, the quality of the therapy alliance usually precedes the client's outcome.

Associations between variables can be concurrent. Concurrent usually means happening at the same time. It can mean occurring together. For example, depression and alcohol might be occurring together, not necessarily in a sequence. Using a symptom model, you might be interested in how depression and alcohol co-occur because of something else. In a grieving context, depression and alcohol are both DV's resulting from how a person might be coping with grief. If something like this scenario occurs, it may not matter which variable is the IV or DV.

Anything "fixed" does not typically change on its own; it is usually an IV. Demographics are usually "fixed," meaning they occur first and then the outcome. Gender, socio-economic status, and years of education are usually demographics and are IVs because they occur first or are "fixed." These variables are generally not the outcome of something; hence, they are usually IVs.

Anything that interests us as a goal or an outcome of interest is usually the DV. Symptom reduction, better relationships, academic outcomes, and abstinence from substances are goals that clients want and are usually the DV. These variables are usually the outcome of something that the client is doing. What is the client doing that leads to symptom reduction, better relationships, better academic outcomes, and abstaining from substance use?

The preceding examples are meant to be intuitive. Under a gender identity fluid framework, someone will always argue that gender is an outcome of biological and socio-cultural influences. That is true. To provide typical examples and illustrate what the IV and DV are, the examples are meant to be intuitive and not meant for people to engage in extensive discussion about what predictors are and what outcomes are.

Second Decision: How do you scale the variables: categorical or continuous? If continuous, ordinal, interval, or ratio?

This step addresses the question of how the variable varies. The variable type is otherwise known as a scale of measurement. What is the scale that we use to measure the variable's variation? This process is essential because we need a way to sort our observations. By sorting, we mean classifying, categorizing, and separating. If we do not sort, our observations become a big mush of variation. Whether we are observing temperature, gender, height, depression, alcohol use, racial discrimination, or suicide risk, we are sorting our observations. We need a framework to guide our sorting process.

The framework starts here - remember there are only two types of variation: by type and by level. Everything we observe varies in only those two ways. Things are different in terms of the type of "it" or the level of "it." Type is always expressed as "this" or "that." And then by level, we usually mean by an amount of something, usually by a frequency count, or by the intensity of something, usually by severity. The level is always expressed as more or less of something.

When creating these variables, you first want to identify the pile of observations you have about something. Let us take a broad question – does therapy work? That question involves a messy pile of observations: what is the therapy, what type of therapy, how long is the therapy, who provided the therapy, what is the mental health problem, who are the clients, what is the outcome, what are the symptoms, what are the goals? First, we sort these observations in the IV and the DV. Second, we sort these observations into variables. Third, we determine how these variables vary, and whether they vary by type or by level. This process is how we determine our variables.

This page titled [3.2: How to Organize Variation With Variables](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

3.3: Categorical Variables- Variables That Vary by Type

The variable type is “this” or “that.” The “it” can vary according to what type it is. Gender varies by type. (Assuming a binary perspective) people are male or female. Racial identity varies by type. (Assuming an erroneous polarization perspective) people are (“labeled” according to the government) as White/Caucasian, Black, Asian, Middle Eastern, Native American, Alaskan, Hawaiian, with Latino/Hispanic as an ethnic identity. Cancer varies by type (at least under current conventional wisdom), such as breast, prostate, skin, and lung cancer.

Type means sorting the data as being one or the other. There is no overlap between the types; you can only belong to one type, not the other. So, people are sorted into males or females, not both, White/Caucasian, Black, Asian, Middle Eastern, Native American, Alaskan, Hawaiian, Latino/Hispanic, but not in several types, or having breast, prostate or skin, or lung cancer. (And yes, these examples assume an erroneous binary, which is not good according to several perspectives).

There are several general and statistical terms for variation by type. We think of type in terms of groups and categories. People are sorted into group A or group B. The terms type, group, and category are used interchangeably. We say, what are the “types” within the gender variable? Or how many groups are there within the race variable? How many types does the category of race consist of? They refer to a variable and then sort it into categories, types, or groups.

Statistically, we think of type as nominal variables or fixed effects. Nominal variables are simply variables that vary by type, and by nominal, we usually think of “name” and not a number. So, gender type varies by name, male or female. Race varies by name, such as White/Caucasian, Black, Asian, Latino/Hispanic. By fixed effect, we mean no variation within each group in a variable. Everybody in the male group is considered male, and everyone in the female group is considered female. There is no one more “male” or “female” than the others within that group (yes – the preceding sentence ignores social constructs of masculinity and femininity, but for now, the example is only used to illustrate a fixed effect).

How do you tell if something varies by type? Well, “it” varies by type if whatever it is can be sorted into categories or groups that are discrete or mutually exclusive. The categories and groups are different from each other. It can only be in one category or another. What makes something a type is if one group consists of members or things with something or attributes in common, AND the other group does not have any attributes: no sharing, no blurring, no spillover of the characteristics.

Gender is a common example. Assuming binary for the moment, we sort people into males and females because we say males have something (yes, biological) that the females do not have.

As you can surmise, gender may not be the most outstanding example of type under a gender-affirming, continuous social construct identity perspective. Because there is overlap among genders, rightly so. A better example is treatment versus control groups. One group gets the treatment; the other gets the control group. There is no overlap between the groups because the treatment group receives the treatment, and the control group gets nothing that resembles the treatment. Even here, rightly so, there could be overlap. If the treatment group gets mindfulness treatment protocol, who is to say that the people in control are NOT getting some type of mindfulness practice in their everyday lives? So, is treatment versus control a fitting example of variation by type? More on this later.

When we think that something varies by type, it is important to note that no one type is higher, lower, more, less, better, or worse than the other. There is no valence or value to the type. The order of the groups or categories within a type is arbitrary. Think horizontal, on an even plane. All groups within that type are on an equal plane. The order DOES NOT MATTER! Emphasis is necessary because order matters a lot when we discuss continuous variables later.

In our gender example, males and females are on an equal level. Males are not more, better, or less than females and vice versa (assuming a gender equality society, which, as we know, is not happening). Which means the order can be interchanged. Females can be listed first, then males, and it does not matter.

What exactly doesn’t matter? The numbers that represent each group or category within that type.

Numbers are codes that represent each type. The numbers don’t matter. The numbers themselves have no value; they are just codes, numbers arbitrarily assigned to each category. Statistics programs cannot process words. So, we need numbers for each category or group within that type for the statistical program to do its computations.

For gender, we cannot enter “male” and “female” into the statistics program. We code them with “male” = 1 and “female” = 2. The order does not matter. We could use “female” = 1 and “male” = 2. The number does not matter either. We could use “male” = 0 and “female” = 1, or “male” = 5 and “female” = 7.

Honestly, we don’t just assign numbers arbitrarily. Intuitively and for organizations, it is best to assign numbers in conventional ways. So yes, a “5” and “7” seem random and don’t help our projects. It does matter for organizational purposes to assign numbers in an order that makes sense for comparison and advanced analyses, such as dummy coding in regression analyses. For example, assigning numbers such as “1” for treatment and “0” for control does make sense in terms of low to high order. We want to see if treatment improves symptoms, so we anticipate that treatment will be higher. We give it a “1” for organization purposes rather than a “0” because we want the treatment to produce higher, better symptom relief than the control group. So, sometimes, we want to intentionally assign numbers to code the distinct groups or categories. We’ll return to this issue.

How many types? Depends on what you know about the variable, how many types are possible, and what types you need based on your research question. Usually, we are concerned about whether the variable is dichotomous, or just two groups, or consisting of multiple groups. Gender is usually thought of as dichotomous, male, or female. Race is usually thought of as multiple groups: White/Caucasian, Black, Latino/Hispanic, and Asian.

The best way to answer how many types are within a nominal, categorical variable is to use your conceptualization of your research question. For gender, you can ask if someone is male or female. But is that all you need for your research question? Does it matter if knowing if someone is male or female explains the variation you are curious about in your outcome? Suppose the psychological outcome is willingness to seek treatment. Being male or female might be an issue because, in general, females tend to seek therapy more so than males. So, we need to ask about gender in the male vs. female category. But if you are interested in how cis-gender and transgender seek treatment, now you are introducing another conceptualization. So, we need to ask about gender as male vs. female, but now include transgender Male to Female, Transgender female to male.

At this point, the rabbit hole can really open up. You may be concerned that you did not include people who are transitioning genders and have yet to consider themselves as fully transitioned to male or female. So, you may include additional category types for gender.

What is correct?

We do have to be mindful about being appropriate and not appearing inclusive. If your goal is to create a census, you may need more categories. For example, if you want to create a census of the number of gender-affirming identities enrolling in a psychology graduate program, then yes, to be inclusive, more gender categories are needed to address all identities.

You need to let your conceptualization of your research question be your guide. Suppose you have enough literature review and concern about comparing just males and females and their willingness to seek treatment. In that case, you may only need male and female categories. Suppose you have concerns that something is going on regarding transgender experiences compared to cis-gender experiences and that something may be stigma or hostile gender climates within schools or religious groups. In that case, you may need to include transgender categories. You should not arbitrarily list every category. Including everything simply gets messy in organizing your variables and data.

What is incorrect is to have an “everything but the kitchen sink” mentality to ensure you are addressing all possibilities to ensure that you do not miss anything. Including all possible gender categories is not good practice. Too many gender categories make it difficult to account for all of them in your analysis. It entails more work because you will have to collapse and re-code the categories if there are few or zero participants in each category. Including a category to get one more participant does not dramatically change your results. If you find yourself in a position where you feel you are being forced to include more categories, types, or groups within a variable to be comprehensive, that is a sure sign to say “stop” and not go any further. Only include those categories, types, and groups that are conceptually useful in answering your research question.

Groups typically collapse into fewer groups after examining the data. Race often collapses because there are not enough participants per category. Suppose we examine race but have most participants in the White/Caucasian category and only a few in the Black, Latino/Hispanic, and Asian categories. In that case, someone might decide to collapse the Black, Latino/Hispanic, and Asian categories into a general “minority” or “non-White” category. Is that good practice? Always use your research question and your desired outcome as your guide. If you are doing a census of racial categories of students seeking counselling at a university

counselling center, and the numbers for other racial categories are low, it is best not to collapse them. In this case, your variable of interest is encouraging students from racial groups to attend counselling, and it does matter which racial groups are underrepresented more so than others. So, it is best not to collapse those groups. On the other hand, if you are examining something where race is not the primary concern in the outcome, such as risk factors for pediatric cancer (acknowledging that racial disparities are present in every issue). If that is the case, then collapsing race into, say, a White versus non-White category might be sufficient to answer the research question.

What does nominal data look like? Put differently, what does the data show for each type within a categorical variable? The data are frequency counts or a “yes” or a “no.” The person or observation is included in that category. In the dataset, we assign a code to each person, subject, or observation. So, we have a participant, and we want to show whether they are male or female for their data record. We enter “1” or “2” for the gender code. The statistical analysis output is a frequency count of how many per type are within that variable. If we say we have 50 males and 50 females, our statistical output is a total frequency count or a percentage of how many subjects are in a particular category.

The numbers we assign to each type within that category are just codes; they do not matter and have no inherent meaning. The numbers are strictly just for labelling each of the types. However, the output numbers are total frequency counts representing the number of persons for each type within that categorical variable.

This decision about whether the numbers mean anything or do not mean anything now matters when we turn to the second type of variables considered continuous.

This page titled [3.3: Categorical Variables- Variables That Vary by Type](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

3.4: Continuous Variables- Variables That Vary by Level

Variables vary by level. By level, we mean a) an amount or frequency, or b) intensity or severity. By level, we mean we are sorting our data and observations as more or less of something. This person has more of something than another person. This person has less of something than another person. This person has more intense something than another person. This person has less intense something than another person.

General terms for variation by level can be considered as amount, frequency, or intensity. It all comes down to more or less. Any adjective we use to describe more or less fits here. By amount, we say, taller versus shorter, faster or slower, older or younger, farther or closer. By intensity, we say, hyper or hypo, “drives me crazy,” or “chill,” funny or boring, delicious or yucky, or spicy hot or bland.

Statistically, “level” as a term refers to a random effect. In contrast to a fixed effect, where the variable can be only one value, random effects mean that the variable can be any value. Granted, random means many things in statistics and research. For now, we are using random effects to represent continuous data compared to a fixed effect, which is usually reserved for categorical data. The fixed effect does mean that we are confident that everyone or everything belongs in that category and no other category. Random effect implies that there is allowance or uncertainty that the entity does vary along a continuum.

There is a distinction between how a continuous variable is demarcated according to more or less of an amount versus more or less of an intensity. When the numbers represent an amount, it is self-explanatory. Age, number of years of education, number of therapy sessions attended, and number of days spent on vacation are all examples where the number represents the amount of something.

The issue gets interesting when the number represents the intensity or severity of something. Here, the number, the starting number, the end number, the range, and the unit get murky. The problem is that when a number represents an amount of something, we can usually identify what that something is. However, for intensity, the number represents our *experience* of something, and we infer what that something is and how we experience it.

What is the best way to use a number to represent depression? We know that depression is on a continuum, from no depression to severe depression. We could divide the continuum more by adding mild to moderate depression. The question then becomes, how many numbers are needed, and what is the range of these numbers to represent the continuum of depression?

What is the best way to use a number to represent your attraction to your partner? We could use a scale from one to ten, with one being no attraction and ten being an intense attraction. There are several interesting issues to raise here. Who decided that attraction is rated from one to ten? What is the difference between a one and a two? What are the indicators that someone is a five versus a six? Does everyone use the same criteria to decide if someone is a one, two, five, six, or ten?

What is the best way to use a number to represent pain intensity? The same issues as the attraction scale apply, but I use this example to highlight the role of context. Does the same one-to-ten scale apply to all pain? Do we use the pain scale for a person with a headache versus someone who broke their arm, someone experiencing a heart attack, or someone who is giving birth to their child? Pain is an individual experience. The same event, such as stubbing their toe, might be a level 10 to some people but a three to others, depending on their pain tolerance. The point is that intensity is often in the eye of the beholder, and the intensity range is arbitrary. A discussion of this issue is beyond this chapter; your conceptualization will aid in deciding how best to measure the intensity of the variable.

These issues are interesting. For now, a good idea to consider is if something does not vary by amount it varies by intensity. The challenge is deciding how to best represent that intensity from a range of low to high. Fast forward to the scaling options, and we will note that the ordinal and the interval scaling options are the choices for representing severity.

The variation is on a continuum from low to high; now, the numbers mean something because the number represents more or less of “it.” In contrast to categorical variables, the numbers are codes and have no inherent meaning. However, for continuous variables, the numbers do have meaning because they represent a value from low to high. The numbers matter because they now have meaning. 2 is greater than 1. The numbers stand for more of or less of something.

And having more of or less of something does matter. Something has more amount/more frequency, less amount/less frequency or more intensity/more severity, less intensity/less severity. Some examples are more or fewer hours studying (14 per week), more or

less stress during the study session (stress: 5 on a scale from 1 to 10), more encounters with micro aggressions (2 per month), more intense responses to each micro- aggression (frustration: 7 on a scale from 1 to 10).

The continuum is one continuous line from low to high. We could say that someone is more of “it” than another person. That would suffice, and we use this language all the time. We say someone is taller than another person, more depressed than another person, or has a greater work ethic than another person. For most discussions, this comparison of more or less is enough. But we need something more for statistical analysis, figuring out associations, and quantifying the comparison.

That “something more” is demarcating the continuous line. Demarcating simply means subdividing the line into units. Like a ruler, we take a continuous line and then subdivide it into units. The questions for demarcating the line are how large and even the unit is or if the units are equal in length. When we demarcate the line, we do so in three ways: ordinal, interval, and ratio, and these are the three types of continuous variables.

This order does matter: ordinal, interval, and ratio. Consider these variable continuous types as arranged from less to more precise in measurement. Precision means how well the numbers cannot or can stand on their own. Less precision in the numbers means you need more information to understand the numbers. More precision means that the number itself gives you all the information you need. The ordinal is the less precise of the three, followed by the interval, then the ratio.

3.4.1: Ordinal Ranks Variables

First, the ordinal. Think of ordinal as a rank variable. Generally speaking, ordinal and rank are the same thing. When you read the term ordinal, think rank, and vice versa. Sometimes, you read the term “ordinal ranks.” Technically speaking, ordinal is referring to “order.” Rank is the noun version of arranging things into ranks. Ordinal and rank are the same thing, and there hasn’t been a major reason to separate the two terms.

Ordinal or rank variables have mutually exclusive categories, but the order of the ranks does matter. For example, we use ordinal variables for the degree of clinical severity: 1 = non-clinical, 2 = sub-clinical, and 3 = clinical. Ordinal variables have not set zero points, so you can assign numbers any way you want. There is nothing wrong with redoing the ordinal codes as 3 = non-clinical, 2 = sub-clinical, and 1 = clinical. However, the order does matter, and sub-clinical does have to be the middle category because it is less than clinical severity but greater than non-clinical severity.

No zero point means the numbers do not stand on their own. That means you cannot tell what a 2 means. You must get more information and context to interpret the score’s meaning. You have no idea what a 1, 2, or 3 means unless someone tells you. You need to know the lowest number in the range and the highest number, and what they represent. If someone gives you the number “2” for an ordinal variable, you must know what the “2” is out of what is the lowest and highest score. Is it “2” out of 3 ordinal ranks, or a “2” out of 10 ordinal ranks? You also need to know if a “2” is good or bad according to what the lowest and highest ranks represent.

We use rankings all the time. In sports, we say a team is in 1st place, 2nd place, or 3rd place. When we rank LGBTQ-friendly campus climate, we say this campus has the #1 best ranking, 2 = #2 ranking, and so forth. We do switch what the first rank means. We could describe the #1 sports blooper error this week or rank the worst albums by a favorite artist. For instance, “Yellow Submarine” is the #1 worst Beatles album. In psychology, we do rankings. Non-clinical, sub-clinical, and clinical. We rank behavioral problems from worst-behaved students to best-behaved students.

What should the numbers for the rankings be? This is an issue that we revisit with interval variables. One way to decide the numbers for the rankings is to figure out if we want the lowest and highest numbers to mean a good thing or a bad thing. The numbers do matter for continuous variables, and they matter in terms of good or bad valence.

Intuitively, we use numbers that align with the rankings. So, if someone is in 1st place, we give that rank a “1.” For a non-clinical, sub-clinical, and clinical ranking, we can use 1, 2, and 3, or 3, 2, and 1. Does it matter for the ranking? It does not.

The problem might be when we associate the numbers with something else. Generally, we increase the number when we say there is more of something. It is a lot easier to say that more of “this” leads to more of “that.” It gets interesting for rankings where the 1st ranking is a good thing. The more hours an athlete spends on drills, the more likely the athlete will get first place instead of second place. So, an increase in more hours means that the direction will decrease from a ‘2’ ranking to a ‘1’ ranking. Is it confusing? Possibly. The problem comes when we think about how to describe associations between good and bad things. For psychology, the

more days a person spends time alone and isolated, the more likely they will be in the clinical range of depression. So, more days alone are associated with increasing the ranking from a “2” sub-clinical to a “3” clinical range. If you want to show improvement, the ranking direction becomes interesting. We want to say that the more time a person spends in treatment, the more the person improves. But the ranking decreases from “3” to “2.” So, we say that the more time the person spends in treatment, the person improves by going from a “3” clinical ranking, decreasing to a “2” subclinical ranking. You must be mindful of the mental flip-flops that sometimes occur when you are associating the numbers as representing low or high amounts of the construct.

An important characteristic of the ordinal rank variable is that the distance between and within the categories is not equal, which varies across the variable's full range. When we say a team is in 1st place, a team is in 2nd place, and a team is in third place, we do not say how much the teams are in those places. A team can be in 1st place by 1 or 10 points. It doesn't matter because the team is in 1st place. A team can be in 1st place by 1 point, and the team in 2nd place can be in that place by 10 points compared to the 3rd place team. The distance between the 1st and 2nd place teams and the 2nd and 3rd place teams are unequal. All that matters is how the teams are ranked by 1st place, 2nd place, and 3rd place.

The distance, or the range, within the ranks, can vary across the ranks. This situation happens all the time in psychology. On the MMPI, we give scores from 1 to 100. The scores of 1 to 55 indicate a normal range, 56 to 65 indicate a sub-clinical range, and 66 to 100 indicate a clinical range. The range of the normal rank is 1 to 55 for a span of 55 points. The sub-clinical range has a span of only 11 points. The clinical range has a span of 34 points. Within the ranks themselves, the ranges differ. In education, passing is 70 to 100, and failing is 69 to 0. So, the passing score range is 30, while the failing score range is 69. The range intervals are different within each rank.

We also group entities or people in terms of ranks, and within that rank, there could be people who just made the cut-off score for the rank and people who clearly vaulted the cut-off score. We do this all the time. We give students an “A” as the top grade. However, within that “A” group are students who got a 91 and just made the “A,” and students who got a 99 clearly got the “A.” To the students, it may matter who got the 99 or the 91, but to assign a grade or a rank, these students get the “A.” What matters is that the group as a whole are ranked above all other students who got a “B.”

In psychology, we use rankings all the time. Usually, we use a three-ranking system: normal range, sub-clinical, or clinical range. Admissions interviews sort candidates into three categories: admit, possibly admit or put on a waitlist, and no admit.

Why use ordinal as a variable type?

That might suffice when you believe that the distribution of “it” can be summarized into groups. For example, suppose you want to measure which students are good, not-so-good, or bad. You could count the number of disciplinary referrals per student, and per month, you could count each referral for each student. However, it might be easier to sort the students into general categories such as the following: 0 = none, 1 = 1 to 3, 2 = 3 to 10, 3 = more than 10.

You can use ordinal variables when you are unsure whether there is an equal interval between the ranks. This situation usually occurs when there are “grey areas.” For situations where you are sorting basketball skills, you may sort them into three groups: make the team as a starter, be on the bench, or not make the team. It is difficult to assign a number that designates how to “score” basketball ability, so we might just sort the players into groups and rank them from low to high.

Ordinal variables are used when the distribution of the variable seems “lumpy,” when the distribution of scores does not have to be equal, or when there are usually more people in one group versus another. For example, when measuring the number of alcoholic drinks someone consumes, you could count the number of drinks. However, when sorting through the drinks, one to two drinks may be equivalent, and the difference between one or two drinks might not matter in terms of the effect of driving home. But the difference between two and three or four or five might matter. So, the cutoff for the range is two, but then two or more up to five may matter in terms of trying to decide if the person is too drunk to drive. Any drinks after six and above don't matter; if a person is beyond six, the person is not safe to drive.

3.4.2: Continuous Interval Variables

The second type is the continuous interval variable type. For this variable, the distance between scores is (supposed to be) equal among all points on the scale. Ordinal variables have unequal distances between the ranks and within the ranks. For interval variables, the scores do have an equal distance between them.

Interval and ordinal variables have one thing in common. Both variable types do not have a zero point. This means there is no true zero for any of the measurement scales. So, you can technically use any number you want to arrange the interval scale. Like the ordinal variable, because there is no zero point and you can use whatever number you want, you must get additional information and context (range, average score) to interpret the score's meaning.

For interval scales, you assume the scores have equal intervals: 1 to 2, 2 to 3, 3 to 4, and 4 to 5 are the same distance. That means the distance from 1 to 2 is 2 to 3. In ordinal scales, the distance between the first and second ranks can differ. A person who is considered a heavy drinker will be ranked first in terms of the number of drinks, compared to a person who is a social drinker who is ranked second in terms of the number of drinks. But the number of drinks for the first person could be 10 beers, and the second person could have 2 beers. However, the number of drinks for another person in the heavy drinker rank could be 15 beers, and the number of drinks for another person in the social drinker rank could be 1 beer. The distance between beers in scenario one is 8 beers, and the distance in scenario two is 14. However, the former persons still receive the same rank – heavy drinker- and the latter persons still receive the same rank – social drinker. Even though the two pairs of drinkers have a different number of beers drank between them.

The classic and **only** example of the continuous interval scale is the Likert Scale. Likert scales are the following – “rate your customer satisfaction on a scale from 1 to 5”; “rate your depression on a scale from 1 to 10”; “rate your child’s temper tantrum severity on a scale from 1 to 5.”

The scale itself can be any number range. These scales can be from 0 to 4, 0 to 5, 1 to 5, 1 to 3, 1 to 10, 1 to 100.

The important note about the interval variable is that you have no idea what the lowest or highest number stands for. On a scale from 1 to 10, 1 could be bad, or 1 could be good, 10 could be bad, and 10 could be good. For example, you could rate someone’s physical attractiveness on a scale from 1 to 10, with 1 being unattractive and 10 being attractive. A 1 is bad, and 10 is good. You can also rate how much a student in a classroom is driving you crazy, with 1 as good and 10 as bad. We rate pain on this scale – 1 as no pain and 10 as excruciating pain. In these cases, 1 could be good or bad, and 10 could be good or bad. Someone needs to tell you what the lowest score and highest score mean. The numbers cannot stand independently because there is no context under which you can determine what values are good or bad.

Why use this variable type?

When you want to use a conventional interval scale to measure something. Scales from 1 to 10 are ubiquitous because converting our experiences into a scale of 1 to 10 or 1 to 5 is easy. We have 10 fingers, so we can easily use them to show our experience of something. The range of one to 10, 1 to 5, or 1 to 7 implies a midpoint where we can say we do not think our experience fits either extreme.

We use this variable type when we really do not have a “standard” measurement for something. Usually, this “something” is theoretical. There are no “standard” measurements of racism, life satisfaction, hope or worry about the country’s economic future, or being a team player. We have measures of psychological phenomenon that are in common use, such as the MMPI, the Beck Depression Scale, the PTSD Inventory, the Positive and Negative Schizophrenia Symptom Checklist, the Autism Diagnostic Observation System, the Pittsburgh Sleep Index, and the Weschler Intelligence Scale. But suppose you really think about the history of how those measures were assembled. In that case, there really is no basis for the interval scoring protocol: “Rate the child’s behavior on a scale of 0 to 3, rate the patient’s delusions on a scale from 0 to 7, have you felt you are no longer interested in things you were interested in?” I would argue that there really are no “standard” measures of depression, alcoholism, PTSD, autism, and so forth. Some measures have more popular use and evidence of their validity, but as far as a standard, the psychology phenomenon does not have it.

At this point, I will remind the reader that the ordinal and the interval variables are the only choices we have to measure the intensity of our experience of the variation of a phenomenon. The ordinal and interval variables are used when the experience of the phenomenon is arbitrary. No rule or standard says that the starting point for the experience is zero or one, or that the endpoint for the experience is 10, 20, or 100. The unit of intensity does not have physical properties. There is no such thing as one unit or two units of a phenomenon such as depression. These issues mean the ordinal and interval variables are the only choices we have to measure the intensity of a phenomenon because the ordinal and interval variables, by definition, are not based on a definitive zero and do not have equal intervals as their unit of measurement.

Note the following stance from this author – interval continuous variables, or Likert scales, are flawed. In psychology research, it is hard to determine whether the intervals are equal between two points. For example, what is the difference between a “4” and a “5” or Agree vs. Strongly Agree on a scale measuring an adolescent’s liking for school? What is the difference between a “9” vs. “10” in pain thresholds for a woman who is giving birth compared to a man who is suffering from a leg cramp? What is the difference between a 90 vs. 100 in a stress index? Researchers have pointed out the flaws of Likert scales, and the discussion is beyond this textbook. Suffice it to say that interval variables are inefficient ways to measure something. However, given that most psychological, educational, social, and cultural phenomena that we encounter are based on perceptions and experiences, and don’t have any physical indication of what is nothing, absent, more, or maximum, we are stuck with interval variables, or Likert scales, because there really is no other alternative.

3.4.3: Ratio Variables

Ratio variables have equal interval measurement; the key distinction is that they have a zero point. The zero has meaning: zero does mean zero, or nothing, or the absence of the “it.” The ratio variables are called ratio variables because the number generated by a measurement is compared to zero. So, the number does have meaning, and the number can stand on its own, meaning there is no need for context to interpret the number.

Ratio variables are used when there are physical measurements: height, weight, distance, and speed. In these cases, the number stands independently and does not need context. A person who is 5 feet 2 inches is 5 feet 2 inches. A person who weighs 150 lbs. weighs 150 lbs. A car that goes 50 miles per hour goes 50 miles per hour. Each value is compared to 0: five feet two inches to 0, 150 to 0, 50 to 0. So, the number can stand on its own.

In other disciplines, we use ratio variables more as frequency counts. The population of a town, the number of churches per neighborhood, the number of flu cases during the winter season, the number of deer in a park preserve, and the number of shoplifting incidents for a franchise retail store.

In psychology, we use ratio variables, especially in terms of demographics. Age, income, and years of education are ratio variables. A person who is 30 years old earns \$75,000 / year with a college degree and has 16 years of education.

Psychology uses ratio variables as frequency counts. The number of counselling sessions attended, the number of student disciplinary referrals, and the number of alcoholic drinks consumed per day.

When do you use this variable? Well, the variable is strict in terms of what it can be used for. The ratio variable is clearly used when there is a clear zero point, and the number represents a frequency count. Everything else would be considered an interval or an ordinal variable.

You might want to be careful, however. The entity might be a ratio variable, but how it is scaled as a variable for a statistical analysis might be an ordinal variable. Consider this scenario. You want to study the number of drinks a person has and their decision to drive afterwards. You create a variable called drinks, and the variable is divided into the following: 0-1, 2-5, 6-10, 10 or more. What is the scaling of the variable? In this case, the variable that is entered into the analysis is ordinal. The number of drinks is originally a ratio variable because it is a frequency count of the number of drinks. But the variable is rescaled into an ordinal variable. The variable has a zero point, but the intervals are not equidistant. The first rank is zero or one drink; the second and third ranks have a range of four drinks, and the fourth rank has an indefinite range of drinks. Why rescale the ratio variable into an ordinal variable? For parsimony, theory, and ease of interpretation. It may not matter if a person has two versus three drinks, but it may matter if they have five or six drinks. It definitely does not matter if the person has 10, 11, or 12 drinks, because at that point, the person is definitely in no condition to drive. For conceptual purposes, entering the variable as a ratio variable does not enhance the understanding of the outcome because it does not matter if the person has two or three or four drinks. It may matter if there is a general threshold, such as two drinks versus six drinks, when the outcome is affected. In this case, the number of drinks as a variable could originally be a ratio variable, but it is re-scaled as an ordinal variable for the statistical analysis.

Why the Variables are Arranged from Ordinal to Interval to Ratio

Continuous variables are arranged from Ordinal to Interval to Ratio in terms of the level of precision in demarcating the variation of “it” from low to high. The ordinal variable is the sloppiest; the ratio is the most precise. Sloppy vs. precision is based on how the variation of the variable is divided into units that we count.

The phrase “units” that we count is interesting. We divide the variable “length” into units called inches (assuming the English system here). We divide the variable “weight” into units called ounces. We divide the variable “time” into units called seconds. We divide “money” into units called dollars. In psychology, we divide the variable “depression” into ... well something.

We know that people are more or less depressed, but what is the unit? We know that some people have a bad day, some are sad based on an external event (e.g., a death in the family or the Cubs losing), and some are chronically sad. But what is the unit? The challenge in psychology is that these variables are internal experiences you cannot see. All we can do is infer that something is there and that something varies in amount from low to high. When we divide “depression” into units, the best we can do is infer something to count. For depression, we could ask the number of days that a person is depressed. The days become the unit. Is that the best way to count the number of units for an internal state such as depression? Putting aside any debate, for now, we divide the variable into units, determine if the unit is equal or unequal and if the number assigned to each of those units can be interpreted with or without context.

You might wonder how a unit can be equal or unequal in terms of the physical world. The measurement apparatus itself needs to be equal in units. However, how we use the apparatus might lend itself to unequal units. Take baking. A heaping tablespoon of sugar is different than a tablespoon of sugar. Levelling off one cup of flour is different from scooping a cup of flour and different than weighing a cup of flour, especially if the flour is all-purpose or bread flour (or strong flour, as the British say). The measurement apparatus needs to be precise in terms of equal units. The act of measuring something might yield unequal amounts of the variable.

Back to the dimensions of sloppiness and precision. The factors to consider are: a) is the variable divided into unequal or equal units? b) do the numbers need context to interpret them, or can they be interpreted without context? By context, we mean someone else must tell you what the number represents.

Ordinal variables are the sloppiest because a) the units are unequal, and b) the number associated with each unit needs additional context. Interval variables are less sloppy because a) the units are equal, but b) the numbers associated with each unit need added context. Ratio variables are the most precise because a) the units are equal, and b) the numbers associated with each unit do not need additional context.

This page titled [3.4: Continuous Variables- Variables That Vary by Level](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

3.5: Nuances in Determining if a Variable Should be a Certain Type

The sloppiness vs. precision continuum does not mean that one variable type, in this case, ratio, is better or more valued than the others, in this case ordinal or interval. The sloppiness and precision are descriptions, nothing more. Variables are not inherently sloppy or precise. Sometimes, a “sloppy” variable is the best we can do. The goal of having a variable is to describe the variation. The process is how we divide the variation. How we determine the process of separating the variation has its own nuances.

What can trip you up is that you think that a variable is inherently one type or another. It’s how you divide up the variable into sections or demarcations. You decide how to demarcate the variation you see into something that matters for your outcome of interest.

Let us take age. The first reaction is – “That’s got to be a ratio variable; there is a true zero point, and age can be divided into equal units called years. And years can be subdivided into months, days, and hours. So, age must be a ratio variable.”

That is true, but you can divide the variables in different ways to suit your purposes. So, you can divide the variable as a continuous variable into an ordinal or ratio variable.

For age, you could ask your participants in your research study to state their age as follows: “Write your age here: ____.” Or you can have your participants pick the age range that they are in: 17 and under, 18 to 20 or 21 and up. You can ask them for their perception of their age range: young adult, middle age, retirement age. The first version divides age into a ratio variable. Each response is placed along a continuum of years: 17, 18, 19, 20, 21, 22. The second version divides age into an ordinal variable. Each response falls into one of the three ranks: 1 = 17 and under, 2 = 18 to 20, and 3 = 21 and up. What matters is that in the ordinal version, the value that is entered into the dataset as the age variable becomes “1”; “2”; “3.” The statistical processor is going to compute their statistics using “1,” “2,” and “3” and not the actual age.

Does this matter? It matters depending on the outcome you are interested in. If you need the actual age, such as 17, 18, 19, 20, 21, and 22, in this case, you are saying that increasing the age from 17 to 18, then 18 to 19, and so forth, matters for what outcome you want. In this case, you might be interested in career choices. Students at 17 are about to graduate high school, so they have their career outlook. Their career outlook might be different when the students graduate at age 18. Then, from 18 to 19, they go to college or work, and their career or employment paths might be different at age 20 when they are in college or have two years of work post-high school. If you think your outcome variable of career or work aspirations changes as a result of age, and the changes in age from year to year from age 17 to 21 matter from a conceptual standpoint, then yes, consider “age” as a ratio variable because you need the variation of “age” to be divided along a ratio scale. The scale has equal units, in this case by year, and the number of years can stand independently without additional context.

Does scaling the variable age as an ordinal variable matter? Perhaps. You might be only interested in ranking the ages 17 and under, 18 to 20, and 21 and up. Why? Well, let’s extend our previous example. If you are examining career aspirations of a population of high school students, maybe you need to know if the student: graduated high school at age 17 and under; or if the student graduated high school and went to college or work; or neither, such as going on a European vacation. If the student is age 21, this presumably means the student has graduated, or is working full time, or still in school. Still, age 21 is the legal age to drink alcohol, and when most people would consider the adult stage, that age range has something meaningful to say about career choices. So, while “age” is the variable, how you divide it is according to what variations you need from “age” to predict what you want.

Age as an interval scale does seem hard to conceive. Age as a categorical variable does seem hard to conceive. So, no examples will be given here.

How about gender, though? Gender is often conveniently thought of as a categorical variable in a binary form: female and male. But as our thinking of gender advances, we know that the male-female binary is not the only option. Gender can be continuous. There is cis-gender male, cis-gender female, transgender male to female, transgender female to male, or non-binary gender. In this context, the variable gender can still be categorical. The codes for the expanded gender options range from 1 to 5. But these numbers are just codes; there is no inherent value in any of the numbers themselves. The numbers are entered into the dataset and the statistics processor as codes, and the statistics processor will not analyze the codes as being higher or lower than the others. There is no such thing as cis-gender female being greater in value than cis-gender male and transgender female to male as being greater in value than transgender male to female.

Could gender be thought of as a continuous variable? For that possibility, we need to consider whether the values range from low to high for something like gender. One scenario is considering gender as a continuum and how people perceive their gender qualities. Perhaps the continuum is masculinity to femininity. Some males might think of themselves as masculine; some females might think of themselves as masculine; some males might think of themselves as feminine, and some females might think of themselves as masculine. In this case, we could ask participants to rate their masculinity-femininity, ranging from low to high. What would the range be? We could use one to five, one to four, one to 10, or one to 50. Does the range matter? It should. Unless you have a conceptual reason to set the range at one to five and decide what you think a one could represent in terms of masculinity or femininity compared to a two, then you should set your range accordingly. You could set the scale from one to 10. In this case, you need to consider what a 9 means compared to a 10 for masculinity or femininity. If you set the scale from one to 50, then what does 45, 46, 47, 48, 49, or 50 mean in terms of the level of masculinity or femininity?

The example calls for an interval scale if you haven't surmised by now. The identification of 1 to 5, 1 to 10, or 1 to 50, as the range is arbitrary. You do not know what a 1 or 5 or 10 or 50 means in terms of level of masculinity or femininity unless someone tells you. Does your choice of interval range matter? According to your outcome, yes, it could. Does it matter if the range is 1 to 10 vs 1 to 50? Does it matter if your participants indicate a 45 vs. an 8? If you set the ranges as 1 to 5, 1 to 10, or 1 to 50, then you need to conceptualize what those ranges represent to defend your choice.

It should be noticed here that the range is from masculinity to femininity. This range, as written, implies that the low end of masculinity is femininity, and the low end of femininity is masculinity. As far as anyone can tell, there is no verifiable evidence that the opposite pole of masculinity is femininity. It is quite likely that gender is not one continuum, but two continuums, masculinity and femininity. In turn, gender is now reconceived as gender expression, and now there are two types or two categories of gender expression: masculinity and femininity. Each type could have an interval scale: 1 to 5, 1 to 10, and 1 to 50. Does this matter? Conceptually, it is best to think of your variable of gender expression as two types because people may consider themselves as exclusively all masculine, exclusively all feminine, or both masculine and feminine, or neither masculine nor feminine. If that structure of gender identity matters for your outcome, then it is best to think of dividing the continuum of gender identity expression in that fashion.

These discussions are meant to illustrate that variables are not inherently categorical or continuous; if continuous, they are not inherently ordinal, interval, or ratio. The demarcations and scaling of your variable should be according to your conceptualization of the variation of the variable and how your variable relates to the variations in your outcome variable. Variables can be categorical, dichotomous or several categories, or continuous, and in ordinal, interval, or ratio form. Depending on the context of what you need that variable to vary to predict your outcome, according to your conceptualization of both the predictor variable and the outcome variable, that conceptualization determines the type of variable, which is based on how you want to demarcate or subdivide, the variation of that variable.

There is no "right way" for a construct to be one variable type or data.

Always think conceptually!!!

This page titled [3.5: Nuances in Determining if a Variable Should be a Certain Type](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

3.6: Making Decisions About the Variables in Your Research Study

There are two reasons why you need a schematic for making decisions about your variables. The first reason is to decide how to structure your variables for your research design. The second reason is deciding what statistical test to use.

The decision you need to make is how you will sort and divide the variation you see in your “it” or the variables you see.

3.6.1: Identifying the Variables

1. First decision – how many variables are there?
2. Second decision - Are the variables independent or dependent variables?
3. Third decision - How are the variables scaled? Is it either Categorical or continuous? And if continuous, is the variable ordinal, interval, or ratio?

For your first decision, decide what is the “it” that you are studying. What is “it” that you want to understand? You want to “operationally define” your variables. You want to determine the definition of “it” and what “it” is not. If you want to study depression, you must ask yourself what type of depression you are curious about. Is it general depression, chronic depression, treatment-resistant depression, seasonal depression, or depression associated with an event such as grief?

The process of “operationally defining” “it” often involves a literature review to gather the conceptual understanding of what “it” is. You may decide what “it” is, but you may think that “it” has several dimensions, factors, or types. In psychology, we assess intelligence. Intelligence is subdivided into other types of intelligence, such as verbal, math, and analytic. We assess for PTSD, but we view PTSD as encompassing an array of symptoms, such as physical symptoms, for example, heart racing, sweating, and lack of sleep; social symptoms, for example, not wanting to socialize; and neurological symptoms, such as flashbacks or nightmares. We choose which dimensions, factors, or types of “it” according to what information we need to predict our outcome.

Then, think about how “it” varies. Recall that not every variable is inherently categorical or continuous. Variables can be scaled in different ways. How you divide the variation you experience in your variables will be according to your research question and your outcome.

A good place to start would be deciding how many different variables are associated with “it.” Start with the different types of “it” that are associated with the overall “it.” For depression, you could start with a self-report of mood, then loss of interest in activities, loss of interest in social relationships, isolation, and suicidal thoughts. So, start with the types. Each of these “types” constitutes its own variable associated with the overall “it.” Then, figure out how each of these variables varies and how you want them to vary by category, type, or continuous, which is by level.

When designing your research study, why make a big deal of this? Because researchers can take a sloppy approach to constructing their variables. Sloppy definitions of “it” and sloppy descriptions of how “it” varies, as well as how to divide the variation of “it,” can lead to sloppy variances. The sloppy variances will lead to a statistical analysis that cannot detect patterns among the independent and dependent variables. Statistical analysis cannot “save” an analysis if the variables are poorly defined and their variations are poorly demarcated. A chef only uses quality ingredients. To make a five-star worthy dish, chefs cannot do much with subpar ingredients. Researchers and statistical analyses need to have quality variables. The “it” is often messy in terms of how the variation of it is described by type and level. There are no straightforward ways to carefully define bullying in terms of type and by level, including more or less amounts and more or less intensities of violent behavior. One of the reasons why these variables can be sloppy is because there are several dimensions to any “it.” Bullying can be divided into social, verbal, physical, and internet bullying. Bullying can be different among classmates, between males and females, between males and males, and between females and females. Verbal bullying could be teasing, shaming, name-calling, or threats. You could count the number of instances of bullying incidents occurring. You could describe the intensity of a specific bullying incident, such as a physical shove or cornering a student in the hallway. You could catalog the bullying on ordinal scales, such as mild, threatening, or hostile. You could rate bullying as not feeling safe on a scale of 1 to 5 or 1 to 10. So, the “it” can consist of a chaotic variation. However, careful planning and review can make the variables and how they are scaled more precise so that the data collection for each variable does not end up as messy data.

3.6.2: Selecting a Statistical Test

1. The decision to select a statistical test has three steps:
2. How many variables are there?
3. Are the variables independent or dependent variables?
4. How are the variables scaled? Is it either categorical or continuous?
 - a. And if continuous, is the variable ordinal, interval, or ratio?
 - b. Consider whether the variation of the variable ranges by amount or intensity.

First step – how many variables are there?

To most, this question might seem like a “duh” question. However, the number of variables is not always clear when deciphering research studies and analyses.

It is not always clear what the variables are. Consider this example. A psychologist wants to examine hospitalized clinical patients and their depression and suicide severity levels. The depression and the suicide severity levels could be two variables – depression and suicide severity. Or it could be one variable where depression is the primary variable; and if the depression is high, it calls for a suicide severity risk. In most cases, research scenarios are not trying to “trick you” about the number of variables, but it is something to keep in mind.

Remember that not all variables are independent or dependent. Some of the variables are what we call covariates. Covariates are like independent variables but are not essential or important for predicting your dependent variable. Most covariates are demographics, such as age, gender, and race. These variables usually predict the outcome variable, but they are not that important because there might be another independent variable that is the focus of your study, such as the number of therapy sessions or the number of times spent doing mindfulness sessions.

You need to decide the number of independent and dependent variables. The only two answers to that decision are one or many.

The remaining steps were discussed in previous sections:

The second step is deciding if the variables are independent or dependent variables; the third step is how the variables are scaled. The fourth step is the variation by the amount or intensity.

Putting all this information together.

Fast forward when we talk about the statistical test, but suffice to say, if there is one independent variable, and the variable is categorical scaling, you are likely looking at using a t-test if there are only two categories for the independent variable or an ANOVA if there are three or more categories for the independent variable (you can actually use an ANOVA for two categories for the independent variable, but more later on that twist). If the one independent variable is continuous and either an ordinal, interval, or ratio scaling, you are looking for a Pearson or a Spearman correlation. The aforementioned tests only have one dependent variable, which is continuous.

If you have more than one independent variable and all are categorical, you are likely using an ANOVA. If you have more than one independent variable and a mix of categorical and continuous variables, or if they are all continuous variables, then you are likely using a regression. In both cases, there is only one dependent variable, and the variable is continuous.

If you have more than one dependent variable, and the variables are all continuous, then you are looking at a MANOVA or Multiple Analysis of Variance. If there are one or more independent variables, they need to be categorical scaling.

If you have more than one dependent variable, and the variables are all continuous, then you are looking at a MANOVA or Multiple Analysis of Variance. There can be more than one independent variable, but they must be categorical scaling.

If you have one independent variable and one dependent variable and both are categorical or if you have more than one independent variable and more than one dependent variable and both are categorical, then you are looking at a Chi-square analysis.

This page titled [3.6: Making Decisions About the Variables in Your Research Study](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

3.7: Lessons Learned About Variables

There are miscellaneous lessons about variables.

Temperature as a variable is a sticking point. Is temperature truly a ratio variable? After all, there is a zero point – 0 degrees Fahrenheit, 0 degrees Celsius. But that zero point is based on a continuum that is human made. Under the Fahrenheit scale, water freezes at 32 degrees; under the Celsius scale, water freezes at 0 degrees. Under the Kelvin Scale, water freezes at 273.15 K. So, there is no established zero point; it is more like there is an established shared state - frozen water, and the different measuring systems have different zero points. So, there doesn't seem to be an established zero for temperature, so can temperature be considered a ratio variable? The discussion might be the equivalent of splitting hairs. Suffice it to say, temperature is a ratio variable. Even though there is no definitive 'zero' point, once you decide which scale you are using, the number can stand on its own without interpretation. Temperature is a ratio variable unless there is a real need to open this discussion.

Returning to the concept of the variable, always remember that when we say variables vary, we really mean there is a change. Change can refer to a change in level, such as low to high amounts or low to high intensity. Or change can be changing from type to type, such as control to treatment or male to female.

It is crucial to know if your variable is considered categorical or continuous. For continuous variables, it is crucial to know if the variable is ordinal, interval or ratio when selecting a statistical analysis. For correlations, it is crucial to know if the variable is interval or ratio versus ordinal. Interval or ratio variables use equal units, and the Pearson correlation is called for. Ordinal variables use unequal units or ranks, and the Spearman correlation is warranted—more on this in the correlation section. However, for regressions, we usually use a regression for all continuous variables, regardless of whether they are ordinal, interval, or ratio. No, that is not good practice, but this issue will be discussed in later regression chapters.

Use your literature review to determine how past studies scaled these variables. How did studies scale the variation for PTSD, bullying, depression, and intelligence? Inventory the variables and how they varied and if those variations helped predict the outcome variables, then catalogue this information as part of your literature review. I would catalog the following.

1. What was the research question? What are the authors trying to predict?
2. What were the constructs under investigation?
3. Do the constructs make sense in terms of how those constructs could predict the outcome?
4. What were the operational definitions? What was the conceptual basis for the operational definitions?
5. Describe how the constructs were defined as variables. Were there several variables per construct?
6. How many independent variables, dependent variables, and covariates?
7. How were the variables scaled? By type or by level. And if by level, they are continuous, and were the demarcations by ordinal, interval, or ratio?
8. If by type, does the number of categories per type align with the theoretical expectations for the variable?
9. If by level, then what do high and low values mean? Do the values mean high or low in amount or high and low in intensity?
What are the score ranges for the variables? Is it clear not only what the low and high scores mean but also what the midrange and all other values mean when measuring the construct?

Cataloguing this information could help sort through your variables as you design your own study.

What usually happens is that researchers tend to add variables. Collapsing them, subdividing them, combining them. This practice happens for many reasons. Sometimes, we combine variables when there are not enough participants per category. For example, for the variable race, we tend to have categories in White, Black, Hispanic, and Asian. However, some contexts do not have participants from all these races, such as certain neighborhoods where the census shows a majority of White persons. So, some researchers may collapse the few participants in the Black, Hispanic, and Asian categories into one "minority group" category. Is this good practice? Possibly, if determining who is Black, Hispanic, or Asian has no bearing on the outcome variable. We will discuss this issue more when we discuss the ANOVA analysis.

I bet reading all of this material gave you more than you could ever want to know about variables. Variables are like your ingredients. We work hard to obtain high-quality ingredients to create a high-quality meal or dessert. We work hard to understand our phenomenon of interest by creating high-quality variables to obtain an outcome that furthers our understanding of the phenomenon.

This page titled [3.7: Lessons Learned About Variables](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

3.8: Discussion Questions

Discussion Questions

1. Create a decision tree about how to demarcate your variables.
 2. Create a table that outlines the factors of the variable types, pros and cons, and which statistical analysis is associated with the variable types.
 3. Take an issue you are interested in. Name the variables you think comprise that issue. Sort the variables by type and amount. How would you measure each of the variables?
 4. What does the concept of “partitioning the variance” mean?
 5. How do you decide which variables are the independent and dependent variables?
 6. What does it mean to scale a variable, and what are the different ways we scale variables?
 7. Take a variable and be creative in the many ways that you could scale that variable. How would you best decide how to scale that variable?
-

This page titled [3.8: Discussion Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

CHAPTER OVERVIEW

4: Distributions

Learning Objectives

By the end of this chapter, you will be able to:

- Differentiate between categorical distributions and continuous distributions
- Understand the steps for reviewing journal articles and their statistical analysis
- Understand the importance of a normal distribution for continuous variables
- What to do if you have a skewed distribution

Key Terms

- Categorical Distributions
- Continuous Distributions
- Normal Distribution
- Skewed Distribution

Recap

Statistics is about studying variation. So, let's see what kind of variation we have in each of our variables. This is part of the overall process of descriptive statistics. To recap, descriptive statistics simply describe the variation we have in our variables. Descriptive statistics is the first step when doing your statistical analyses. Inferential statistics is your second step. Inferential statistics refers to conducting statistical tests to determine if the variations in our variables predict or help us understand the variations in other variables. It is your classic statistical testing involving the common analysis: t-test, ANOVA, correlation, regression, and Chi-square.

Always do your descriptive statistics first, then your inferential statistics. The goal of descriptive statistics is to assess the quality of the variation in your variables. It is tempting to dive into your inferential statistics and start running statistical analyses. But just like in cooking, you want to take inventory of your ingredients, make sure you have the quality ingredients, make sure you have the right amounts, and organize how you want to use your ingredients before you start the cooking process. It is the same with statistics. We need to know the quality of our variation for each of our variables before we begin analyzing our variables. That is why we start with descriptive statistics before we conduct our inferential statistics.

There are many ways to proceed with descriptive statistics. For statisticians, the goal of descriptive statistics is to satisfy the assumptions of the statistical test. What statisticians mean when they state this goal is they need to examine the variables and their variations to determine if these assumptions are met. If the assumptions are not met, then the statistical test cannot be conducted.

What are these assumptions? How do you meet them? How do you know if the assumptions are met or not met? What do you do if the assumptions are met? What do you do if the assumptions are not met? If the statistical test cannot be conducted if the assumptions are not met, then what happens?

We will gather answers to these questions as we proceed. Suffice it to say, descriptive statistics is a process involving our critical thinking about how to evaluate the quality of the variations in our variables.

[4.1: Distributions – A Picture of Variation](#)

[4.2: Categorical Distributions](#)

[4.2.3: Evaluating the Plot](#)

[4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables](#)

[4.4: Continuous Distributions](#)

[4.5: Evaluating the Quality of Frequency Distributions](#)

[4.6: Normal Distribution for Continuous Variables](#)

[4.7: Skewed Distribution- The Opposite of Normal Distribution](#)

[4.8: Evaluating a Continuous Distribution](#)

[4.9: Discussion Questions](#)

This page titled [4: Distributions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

4.1: Distributions – A Picture of Variation

A good place to start with descriptive statistics and to satisfy assumptions is to look at the variation of your variables. It helps to analyze variation by looking at it. And by looking at it, I do mean looking at a picture of it. In descriptive statistics, we need to look at a table of values to generate a picture of the distribution, otherwise known as a distribution plot.

Unless someone is setting you up on a blind date, you want to see who you are dating before you even go on a date. If you want to see who you are dating, you need a picture. You will scan that picture and look for something that you want to see based on whatever outcome you want out of your first date. Wow, this paragraph seems so serious in tone. But the paragraph is true, right?

This visual examination process is how we examine distributions. Distributions are pictures of the variations we see in a variable. What are we looking for when we look at variation? We are looking for quality distributions. We want to see how varied the variation is. Is there a wide spread in the variation, or is there no variation at all? Is the variation smooth, clumpy, or broken in spots? As always, our gauge for evaluating the quality of our variation depends on our research question.

There are two types of distributions of variation.

The first type, categorical distribution, describes the variation of a categorical or nominal variable.

The second type, continuous distribution, describes the variation of a continuous variable, including the ordinal, interval, and ratio variables. There are different criteria for evaluating each distribution.

To evaluate distributions, you need two things. The first is a table. The table has two columns. The first column contains the variable's values. The second column is a frequency count of the number of times you observe each value. The times you observe each value consist of the number of participants who provided that value or the number of counts per value.

Table One displays frequency distributions for a gender-categorical variable.

Table One - Frequency Chart for Gender Categorical Variable

Group	Frequency	Percent	Valid Percent	Cumulative Percent
Male	10	13	13	13
Female	64	83	83	96
Other (please specify)	3	4	4	100
Total	77	100	100	

Table Two displays frequency distributions for a continuous variable.

Table Two - Frequency Chart for Emotional Expression Continuous Variable

Value	Frequency	Percent	Valid Percent	Cumulative Percent
34	1	1.3	1.3	1.3
43	1	1.3	1.3	2.6
44	2	2.6	2.6	5.2
45	1	1.3	1.3	6.5
46	4	5.2	5.2	11.7
47	2	2.6	2.6	14.3
48	5	6.5	6.5	20.8
49	6	7.8	7.8	28.6
50	4	5.2	5.2	33.8
51	1	1.3	1.3	35.1
52	7	9.1	9.1	44.2
53	4	5.2	5.2	49.4
54	6	7.8	7.8	57.1
55	5	6.5	6.5	63.6
56	3	3.9	3.9	67.5
57	4	5.2	5.2	72.7
58	4	5.2	5.2	77.9
59	1	1.3	1.3	79.2
60	2	2.6	2.6	81.8
61	3	3.9	3.9	85.7
62	2	2.6	2.6	88.3
63	1	1.3	1.3	89.6
66	1	1.3	1.3	90.9
68	3	3.9	3.9	94.8
69	1	1.3	1.3	96.1
71	1	1.3	1.3	97.4
72	1	1.3	1.3	98.7
73	1	1.3	1.3	100

The second thing you need is a plot. The plot helps you observe the variation. The plot consists of an X-Y plot figure. The X-axis, or the horizontal line, is each of the variables' values. The Y-axis, the vertical line, is the frequency count. The columns along the X-axis represent the number of times you observe that variable's value. Figures One and Two present frequency plots for the gender and emotional expression variables.

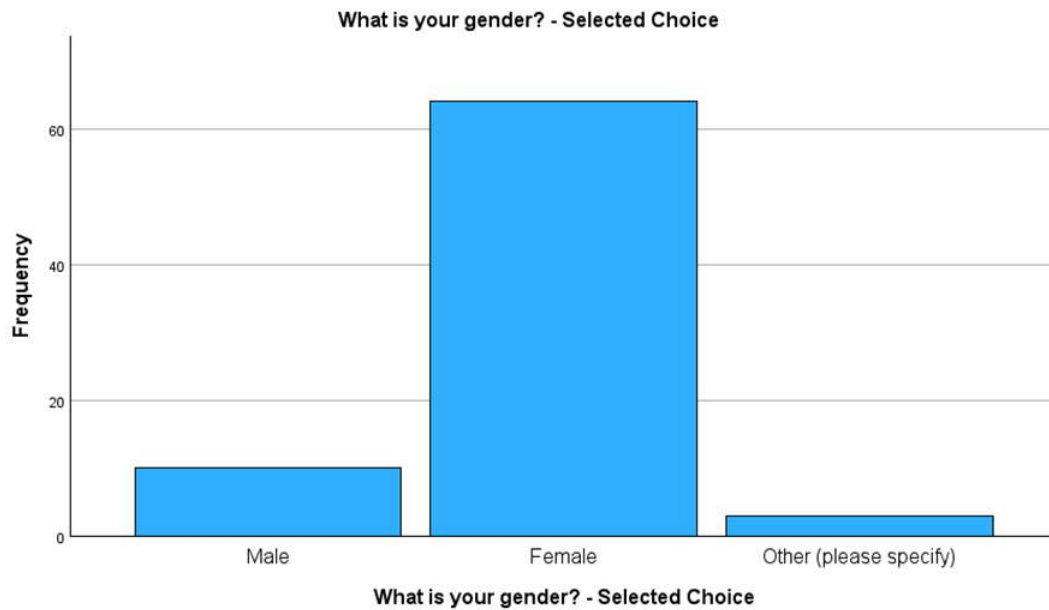


Figure One: Frequency Plot for Gender Categorical Variable

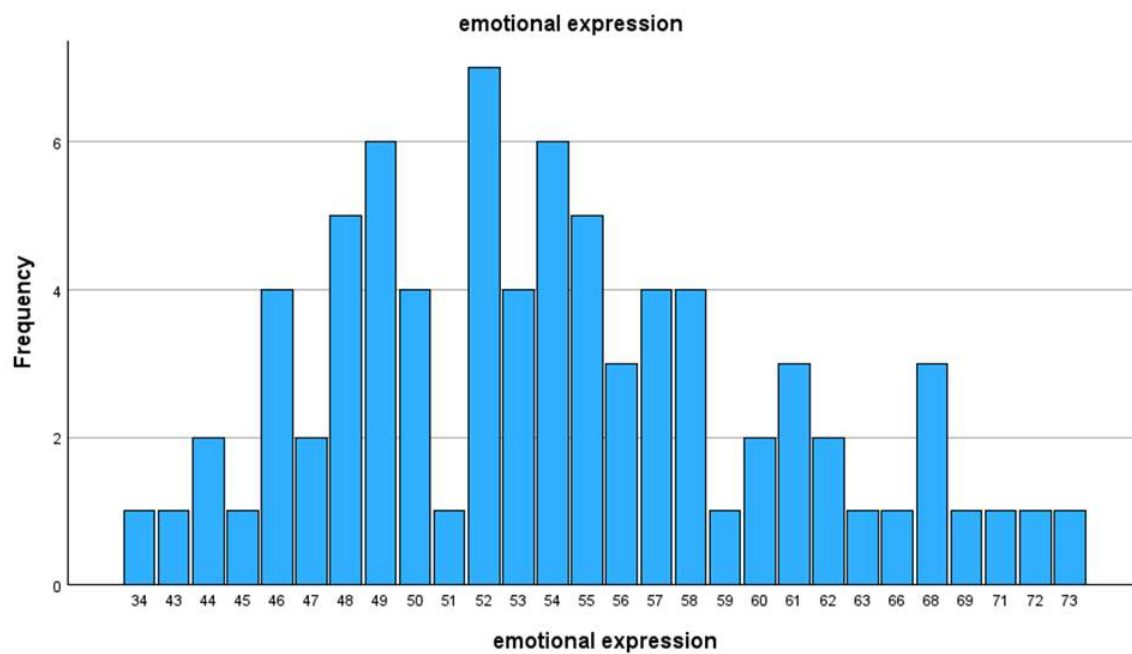


Figure Two: Frequency Plot for Emotional Expression Continuous Variable

This page titled [4.1: Distributions – A Picture of Variation](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

SECTION OVERVIEW

4.2: Categorical Distributions

Categorical distributions are for categorical, nominal variables. Usually, these variables are demographic variables. Examples are gender: males, females (assuming a binary world, of course), diagnosis: anxiety, addiction, PTSD diagnosis, race: Black, White, Hispanic, Asian (assuming these are the only racial categories), academic major: Liberal Arts, Physical Sciences.

For categorical variables, the data consists of counts, or the number of times you observe someone to be in one of those types or groups within the nominal category. This distribution consists of the number of times you counted that someone or something was in a particular category.

The numbers assigned to each category type are just codes. Gender: 0 = males, 1 = females (assuming a binary world, of course), Diagnosis: 1 = Anxiety, 2 = Addiction, 3 = PTSD diagnosis, Race: 1 = Black, 2 = White, 3 = Hispanic, 4 = Asian (assuming these are the only racial categories), Academic Major: 0 = Liberal Arts, 1 = Physical Sciences.

Notice the codes. Some codes start with 0 (males) or 1 (anxiety). You can number the codes any way you see fit. For general ease, it seems easiest to start with 1 and 2, and so forth. But the number, mostly, does not matter for the purpose of entering the codes into a statistical processor. I say this mostly because, in later analysis, the order of the codes could matter in terms of comparison. One of the analyses, or statistical procedure, is dummy coding. Here, one of the categories has to be 0, and the other categories are 1, 2, 3, and so forth. We will cover this issue in regression analyses and other analyses. For now, let us keep it simple and say that as codes only, the numbers assigned to each group within a categorical variable can be anything, and for conventional purposes, it is easy to number them as 1, 2, or 3.

You would be forgiven if you were confused about the codes assigned to each group within a categorical variable, and the data summaries. The data for nominal variables are summarized as frequency counts. It is important to discern between the two. So for our examples, the following are summarized frequency counts for the group – Gender: males, $n = 16$, females, $n = 16$; Diagnosis: Number of patients with Anxiety, $n = 35$, Addiction, $n = 23$, PTSD diagnosis, $n = 16$; Race: Black, $n = 345$, White, $n = 125$, Hispanic, $n = 163$, Asian, $n = 142$; Academic Major: Liberal Arts, $n = 93$, Physical Sciences, $n = 76$. Side note: the n is lowercase and italicized to represent the sample size for groups, while the N represents the total sample size (APA manual, section 6.44, p. 187). Those counts are results and not the codes for each group.

Recall that these variables have variations by type and group, are mutually exclusive variables, and are discrete. In Figure One, or Figure 1, the frequency plot, along the X-axis, the participants are distributed horizontally. There are no values, and no one group, or type is higher or lower than the other. The numbers assigned to each group are just codes; there are no values or valences assigned to each number.

Why is it important to recognize that there are no numbers with values, just codes? Because it means the variation by type can be in any order along the X-axis. The order is arbitrary. If the order of the arrangement of the types along the X-axis is arbitrary, that means the shape of the distribution can be anything.

Why do you give a rat's ass about that? In contrast to categorical distributions, the shape of a continuous distribution does matter, and we hope to see a normal bell curve distribution. However, categorical variables do not have to follow that distribution because the order of the types is arbitrary.

4.2.1: Creating the Table

The first step is to create a table of the nominal variables for each group within that nominal variable using SPSS.

4.2.2: Creating the Plot

The second step is to create a distribution plot of the nominal variables for each group within that nominal variable. This is repeated from above. The distribution plot consists of the X-axis and the Y-axis. The X-axis lists all the groups for the nominal variable in the order of the codes. The Y-axis is a frequency count of how many participants or observations fall in each group.

4.2.3: Evaluating the Plot

This page titled [4.2: Categorical Distributions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

4.2.3: Evaluating the Plot

What are you looking for and why?

Evaluating the categorical distribution plot means determining if the sample characteristics are a good estimate of the population's characteristics.

Remember that all statistics involve comparison. In this case, you are comparing your sample characteristics to the population characteristics. The question is this – Did we get the “right” sample?

Evaluating categorical distributions is good for determining if you have possible confounds. Common confounds include having too few participants in certain categories, or an imbalance of participants across the categories. Briefly, confounds are issues in your data or your research design that would invalidate the results of your research study. A goal of doing a descriptive statistical analysis is to discover if there are confounds to address that might affect your results. This issue will be continuously revisited.

Evaluating the distribution plot.

To evaluate the distribution plot, keep in mind that the shape of the distribution plot for a categorical variable can be anything. The reason the shape can be anything is because the codes we use for each group are in arbitrary order. For example, gender. On the X-axis, males can go first, females can go second, or females go first on the X-axis.

Along the Y-axis are the frequency counts, or the number of participants in each group. Continuing our example, gender: males, $n = 16$, females, $n = 16$. The plot is listed below. How do you evaluate it?

Recall that all evaluations involve comparisons. Most comparisons are basically in the form of good or bad in comparison to something.

Comparing the sample to the population characteristics.

In this case, one method of comparing your distribution plot is to follow the framework of how the sample compares to the population. The distribution plot represents your sample, and we want to compare if the sample distribution plot is the same as the distribution found in the population. We need the population characteristics in mind to make the comparison.

Remember – you are always looking to compare what you are looking at to something. In this case, you want to ask yourself the following questions. Does the sample (what you have in front of you) represent the population (what you expect to see?), or do the sample statistics represent the population parameter characteristics? Are these sample characteristics expected or unexpected, given the population and the research question?

Suppose we are evaluating a college classroom and gender distribution to determine associations between gender and grade. In this case, males, $n = 16$, females, $n = 16$, might be a good representation because if there are 32 students, there is an equal number of genders, and it looks like we got all the students. In this case, the sample characteristics are expected given the proportion of males and females, which is 50/50, we would expect in the population.

But what if the context is students who receive clinical counseling services from a college counseling center? Suppose we got the following results: males, $n = 16$, females, $n = 32$.

Is this a good or bad result? Well, it seems good. Here's why. In the population, more females than males use counseling center services. So, while it is tempting to think that there is not an equal balance of the number of males and females, given the context of a counseling center population, this imbalance and these numbers are probably what's expected. To evaluate this distribution of this nominal variable, it seems fine given the context of the population, which is that higher percentages of females will seek counseling compared to males. In this case, the sample characteristics are expected, given what we know about the population.

What if the context was adolescents in a juvenile detention center, and we got the same results: males, $n = 16$, females, $n = 32$? Is this a good or bad result? Well, it seems bad. Here's why. In the population, more males than females are in juvenile detention centers. So, it is odd that there are more females than males because we would expect more males than females. Especially if we picked the sample at random because if it was a random selection, there should be more males than females. In this case, the sample characteristics are unexpected, given what we know of the population.

Are the previous scenarios a problem? It depends on the research question and if the variable you are interested in is the main focus of the research question. In the college classroom example, suppose the research question is to compare the males and females and their grade performance in the class. In this case, the results are good because you want an equal number of males and females to establish a balanced comparison. However, suppose the research question was about the number of hours studying for the course and their grade performance. In that case, the gender variable is not germane to this question, so whatever result you obtain about the distribution of males and females is not relevant.

In the college counseling center example, if the research question is about gender differences in seeking counseling services, then yes, it is possible the imbalance with more females than males might be an issue. However, the point of that research question is that it is expected that more females would seek therapy than males. Despite the imbalance, it does reflect the reality of the situation, so it might not matter that you have the imbalance.

The third research question is interesting. It is not expected that there will be more females than males. If the research question is to compare gender differences in therapy outcomes in juvenile detention centers, then you might wonder if this detention center with more females than males possesses any unique characteristic that might affect the results. It's hard to say what this characteristic might be, but it does raise some questions, and it is best to obtain a consultation to be certain if the data are suitable for answering the research question.

So why is it good to examine frequency distributions of your nominal (categorical) variables? It is to determine if your sample characteristics represent the population characteristics.

If the population's characteristics are not well known, what else should you be looking for? Without any preconceived notion of population characteristics, we typically look for imbalances across demographic groups. For a demographic variable like race, you might be looking for equal numbers of participants across all races. If your research question is about using race as a predictor of an outcome, such as mental health, then yes, you would need to determine if the races are balanced in numbers.

What does it mean when the population characteristics are not well known? There are some populations for which we simply do not know all the usual demographics. For example, for an undocumented immigrant population, we do not know if our distribution of gender, age ranges, race, economic status, or geographic location for our sample is similar to the population. Recall that the population is unknown because it is impossible to know everything about a population. For a population such as undocumented immigrants, it is difficult to know the true distribution of the population characteristics.

Sometimes, the imbalances can be considered as confounds for your research question. For example, you are looking at race as a predictor of mental health outcomes. If your race variable has White, Black, Hispanic, and Asian, and there are 50+ in the White, Black, Hispanic group, and only 10 Asians, then consider whether including the Asian participants is worthwhile. It is difficult to establish a trend in any variable with 10 participants. This issue will be discussed in the sample size section. For now, you could remove those 10 Asian participants or conduct the analyses and indicate as a limitation that conclusions about Asians as a predictor of mental health outcomes are inconclusive. But at least when you examined the distribution of participants across the groups in the race variable, you spotted an imbalance. That is a good way to review your data and then bring the concern to a consultant to determine if the imbalance will pose problems in your statistical analysis.

What sometimes happens is that we collapse groups when there are too few numbers in each group. For a demographic such as religion, we might list several religious denominations. But if there are too few within a denomination, we might just add them together to retain the participants. If you have religious denominations such as Protestant, Lutheran, and Catholic, but there are few participants in the Protestant and Lutheran denominations, you might consider putting all three denominations under a broader Christian religious group, or you might combine the Protestant and Lutheran groups. It depends on your research question and how you conceptualize the differences between the groups. If the groups are similar enough, collapsing the groups might not make a major difference in your outcome.

How do you collapse groups? Collapsing groups is synonymous with combining groups. To combine groups, you add the groups together. A statistics program would allow you to substitute a value for the combined groups. Suppose you want to collapse Blacks and Hispanics into one overall race. If Black was coded as "2" and Hispanic was coded as "3," you would relabel the races as a "2," or Black coded as "2" and Hispanic coded as "2." Be mindful of haphazardly collapsing groups. Combining Blacks and Hispanics from a racial framework makes little sense because of the racial differences between the two groups. However, combining the two groups could work if you are examining general disparities in accessing mental health care for minorities. So,

the collapse could work here. Always think conceptually when collapsing t groups. Doing so will help you defend your rationale for collapsing them. If you do not have a good conceptualization, then leave the groups alone and do not collapse.

Recap

For categorical distributions, you do evaluate the distribution based on the shape of the distribution. You are not looking for a bell-shaped curve. Categorical/nominal variables are variations by type, and type has no value in terms of variation from low to high. The order of the groups for each categorical variable is arbitrary. This means the frequency count plots for each category can be in any order, and the resulting curve or distribution can take any shape. You cannot rely on the shape of the distribution as an indicator of the quality of the distribution. There are no evaluation criteria for the quality of the distribution of categorical variables.

The only way to evaluate the quality of the categorical distribution is to determine if the variable distribution you obtained as your sample is the same as what you expect from the population under consideration. If the proportions of the frequency counted across the groups in your sample are like what you might expect from the population, then you have a good quality distribution. If they are not, then you might have a problem, and it is best to check with a consultant about options. The only two answers you can obtain from evaluating the categorical distribution are: (a) yes, the sample distribution of the participants across the groups for a given categorical variable is expected given the distribution of the participants across the groups for the population, or (b) they are not expected.

Definitely do not rely on having a balance of participants per group. Sometimes, we expect equal numbers of males and females in a group; sometimes, we do not. Equal numbers across the groups are not a good way to evaluate the quality of the distribution for a categorical variable, simply because the distribution may not be equal in the population.

One option is to drop a group if there are too few participants in that group. Or you can collapse that group into other groups if it conceptually makes sense to do so. Or you consider that too few participants in a group are a limitation that your study cannot address when considering how the results support your hypotheses.

Above all else, think conceptually. If the distribution across your groups is for a nominal categorical variable that is part of your research question, then you have to consider if the imbalance is a confound that will disrupt your ability to answer your research question. Remember, confounds are any statistical or research factor that makes others doubt that you arrived at a valid or correct answer to your research question. Others may panic and state that imbalances in the distribution might result in an incorrect outcome. These imbalances likely do not affect the veracity of the outcome. If the distribution across your groups is for a nominal variable that is not essential to your research question, then it is likely that not adjusting the distribution by collapsing categories and simply leaving it alone would be just fine.

This page titled [4.2.3: Evaluating the Plot](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables

When reviewing journal articles and their statistical analysis, reviewing categorical distributions is a good first step in determining how the results answer a research question. Here's a sample checklist of questions when reviewing.

1. Identify the research question and the purpose of the study.
2. What kind of population is the focus of the study?
3. Your aim is to address this question: do the sample demographics represent the population demographics?
4. Is this sample a good estimate of the population parameter? Does this sample represent the population?
5. Fourth step – review each demographic variable. Does the number or percentage in the sample represent what you expect to see in the sample?
6. The answer is (a) yes, it is expected, (b) no, it is not.
7. If (b) Is there an anomaly? i.e., Is the value too high or too low?
8. Is the category demographic underrepresented? Too many of a certain category?
9. If it is not expected, then explain how that happened. Review the recruitment procedure and review the study's setting.
10. Fifth step – conclusion. Did the sample characteristics (i.e., the demographics) represent the population?
11. Any anomalies? If so, do the anomalies matter? Would these anomalies undermine any conclusion you are trying to draw?
12. Did the researchers get a good sample representation?

This page titled [4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

4.4: Continuous Distributions

Continuous variables are the ordinal, interval, and ratio variables. The evaluation of a continuous distribution is different from a categorical distribution. Recall that continuous distributions involve numbers that now have value. Two is greater than one, three is greater than two, and so forth. With a continuous distribution, the shape of the distribution now matters. Now, the numbers mean something. If you were to take independent observations of a phenomenon (i.e., a variable) on a continual basis, eventually, you would find that the frequency of all possible observations would be distributed along a normal bell-shaped curve.

This page titled [4.4: Continuous Distributions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

4.5: Evaluating the Quality of Frequency Distributions

To evaluate the quality of a continuous distribution, you are looking for something specific to improve its quality. What you want is a normal distribution when there is a bell-shaped curve and scores are evenly dispersed around the mean.

By the way, good luck trying to find a bell-shaped curve. It is rare that you get a bell-shaped curve that represents a normal distribution. That means that even if a distribution does not look like a bell-shaped curve, it's probably close enough to resemble a normal distribution.

To evaluate the distribution of continuous variables, first create a histogram. A histogram is a graphic representation of the frequency distribution. The histogram answers the following questions. What is being distributed? The scores along the continuum of the variable range. How many participants have a particular score along the continuous variable? How do we distribute the variable? First, we distribute it along the x-axis. The x-axis displays the range of the variable or all possible scores. Who is in this distribution? Along the Y axis, the number of participants with a particular score is shown.

This page titled [4.5: Evaluating the Quality of Frequency Distributions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

4.6: Normal Distribution for Continuous Variables

Statistics is the study of variation. You want to see the kind of variation and the quality of the variation. You need to see that the variable actually does vary. The opposite of variation is constant. Constants do not help us because that means everything is the same, which does not help us predict why anything is different.

The reason you need to see a normal distribution in your variables is because it establishes what is significant and what is not significant. Put differently, the normal distribution establishes what is normal and not normal. Or what is expected and not expected. Recall that a hallmark of statistics is variation, and we get variation through comparison – “this is different than that.” To establish that something is significant, we must compare it to something that is not significant. To determine if something is unexpected, we must compare it to something that is expected. To determine if something is not normal, we must compare it to something normal. Hence, we need a normal distribution.

The other reason you need a normal distribution is because it satisfies the assumption of statistical tests. You will hear the phrase “satisfy the assumption of statistical tests.” Assumptions are criteria that are met before proceeding with a statistical analysis. Satisfying those assumptions means we can trust the validity of the statistical results. Not satisfying those assumptions means we cannot trust the validity of the statistical results. More on what this means later. One of the assumptions is a normal distribution. Briefly, if we do not have a normal distribution, one problem is that the statistical results might be due to an outlier or an extreme score. The results are due to exceptions rather than the trend of the group. These are just examples. The main point for now is that we examine the distribution of variables to determine if we can proceed with the statistical test. Normal distribution means we can proceed with the statistical test. A non-normal distribution means we cannot proceed with the test.

When you have a normal distribution, you can use parametric tests. Briefly, these are your common, well-known tests: t-test, ANOVA, correlations, regression. No normal distribution means you might consider a non-parametric test. These non-parametric tests include the chi-square.

This page titled [4.6: Normal Distribution for Continuous Variables](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

4.7: Skewed Distribution- The Opposite of Normal Distribution

The term of a distribution that is not normal is a skewed distribution. In a skewed distribution, the majority of scores are clumped at one end of the distribution, and there is an elongated tail on one side of the distribution. The term “skew” refers to debris, leftovers, extraneous scatter. The skew refers to the tail of the distribution.

There are only two types of skewed distributions. Positive skew: the tail end is on the right, where there are higher scores. Negative skew: the tail end is on the left, where there are lower scores.

An example of a positively skewed distribution would be student discipline data for an elementary grade. Most students in the elementary grades are good kids, but alas, there will always be a few kids who are troublemakers. If we plot the number of students with, e.g., Discipline data, more infractions on the right, positive or more, end of the scale. An example of a negatively skewed distribution could be self-esteem. The positive right end is usually high self-esteem, and the low end is low self-esteem. Generally, in a sample of elementary-grade students, there will be more students with high self-esteem and very few with low self-esteem. These distributions can happen, and we have to decide if we should proceed with that distribution.

Recall that normal distributions are as common as The Cubs winning the World Series. In other words, it never happens. So, if we rarely, if ever, get normal distributions, we always get skewed distributions. So, what do we do? The question is this – “if all distributions are not normal, and if this distribution is not normal, how not normal is this distribution?”

This page titled [4.7: Skewed Distribution- The Opposite of Normal Distribution](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

4.8: Evaluating a Continuous Distribution

The answer to the previous question is that we evaluate how normal or not normal the distribution is.

What are you looking for in the distribution? You need a picture —a histogram— and a table —a frequency table showing how many people obtained each score. You are looking for the shape of the distribution. There should be a bell curve and tails on each side.

First step – look at it.

The first step is to look at the distribution. If you see a normal distribution with a bell-shaped curve, then you have a normal distribution. If you see there is a “lean” or the clump of scores more to the left or right of the graph, not in the center of the graph, and if there are elongated tails on the left or right side of the graph, the distribution is skewed. When you look at the distribution, look for “zits” or bumps on the distribution. We call those bumps outliers. Think of outliers as zits — small bumps on the face. You’ll see clumps of data, then a gap, then a single bump at the end of the distribution.

This is a heuristic or visual process, and by no means is it a definitive process, but it serves as a first step to see if your distribution is close or far away from normal.

Second step – get a skewness value.

A second step is to obtain a skewness value (from your statistics program, like SPSS). If the skewness value is 0, the distribution is symmetrical, and you have a normal distribution. Positive values mean that the distribution is positively skewed. Negative values mean that the distribution is negatively skewed. A positive or negative value simply indicates direction. However, the actual value does matter. Some say look for skewness values above or below 1.0 as indications of a skewed distribution. My recommendation is to look for values above or below 5.0. More about this later.

Third step – look for zits or outliers.

A third step is to look for outliers, otherwise known as those “damn” outliers. The presence of an outlier is an indicator of a possible skewed distribution. Outliers are instances where the observation is extremely high or low compared to the mean score. An “outlier” must be a considerable outlier. Determining what is a “considerable” outlier is an analysis that is its own beast. More about this later.

Fourth step – Standard deviation greater than the mean.

A fourth step is this rule of thumb – if the SD is greater than the mean, that’s bad. This is a big rule, in my opinion. It is essentially this: $SD > M$. For those of you familiar with the mean and standard deviation, the mean is the average of the scores, and the standard deviation is a measure of the spread of scores above and below the mean.

This situation is a big deal because you get values that do not exist on the scale. If we have a mean of 3, a standard deviation of 1, and the scale range is 1 to 5, then the spread of scores around the mean makes sense because it is within the scale range. Here, the mean is 3, and the standard deviation indicates that the score variability is 2 (one standard deviation below the mean) to 4 (one standard deviation above the mean).

The situation gets interesting when you have a mean of 3 and a scale range of 1 to 5, but the standard deviation is 4. Now, what happens? Then, the spread of scores around the mean does not make sense because your upper and lower spread is outside of the scale range. Here, the mean is 3, and the standard deviation indicates that the score variability is -1 (one standard deviation below the mean) to 7 (one standard deviation above the mean). However, a value of -1 cannot exist in the scale range, nor can a value of 7. Both values are definitely outside of the scale range. So, there is no way those values can exist.

Are there situations where the SD is greater than the mean? Yes, especially when there is no upper bound, like a ratio variable, for example, cancer cells, income, or number of arrests. Especially when the lower bound is 0, where the values can only go up, can you get a negative number below the lower bound of 0? Yes, if there is meaning to the negative number. For example, financial or economics. A negative number usually means debt owed or a trade deficit.

If the SD is greater than the mean, then the distribution is definitely skewed. At this point, you would need to consult with a statistician.

Fifth step – low sample size.

A fifth step is to check the sample size. Small sample sizes tend to result in skewed distributions. Few data points mean more chances for an outlier to emerge because there are not enough points to establish a center. More data points mean more chances for scores to fill out the distribution, so there is less chance for an outlier to emerge because there will be data points all along the distribution. A sample size rule of thumb is 30 is ideal. Anything below 20 is a problem. Why? We will revisit this issue when discussing the central limit theorem.

Sixth step – is the distribution supposed to be skewed?

The sixth step is to ask yourself – “Does the skew make sense?” There are plenty of instances, variables, and constructs that are not normally distributed. In fact, I would be very surprised if some of these variables were normally distributed. For example, the number of disciplinary referrals per student in each grade. We know that most students are good kids; there are only a few not-so-great kids. Most students with zero disciplinary referrals will be on the low, zero-end. Some students will be about one or two disciplinary referrals. But then some will have, say, 10 or more. The distribution will be positively skewed because the tail and the skew will be toward the right or high end of the distribution.

It's the same scenario for the number of drinks on a given weekend on a college campus. Most students will likely have zero or one beer. Then, you get a crowd of students who will have a keg of beer — another positively skewed distribution. Take HIV viral load counts. In a given clinic, it is likely that most patients will have low HIV viral load counts, but there will be a few starting treatments with very high HIV viral load counts (and hopefully, the load will go down and recover). It would be surprising if these constructs were normally distributed. You don't want an average number of drinks on a college campus, whatever that average number will be. Same for HIV viral load. So, these skewed distributions are supposed to be skewed, so there is no need to be alarmed if the distribution is skewed; in fact, you might be alarmed if there is a normal distribution.

Are continuous distributions different for ordinal, interval, and ratio variables?

The evaluation criteria are the same for each of those variables. The variables have different propensities, resulting in a skewed distribution. Ordinal and ratio variables tend to have skewed distributions. Interval variables tend to have normal distributions.

Ordinal variables might have skewed distributions depending on how many participants tend to get certain ranks. Suppose a lot of participants get the top ranking, and the top rank is the number one ranking. In that case, you get a positively skewed distribution because most participants will be clustered toward the left at the low value, and only a few are on the right, higher ranking.

Although the number of ranks can be unlimited, most ordinal variables have narrow ranges of ranks. Sports usually have just first, second, third, fourth, and fifth place. GPA is basically five ranks: A, B, C, D, and F. In Other situations, the ranking system is not so apparent. We rank expertise as novice, intermediate, advanced, expert, and only four ranks. We rank clinical distress in three categories: normal functioning, sub-clinical stress, and clinical stress. For these variables, the distribution tends to be skewed towards one or two ranks at the high or low end. But does it mean that the distribution will be skewed to the point where it is a concern? Not likely, because with most ordinal variables, there is a low number of ranks, which means there are fewer chances for outliers and values to be in the extreme ranges. Scores in the extreme ranges tend to pull distributions toward the extreme, and skewness tends not to be severe with ordinal variables with a small number of ranks.

Interval variables are basically Likert scales. Like ordinal variables, the interval is important, and Likert scale ranges are usually limited. The common Likert scale ranges are 0 to 4, 1 to 5, 1 to 7, and 1 to 10. Like ordinal variables, the limited range means there is less opportunity for an extreme score to occur. Do skewed distributions happen with interval variables? If the range is expanded, yes it could occur. But you have to admit, there is very little use for such an expanded range, and the skewed distribution is the result of an expanded range, not because the variable itself has a distribution that is already skewed. For example, if you have a variable where the interval scale is one to 100. Let's say the construct is depression, with 100 being the most severe depression. What is the difference between 99 and 100, or for that matter, 90 and 100? Both are indicators of severe depression; the difference of one or 10 at that extreme end does not give us any additional and precise information about the nature of the depression. The sampling context matters. If the majority of your sample came from a general community, most have none to mild to moderate

depression. Will a participant from the community have a depression score in the high 90s? It is possible, and if that occurs, that the participant should be removed because the participant's depression score represents severe depression requiring hospitalization and is unlikely to share the same characteristics as the community sample. The point is this: Likert scales tend not to have skewed distributions because most Likert scales have limited ranges, and limited ranges mean fewer opportunities for extreme scores, resulting in a severely skewed distribution.

Ratio variables tend to have skewed distributions because the range is usually unlimited beyond the zero point. Under most contexts, the range is expected. The number of points scored in a basketball game, the number of alcoholic drinks one usually has at a wedding, and the number of miles a trainee travels to a practicum site. Sometimes, the variable is undefined, and the unlimited range tends to result in a skewed distribution. The number of violent acts is difficult to define, and in a prison setting, because so many acts would likely qualify as a violent act, it would not be surprising if the distribution of the number of violent acts is positively skewed.

Positive versus negative skew.

Is there any need to compare positive and negative skewed distributions? Positive skewed distributions are more likely to occur than negative skewed distributions. Ratio variables are quite common with their unlimited upper bound, which means that the distribution will likely be positively skewed because the extreme values will be found in the upper ranges.

Negatively skewed distributions are less frequent because there are very few cases where the upper bound is fixed, and most participants score in the upper bound. One example would be a Likert scale from 1 to 5 for customer satisfaction, with 1 being not very satisfied and 5 being very satisfied. Most customers give fours and fives, and few customers give ones and twos. So, the distribution would be negatively skewed because most of the responses would be on the high end, and few responses would be on the low end. In these cases, the skew would not be too severe because the Likert scale is a limited range. In most cases, you could collapse the categories, such as collapsing categories one and two into one "not satisfied" category, leave category 3 as is, and collapsing categories four and five into one "satisfied" category. The skew would be reduced because the range of scores is collapsed into only three values, and there would be no opportunity for an extreme score to occur.

So, positively skewed distributions are more common than negatively skewed distributions. But there really are no advantages to the direction of the skew or how frequently a positive or negative skew will occur.

After going through the steps, now what?

So, after your analysis, then what? You need to decide if your distribution is normal or skewed. If it is normal, you can proceed with your statistical analysis. If it is not normal, then you have some decisions to make.

Here is your answer to your question about whether the distribution is normal or skewed – Your distribution is normal enough for your purposes.

Recall that a normal distribution is actually very rare to find. All distributions have some deviations from the normal distribution. And it really takes a LOT of problems for a distribution to make it really skewed. There are no criteria such as "when four of the six steps indicate a distribution is skewed, it means your distribution is skewed." The six steps above are just ways to guide your analysis of evaluating the quality of your distribution.

Most of the time, your statistical results will be robust even if the distribution shows signs of skewness. In statistical terms, we say the distribution is robust, which means that despite having a distribution that has a few violations of normality and is possibly skewed, you probably get the same result or the same statistical test result, which is significant or not significant, whether you have a non-normal or normal distribution. In other words, you could probably use your variable's distribution as part of your statistical analysis.

Let's start from the beginning.

Your main concern is if you have a distribution that is supposed to be somewhat normal, but it turns out that the distribution is quite skewed. Could this scenario occur? Yes. Suppose you are examining the number of months that patients suffering from alcoholism stay sober after discharge. The mean number of months is usually about three months. This distribution will likely lean towards a positive skewed approach because the mean of three months is on the low end of the scale, and there will be a few patients who do

better than three months, say, up to one year of sobriety. Overall, though, the distribution is possibly close enough to be normal because you likely do not have that many instances where patients have extremely long sober periods after treatment. Is it possible to have a patient who will be sober for five years after discharge? It's entirely possible. Some patients might just do very well. It is likely that the distribution is positively skewed because the skew represents the concept that the patient just happened to do very well in treatment.

In this case, the issue is not necessarily a statistical issue. The issue might be a sample size issue. If you simply collect more data on patients who stay sober over time, you might find patients who stay sober for two years, three years, or four years. In this case, the patient with five years sober does not look like an outlier; rather, the more data collected about the patient's length of time sober will eventually fill out the distribution, so the patient with five years does not look like much of an outlier anymore. In these cases, you might have a research design and data collection solution. It is not a problem that statistics can fix.

You will encounter folks (such as faculty) who insist that you should transform the distribution when you have a non-normal distribution. Transforming the distribution means using an algorithm (e.g., math operations such as square root or logs) to reconfigure the non-normal distribution into a normal distribution. There are concerns about this approach. One concern is that it is difficult to interpret transformed data. How do you interpret the square root of a variable versus the actual values of the variable? In my experience, the transformation algorithm does not magically transform your non-normal distribution into a normal one. Usually, the statistical results from a transformed distribution compared to a non-transformed distribution are the same. If they turn out to be different, then there is a problem. What do you do if a non-transformed, non-normal distribution produces a significant result, but a square-root-transformed distribution into a normal distribution produces a non-significant result? Should we go with the significant result but with the non-normal distribution, or should we go with the non-significant results with the transformed distribution that became normal? That's a difficult decision, and there are no great guidelines for this scenario. Hence, that is why I stay away from using algorithms to transform distributions.

You will encounter folks (such as faculty) who insist that you must switch to a non-parametric test. More about non-parametric tests is in the non-parametric chapter. Suffice to say, non-parametric tests are used when the variable distributions do not meet the assumptions of a parametric test, and in this case, the normal distribution assumption was not met. Every parametric test has a non-parametric equivalent. For example, the parametric test of t-test has a non-parametric equivalent called a 2-sample t-test: Mann-Whitney test. Is this a better solution? For me, not really. Using the non-parametric tests doesn't dramatically fix anything or give you a significant result when you get a non-significant result with the non-normal distribution.

To jump ahead, if your distribution is really skewed or is meant to be skewed, you need to consult a statistician because none of the conventional statistical analyses will produce valid results. The problem is not a skewed distribution or finding the right statistical fix.

The problem is that your variable's distribution simply cannot be analyzed with conventional statistics. If you have a variable whose distribution is erratic, then conventional statistics or transformation simply won't work. For example, if you are examining the number of reported abuse incidents during childhood for abuse survivors, that distribution likely becomes very lumpy and skewed. Some survivors have only one incident; some have a few, some have several, and some have hundreds. Not all those incidents are similar. They will vary in intensity and duration per incident.

For a distribution like that, it is difficult to say what would be a better analysis, but it is highly likely that this variable's distribution needs to be analyzed using an entirely different statistical model. This idea will be developed when we discuss non-parametric statistics, but for now, parametric statistics follow what is known as the general linear model. This model says that all relationships between variables follow a linear trend: as X increases, Y increases or decreases. For a variable such as abuse incidents, it is quite likely that this variable does not follow a general increasing trend. There is no way to determine what the analysis could be here, so you will need to consult a statistician.

Conclusion

When evaluating distributions of variables, consider the following.

1. Is your distribution categorical or continuous?
 - a. If categorical, then your evaluation criteria will consist of the following:

- i. For your sample, is the distribution of the number of participants per group for a given variable like the population under consideration?
 - ii. There are only two answers to this question – is the sample similar and expected, or is the sample not similar and unexpectedly different from the population under consideration?
- b. If the sample meets the population expectation, then proceed with your analyses.
- c. If not, then there is not much you can do for a statistical fix.
 - i. You would have to collect more data and basically recruit more participants to achieve the number of participants you want for each group for the variable.
 - ii. You would have to consider the context and determine if this sample distribution is the best you can do under your research context. If so, then let's leave it alone.
2. Is your distribution a continuous distribution, for ordinal, interval, and ratio variables?
 - a. If continuous, then your evaluation is the question: “Does my distribution resemble a normal distribution?” The only two answers to this question are “Yes, it does” or “No, it does not; it is a skewed distribution.”
 - b. Follow the six guidelines to determine if you have a normal or skewed distribution.
 - c. If it is normal or the skew is not that bad, proceed with your statistical analysis or your conventional parametric analyses of t-test, ANOVA, correlations, and regressions.
 - d. If it is skewed and there are concerns....
 - e. You could transform the variable by using square root or logarithmic transformations. But as noted, I don't like this option.
 - i. Examine your research context and determine if you need to collect more data.
 - ii. Or use a different statistical analysis that is not conventional.
 - iii. You need to consult with a statistician because there are no easy solutions for this.
3. In the end.... Above all, leave it alone.
 - a. Most distributions are going to be skewed to some degree
 - b. The skewed distribution is likely not going to be a single determining factor that severely affects your statistical analysis.

This page titled [4.8: Evaluating a Continuous Distribution](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

4.9: Discussion Questions

Discussion Questions

1. What are the conceptual differences when evaluating a categorical distribution and a continuous distribution?
2. How do you evaluate a distribution for a categorical variable?
3. How do you evaluate a distribution for a continuous variable?
4. What do you do if you have a skewed distribution?

This page titled [4.9: Discussion Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

CHAPTER OVERVIEW

5: Measures of Central Tendency

Learning Objectives

By the end of this chapter, you will be able to:

- Define and differentiate between the three types of central tendency
- Explain from a conceptual perspective the problem of outliers
- Examine the characteristics of positive and negative skewed distributions

Key Terms

- Central tendency
- Outliers
- Mean
- Median
- Mode

[5.1: What is Central Tendency? Why Do We Use It?](#)

[5.2: Three Types of Central Tendency and Why We Need Them](#)

[5.3: The Mean, Median, and Mode in Normal and Skewed Distributions](#)

[5.4: Discussion Questions](#)

This page titled [5: Measures of Central Tendency](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

5.1: What is Central Tendency? Why Do We Use It?

We use central tendency because when we collect data, we have a series of numbers. We call that series of numbers a dataset. These numbers are our observations of that “it” that is varying. This series of numbers is all over the place. There are likely several of them; they are taken in different contexts, and their range can be anything. It’s a lot of information to process. What do we do?

Recall that descriptive statistics is the first step of any statistical analysis; its purpose is to describe the data. What we do is find a way to summarize all that data and organize it. This process is often called summary statistics, and it is just another name for descriptive statistics. There are no strict definitions or distinctions between descriptive and summary statistics. For the purpose of this chapter, I will state that descriptive statistics is a process of describing the variation in our data, and summary statistics is the process of simplifying and organizing the variation in our data.

Data are observations of one “it” at one instance and one point in time. We rarely rely on just one observation of “it.” Even in carpentry and clothes tailoring, we “measure twice, cut once.” We do this because we know that one number, one data point, does not represent “it.” We know that we take a temperature reading and that temperature can fluctuate within the hour. We know that one data point does not represent everything or everybody because not everyone and everything has the same score. The number that represents “it” can fluctuate with another observation of “it.” That’s why we “measure twice, cut once” because our measurement of “it” can be slightly off. We usually take multiple observations of “it,” and that process results in a dataset.

What do we do with that data? Because that is a lot of data to process, we find a way to do something with it. What we do is summarize. Or reduce all these numbers to just a few numbers. We want those numbers to describe the variation. It is just easier to understand a few numbers than to look at all the data at once.

The overall family of values we use for summary statistics are a) central tendency and b) measures of dispersion. Remember: A parameter is a characteristic that describes the normal distribution. Instead of looking at pictures of distributions or a sea of data, a simpler way of describing the distribution is to obtain two values, the central tendency and dispersion. The common terms for central tendency and measures of dispersion are a) the mean and b) the standard deviation. In this chapter, I will focus on central tendency.

Why do we use central tendency?

Central tendency is a statistic that tells us where the scores of a sample are converging. Just to mention here, statistics are values that describe our data. In this case, the value of the statistic describes something about our data. The key issue with central tendency statistics is the term “converge,” which means the tendency, the trend, of where we expect scores to be.

Central tendency helps us do several tasks in statistics.

The central tendency is regarded as the middle, the starting point. Recall that statistics is about the study of variation, and the only way to get variation is to compare “it” with another “it.” You compare something to something. In this case, what we are doing is comparing our observation with the central tendency or the mean. The mean is our starting point for comparison. That’s one reason why we are keen on establishing the central tendency in our sea of varying data.

The measure of central tendency allows us to use one number, for example, the mean, to indicate the performance of the group, otherwise known as the sample.

The term “central tendency” is important as a description of variation because we need the central tendency to establish a focal point in the variation. This focal point is important because most people have scores around the central tendency or the mean. The area where the most scores are at is the hump in the middle of the normal distribution. It is where we will see the majority of scores clump around the mean.

With this focal point as our central tendency, we can address many questions. We need that focal point to indicate where most people will have scores because it helps us determine the significance. The focal point, or central tendency, is our way of comparing what is expected and what is unexpected. The mean is the point where we find the expected scores. Very few people score much higher or lower than the mean.

The central tendency is good for prediction or forecasting. If we do not know what will happen next or what the next data point will be, it is good to use the mean as a predictor of the next data point. The mean is a good summary of what happened before, and what

happened before predicts what will happen next. In terms of sampling, the mean is a good indication of how the sample performed, and the mean is a good indication of how the next participant we sample will perform. For example, if we take a sample of fifth graders in a classroom and measure their reading level, we produce a mean reading level score. Then, the next fifth grader enters the classroom, and we measure their reading level. Without any other information at our disposal, our best guess, or estimate, of what the reading level score of that fifth grader will be is the classroom mean score. Central tendency represents our best guess as to what score will occur, which is the central tendency score (e.g., the mean).

The central tendency is good for making comparisons, usually across groups or at different points in time. We can simplify the comparison of two more groups by comparing their central tendency or mean scores. Instead of comparing everyone to everyone, just compare the means. Compare the mean of group A to the mean of group B. Simple. We use this process ubiquitously in psychology research; we compare the means of males to females, treatment versus control, and people who have trauma to people who do not have trauma. We use the same procedure for making comparisons over time. We compare the mean of a group at time one to the mean of the group at time two. Just like for groups, we use this process in psychology research. We want to see how a group performs at intake, then at discharge, then three months later as a follow-up. We want to see how adolescents develop their emotional awareness in sixth grade, then eighth grade, then tenth grade, and then at graduation. The point is this: To simplify your understanding of psychological research and its findings, all we do when we compare groups or something over time is compare the mean scores. Nothing more.

This page titled [5.1: What is Central Tendency? Why Do We Use It?](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

5.2: Three Types of Central Tendency and Why We Need Them

There are three types of central tendency: mean, median, and mode. You are likely familiar with the definitions of each:

- The mean is the sum of all the values divided by the number of values.
- The median is the midpoint of all the values.
- The mode is the most frequently occurring value. There are pros and cons to each of these values, and these pros and cons are the reasons why we have all three.

5.2.1: The Mean – the Most Honest Central Tendency

The mean is calculated by taking all the values and dividing it by the number of values. Done.

What is the implication of this calculation procedure? The calculation means that one of the pros of the mean is that the mean scores use all the data. In other words, it is “honest” because it accounts for everything. No data are left out. Everything counts.

The con of the mean is that because it does not involve all of the data, the mean is influenced by outliers. Outliers are scores at the extreme high or low ends. The outlier is like a magnet; it pulls the mean towards the outlier. The outlier then distorts the mean because it is not an accurate or best representation of the overall performance of the sample. Rather, the mean represents the performance of the outlier and does not really represent the group.

This situation is common. We see in neighborhoods most of the houses are in a certain price range. But then someone comes along and builds a mansion-style house in the neighborhood that drives up the average price of the homes. In a college course, the students dread being graded on the curve because there is always someone who thinks they are so smart and gets a high grade and raises the curve, creating misery for everyone else in the course. In baseball, we have the batting average of the team, but there is one superstar. These days, there are sometimes two superstars with high batting averages that boost the overall batting average of the team. Outliers are found in many situations that artificially drive up the mean, making the mean less of a representation of the group and more of the outlier.

In actuality, the presence of the outlier may be permissible from a conceptual standpoint. Continuing with the baseball example, a team does have one superstar. The opposing team has to contend with the overall batting average of the team, plus the superstar. In a given classroom, there is always a standout student. It is not like the standout student is an anomaly. Although I do not know anything about real estate, if a person moves in with a high-priced mansion that drives up the housing prices, that situation is likely to occur, and maybe as a trend in the real estate market, you want property values to increase, because you want your property value to increase.

In psychology, outliers are possible. On a given hospital ward, there is one patient who might have severe symptoms far beyond those on that ward. Why is that patient there? Likely because there is nowhere else to admit the patient. On a given night for crisis calls, most of the calls might be routine, but a call might be received where the crisis is far beyond the usual crisis procedures. Conceptually, the issue and the context likely result in outliers. The presence of outliers may actually be expected as part of the sample that gets drawn from the population and phenomenon of interest.

The outlier is usually so far above or below the mean that it conceptually does not make sense to include the outlier in the sample. For example, I do not know anything about HIV issues, but if we were sampling HIV viral loads in patients, and if the clinic usually provides care to patients with moderate HIV viral loads, if a patient arrives with a viral load that is extremely high, in the millions, then the patient’s data should not be included in the mean score. The patient might represent a treatment-resistant case or an unusual case of HIV, and the usual treatment protocols will not apply. However, if a patient has a viral load in the upper thousands, it might be high, but it does not exceed the range of expected levels and could be within the confines of the facility’s treatment offerings. If we were sampling persons convicted of driving under the influence or drunk driving, and the usual range of cases involves one to 10 convictions, and we obtain a person who has 300 convictions, then that person should not be included in the mean score. Someone who has 300 convictions over 365 days is clearly off the charts in terms of drinking and driving behavior. The best way to evaluate if an outlier is adversely affecting the mean score is to do so from a conceptual perspective.

If there is a huge outlier, you wonder whether the outlier really belongs in the sample. If a patient arrives at a hospital ward and is extremely belligerent and psychotic, perhaps that patient belongs in an entirely different facility and should not be counted as part of the sample of patients on the hospital ward. If a crisis call comes in and the authorities need to be notified because it is a medical

crisis, not a psychological crisis, then that call event should not be considered part of the parameters of the crisis call hotline. The decision to remove an outlier, because it does affect what the sample mean represents, should be based on conceptual reasons, not just because it is an outlier per se.

In reality, outliers really don't dramatically damage the outcome of a statistical test. Notice that I stated the outcome of a statistical test, not the mean score. The mean score might be affected, but when these scores are entered into a statistical test, the results usually do not dramatically change. Honestly, I do not have any citations or statistical proof for my assertion. All I have is years of experience, and every time I receive a dataset, and there is an outlier, despite everyone's concern about the presence of the outlier, I find when I conduct my statistical tests in every way possible to account for the outlier, the results do not dramatically change. I believe that when there is a sample that is a sufficient representation of the population, it really takes a huge outlier that is dramatically different from the population to shift the statistical result into something completely different. It is just simply rare for one single participant with a high outlier to completely alter a result.

The bottom line is that people think that outliers are an invalid representation of the sample, but in reality, this perception is really overblown.

Back to the mean. The mean is the most ubiquitous statistic. We use the mean because it does use all the data, and we want to ensure that we are accounting for all the variation when calculating our statistics. It is used all the time and is a common statistic in most statistical analyses and equations.

5.2.2: The Median – More Dishonest than the Mean

The median is the point of distribution at which an equal number of scores (50% of scores) are above and below the median point. The median is not based on the score values; it is the number of scores you have. The median is the point along the continuum of numbers where there are 50% of scores above and below that number. That point is not necessarily one of the numbers in your set of data.

The median, then, is typically useful when there are outliers and you do not want the outliers to influence the value or the central tendency score that you want to represent your group or your sample. Essentially, the median ignores the outliers. Because the median ignores the outliers, you are manipulating the data by dropping information; you are not using all the data or information in the variable. The median is the opposite of the mean because while the mean uses all the data, the median does not. In this way, the median is more “dishonest” than the mean because the median ignores the extreme scores.

In practice, the median is used mostly for summary statistics purposes. In contexts where you want to ignore the extreme scores, the median is often reported rather than the mean. The median is often used for summary statistics only. I have never encountered a situation where the median, instead of the mean, was used as part of a statistical test computation. Instead, the mean is what we use for statistical test computation.

The median is typically used for variables with no upper or lower limits. This means that scores can be anything, which means extreme scores can occur. This situation occurs when there are ratio variables because, as you recall, there is a zero point, and the values can be unlimited. For example, length, money, weight, and time have no upper limit. What typically occurs with ratio variables is that there are gaps all along the continuum. The number of beers drunk can have a gap in the continuum. Most guests at a wedding just have a few drinks, such as five, but then there is one drunk uncle who has way too many at the open bar, such as 20. There will be a gap between drunk uncle and wedding guests in the number of drinks. So, instead of using the mean, which will include the drunk uncle's drinks, using the median is helpful here because it will not account for the extreme score or the drunk uncle.

Interval and ordinal variables rarely use the median because the ranges of those variables are fixed. For example, interval variables are usually Likert Scales, and we often say, “On a scale from one to five, rate your satisfaction.” The range is limited from one to five, and outliers are likely to occur. Ordinal variables typically have few and fixed ranks, such as first, second, and third place, or grades, such as A, B, C, D, and F. So, it is unlikely an outlier will occur in terms of an observation that is far away from the rest of the group because these variables have a small range.

Using the median because there are outliers or a skewed distribution might signal something wrong with the distribution or with the nature of the variable under investigation. We will discuss skewed distributions in the next section. For now, the median is not just a solution to representing a skewed distribution. Something might be awry with the distribution if you expected the distribution to

be normal, but it turned out to be skewed normal, and you need to use the median rather than the mean. Using the median will not solve your problem of having a skewed distribution. A better approach is to examine the distribution and decide if it is meant to be skewed or if something is awry with how the data were collected. In this case, contacting a consultant would be best, especially if you thought the distribution was supposed to be a normal distribution.

5.2.3: The Mode – the Most Useless Number

The mode is the most frequently occurring number in the distribution. It is a useless statistic. I have never used the mode, I have never reported the mode, I have never discussed the mode, I have never read about the mode in research journal articles, I have never used the mode in a statistical calculation, I have never seen the mode except in statistical textbooks. It is totally worthless. Just like watching a Cubs game, it is a total waste of time. OUCH

The only time I needed to learn about the mode was when I had to answer questions about it in the statistics section of a licensing exam. The mode becomes an issue when answering questions about normal and skewed distributions. That is all you need to know about the mode.

This page titled [5.2: Three Types of Central Tendency and Why We Need Them](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

5.3: The Mean, Median, and Mode in Normal and Skewed Distributions

In a normal distribution, the mean, median, and mode are at the same point in the distribution; stated differently, the mean, median, and mode have the same value.

Normal distributions are rare because most distributions have imperfections in their distribution, and they become skewed. The skewed distribution is the reason we have three measures of central tendency. There are two skewed distributions, positive and negative.

A positively skewed distribution will ALWAYS have the following characteristics:

- The tail is on the high end and to the right of the distribution
- The order of the central tendency values (from left to right): Mode, Median, Mean.
- Mean is pulled to the high end, where there are high extreme scores
- The mean is higher than the median
- The median is the midpoint. Think of the median as ‘in the middle of the road.’ The median is always in the middle between the mean and the mode.
- The mode is the highest point in the distribution, or the “hump” in the distribution

In a negatively skewed distribution:

- The tail is on the low end and to the left of the distribution
- The order of the central tendency values (from left to right): Mean, Median, Mode.
- Mean is pulled to the low end, where there are low extreme scores
- The mean is lower than the median
- The median is the midpoint. For example, the median is in the middle of the road. The median is always in the middle between the mean and the mode.
- Mode, which is the highest point in the distribution, or the “hump” in the distribution.

Figure One depicts a skewed distribution with the mean, median, and mode. The positively skewed distribution is on the top, and the negatively skewed distribution is on the bottom. Identifying skewed distributions was addressed in Chapter III. For now, you only need to know the characteristics of the positive and negative skewed distributions. The only reason you need to know this information is for your knowledge base for a licensing exam that addresses this topic. Otherwise, there is little need for you to apply this knowledge when conducting statistics.

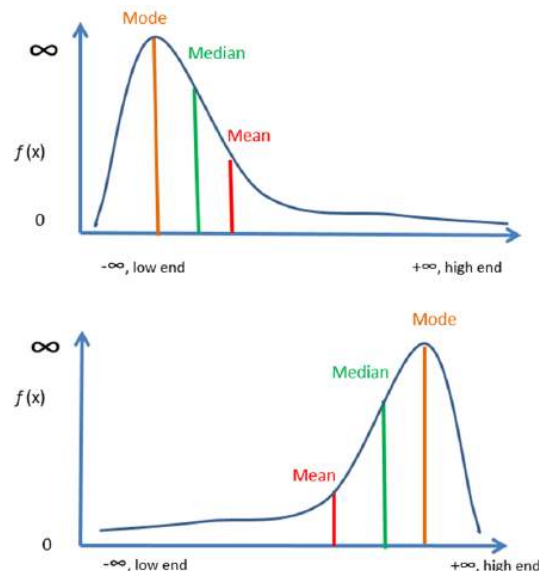


Figure One: *Positively Skewed Distribution (top figure) and Negatively Skewed Distribution (bottom figure) With Mean, Median, and Mode.*

Conclusion

The central tendency is an important statistic because it establishes the point of comparison. Statistics is all about comparison, and we need a central tendency to establish how to compare something to something. Establishing that central tendency is not easy, and the presence of outliers, leading to skewed distributions, means that we need three versions of central tendency. We have the mean, median, and mode, and there are pros and cons to using each mean or the median, except the mode, which is completely useless.

This page titled [5.3: The Mean, Median, and Mode in Normal and Skewed Distributions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

5.4: Discussion Questions

Discussion Questions

1. Why is it important to establish a central tendency?
 2. What is an outlier, and are outliers really a problem?
 3. What are the three types of central tendencies?
 4. What happens to the three types of central tendencies in skewed distributions?
-

This page titled [5.4: Discussion Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

CHAPTER OVERVIEW

6: Measures of Variability

Learning Objectives

By the end of this chapter, you will be able to:

- Understand how to determine variance
- Differentiate between error variance and true variance
- Explain the statistical measures used to determine the location of scores

Key Terms:

- Variance
- Standard Deviation
- Range

[6.1: What is Variance?](#)

[6.2: How Do You Determine Variance?](#)

[6.3: Partitioning of the Variance](#)

[6.4: Standardized Values of Reporting Scores](#)

[6.5: Discussion Questions](#)

This page titled [6: Measures of Variability](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

6.1: What is Variance?

We use the term mean (central tendency) to describe the general trend of a sample's performance. By performance, we mean how we predict what is going to happen next. We continually sample, make observations, or recruit more participants; we say that the next time we observe, our best guess about what will happen next, or what the next participant's or observation score will be, will be the mean.

But the mean does not give us the full story. Not everyone in a sample has scores at the mean. We need a statistic to describe the spread of observations around the mean in that sample. That statistic is called variance.

Variance is a fancy statistical technical term that represents variation. Think of variance as a value describing the spread of scores around the mean. By "spread of scores," we mean a score's distance from the mean or the location of a particular score away from the mean. The variance values include the following terms: variance, standard deviation, and range.

Why should you give a rat's ass about this concept? Eventually, you will learn that variance is an important concept for understanding why we have significant or non-significant results. Variance helps us determine if something is significant or not significant.

This page titled [6.1: What is Variance?](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

6.2: How Do You Determine Variance?

To determine variance, think, “How do we determine the spread of scores around the mean?” or “How do we determine how far these scores are away from the mean?” To determine variance, we start with deviation scores. A deviation score is a score’s value minus the mean score. We calculate a deviation score from every value in our variable and determine how far away it is from the mean.

The sum of deviation scores is always zero. Deviation scores are always deviations above and below the mean, so those scores will cancel each other out. Hence, the sum of the deviation scores is always zero. That is not helpful for determining variance. To solve the problem of deviation scores cancelling each other out, we take each deviation score, square it so the score takes on a positive value, add them, and then divide by the number of scores.

Notice that we do not take the absolute value of the deviation score; we square the deviation score. Why? Because, for some reason, the absolute value doesn't work in mathematical equations for statistics. Go figure...

The better reason for this is— that variance consists of the mean sum of squares of the deviation scores because we square them because we want the area underneath the curve, otherwise known as the normal distribution. You are going to hear and read a lot about the sum of squares. The sum of squares refers to the deviation scores.

Recall from geometry that area is calculated by length times width. Or, in the case of a square, length squared. So, we square to calculate the area.

Why do we calculate the area? Because we are interested in the area underneath the curve or the normal distribution. The area is the percentage of scores that are expected to fall under that section of the distribution.

Why is this important? Fast forward to chapter eight on significance. Knowing the area underneath the curve is important because:

- Scores that have a high percentage are likely to occur.
- Scores with a low percentage, perhaps less than 5% of the time, are less likely to occur strictly by chance.
- We are interested in those scores that do not occur that often, such as less than 5% of the time.

As you can surmise, calculating the area underneath the curve is part of the process of understanding which results are significant. We know if a result is significant if it does not occur that often. How often? We say an observation is significant if it only occurs 5% of the time or 5% of all of our observations. That percentage sign is basically a proportion of the area underneath the curve. We calculate area by squaring the deviation scores, which is basically the length of each “square” underneath the curve. Fast forward, if you recognize that 5%, it is because it is ubiquitous, $p < .05$. More on that later in chapter eight on significance.

Knowing that variance is calculated based on the sum of squares and assuming a normal distribution, we know that the following percentages occur underneath the normal distribution.

The reason we make a big deal about determining if our distribution is normal is that these percentages are only valid if the distribution is normal. These percentages are necessary for establishing what observations are significant or not significant. If the distribution is not normal, then we cannot establish these percentages. Some other model of our variable’s distribution is needed, and that means you need a completely different set of statistical analyses, and this book will not address those analyses. And that is when you need a statistical consultant.

This page titled [6.2: How Do You Determine Variance?](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

6.3: Partitioning of the Variance

You will see the phrase “partitioning of the variance.” What that phrase means is dividing or separating the variance into parts. Understanding why variance is partitioned is important for understanding statistical significance. We always divide the variance into two main parts: **error variance** and **true variance**.

6.3.1: True Variance

The first source of variance is “true” variance. True variance means that the scores representing the phenomenon of interest vary because the phenomenon itself does vary. Take height. We measure height for a set of 10 people, and we have 10 different scores because the height of each person does vary. In psychology, we measure depression. If we assess depression for a set of 10 people, we have 10 different scores because the depression for each person does vary. Some people are more or less depressed than others. In both cases, the scores should vary because they reflect differences in the phenomenon. This situation is an indicator of true variance. The scores vary because the phenomenon varies. Fast forward to the next section, true variance is the opposite of error variance because if the scores vary, the scores might vary because of something else, other than the actual phenomenon varying. Hence, the opposite of true variance is error variance, which means that something else is making the scores vary.

True variance is otherwise known as actual variance, real effects, real differences between groups, and real relationships among variables. Sometimes referred to as between-subjects variance, true signal, strong signal, or a detectable pattern. Think of true variance like a radio dial. Before the internet streaming service, in my day, we had car radios. We had to tune in past the static to find a clear signal. That clear signal is a “true” variance, indicating that you found something. Everything else is noise, static, and error variance. How much true variance we will likely obtain is discussed in later chapters.

6.3.2: Error Variance

In contrast with true variance, error variance means that the scores vary because of something else that is varying and not due to the phenomenon that is varying. Take height. We measure height for one person and get a score that represents how tall that person is. However, we can take another measurement of that same person and get a slightly different height. How does this happen? We might place the ruler, which is on the floor, not quite at the start. We might have the person take off their shoes. We might measure up to their head, but if they have a lot of hair, finding the top of their head might be a challenge. We might put their body against a wall to make sure they are standing up straight. These idiosyncrasies can result in the measurement of height, producing slightly different numbers.

In psychology, we measure depression. If we assess depression for one person at a given time, we get a score that represents their depression. But if we take another measurement of their depression on a different day, the person’s mood might lift, or the White Sox win a game, and they are happy, or we just happened to catch them after getting a good night’s sleep. Or we give them a different measure of depression, and they obtain different scores because the items are different. Or we provide them with a measure of depression that was framed according to a non-Western perspective rather than a Western perspective, and their depression might not be as severe as the original measurement. Or the person has a bad attitude toward taking a depression scale and provides random answers. Or a person could misread an item and not answer as they normally would. Or two psychology interns can rate the severity of the person’s depression and have two different ratings, or scores, that represent the depression. There is a myriad of ways the scores can vary, and the variations have nothing to do with the construct of depression actually varying. Something else was responsible for varying depression scores.

With error variance, something else can result from random effects, mistakes, systemic errors, or unknown errors. What we try to do in statistics is determine how much of the variation in scores is due to true variance and how much of the variation is due to error variance. Then, we try to determine how much of that error variance is due to pure randomness or something systematic. Trying to divide or determine how much of the error variance is due to various sources is the process of partitioning the variance.

Random effects can refer to random chance, pure error, no pattern, or within-subjects variance. There will always be an error or “noise.” There will always be data or observations that we cannot account for because they happen due to random chance alone. In other words, people get lucky. The basketball luckily bounced into the net. The Cubs player hits a ball.

Error variance is also referred to as an unknown error. Put differently, error is something we cannot predict. Remember that prediction always involves two variables: the independent and the dependent variable. Prediction can be expressed as: “knowing

the level of the IV, we can predict the level of the DV.” By knowing how many hours you study for the test, I can predict what grade you will get on the test. But you know that not every outcome follows that prediction. Someone can study for one hour, or not at all, and get an “A” on the test. So, that observation is an error because our prediction failed to predict what grade the person should have received. So, error in this case is not just random luck; it says that we failed to make the right prediction. Unknown means something else is going on which accounts for the outcome. For a given outcome, we try to get all the variables we can to predict an outcome. We cannot account for everything, though. Sometimes, we do not know what other variables could account for the outcome; sometimes, we cannot collect the data for these variables; sometimes, we do not have a measurement method for these variables.

Think of this problem as trying to decipher a mysterious recipe. We taste a delicious dish and then work backwards to figure out what ingredients are in that dish. We could guess most of the ingredients, but we might not guess the secret ingredients. Or we might not know what cooking technique was used to create the dish. These represent the unknown. In psychology, we observe a client in mental health stress. We know about the usual stress inputs. We may not know about other stress inputs, such as cultural stigma, past abuse history, environmental stressors, and concerns about gender or sexual orientation identities. We cannot collect data on everything, and we are limited in how we measure these variables. Anything we cannot account for is an unknown variance or error variance.

In psychology, error variance is abundant because so many variables affect a client’s well-being and progress toward recovery. This situation is nothing to fret about because it is a normal occurrence and expected, given the state of how we measure variables related to well-being. Later chapters discuss how much error variance we are willing to accept.

The other source of error is mistakes, which are not random. Mistakes include miskeyed data, misreporting data, or participants carelessly responding to survey questions. Some mistakes are akin to “Murphy’s Law,” where accidents happen, and something goes awry in data collection. Other mistakes are due to an error in the data collection process, such as forgetting to include an item or someone forgetting to include a test in the test battery. Mistakes are less of a statistical issue and more of a research design issue, specifically, a data collection issue. There is not much you can do about this error from a statistical standpoint. Better off reviewing the data collection to see if the errors can be removed.

Systemic error is another source of error. Systemic error is something that is consistently going wrong, and it is usually a research design or method issue. For example, the researcher collects their own data and is biased towards seeing the results they want to see. Others are research design confounds such as the assessment practice effect. These issues are best addressed as research method issues. Here, too, there is not much you can do about this error from a statistical standpoint, and a review of the research method is warranted to address this error.

Recall that a hallmark of statistics is that a statistical test is basically a ratio. This ratio consists of truth over error. Stated differently, we want our true variance to be larger than the error variance. As a ratio, the true variance is always on top, the numerator of a statistical ratio. The error variance is always on the bottom, the denominator of a statistical ratio. Basic math operations state that when the numerator is larger than the denominator, we get a number greater than one. For now, the lesson learned is that statistical tests, and the numbers they generate are basically ratios. The ratio consists of the true variance, or signal, or pattern, over the error variance, or noise, static or randomness. Larger numbers mean there is truer variance, or pattern or signal, and less error variance, and that result is good. How large is that number, and if it matters, are issues discussed in later chapters?

This page titled [6.3: Partitioning of the Variance](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

6.4: Standardized Values of Reporting Scores

The reason we learn about standardized values for reporting scores is for conversion purposes. We take a person's raw score and convert it into z-scores, t-scores, and percentiles. For those taking the psychology EPPP exam, there are questions that ask you to compare which value, z-score, t-score, or percentile, is higher or lower.

6.4.1: Standard Deviation

The standard deviation is a ubiquitous statistic. When you see a mean, you should see a standard deviation value reported alongside the mean. Both values, in concert, provide good information about the quality of the distribution of the scores for a variable.

The formula for the standard deviation is the square root of the variance value. The variance value is expressed as squared units or sum of squares. It is hard to interpret what a good high value is. The variance value is based on the measurement system of the variable. If a variable is a Likert scale from one to five, it will have a different variance value compared to an interval variable such as one to 100. So, a variance value of 50 cannot be compared to a variance value of 500 because the values are based on different variable value ranges.

Taking the square root of the variance gives us the average deviation of scores from the mean. This value is a metric that is used as a gauge to measure how spread apart the scores are around the mean. The metric allows for easy communication about the spread of scores or the distance from the mean. Think of the standard deviation as a distance measurement unit. It tells us how far the score is from the mean. The standard deviation is an easy communication system. Standard means that everyone knows what you are talking about. That is why we call it a "standard" deviation. The term standard means that everyone agrees on what it is and its properties.

These are the two common ways we use standard deviation: location and range.

Location, location, location.

First, we use the standard deviation to communicate the location of a variable's score. When you see this set – Mean = 10, SD = 2. What do you do?

First, what you need is a particular score. Your goal is to determine the location of a score. If someone tells you that a score is 12, you want to know how you interpret that score. Is that a good or bad score? To do so, you take the score and determine if it is above or below the mean. Add the SD to the mean; in this case, you get 12. The score under question is one standard deviation above the mean. That seems like a good score because the person's score is higher than the mean by one standard deviation. If someone tells you that a score is 8, is that a good or bad score? You take the score and determine if it is below the mean. Subtract the SD from the mean; in this case, you get 8. The score under question is one standard deviation below the mean. That seems like a bad score because the person's score is lower than the mean by one standard deviation.

If someone wants to know what a score would be if someone scored two standard deviations above the mean, then you add the standard deviation value to the mean. In this case, you add two standard deviation values, and $10 + 2 + 2 = 14$. For scores that are two standard deviations below the mean, subtract the values: $10 - 2 - 2 = 6$.

We use the standard deviation to communicate if a particular score is above or below the mean. All values of the SD are + or - because you can go above or below the mean. The sign of the standard deviation communicates the location above or below the mean.

The number of SDs communicates how far away a score is from the mean. We use the number of standard deviations and the value of the standard deviation to communicate how far the score is from the mean.

For example, a given score of 12, with a mean of 10 and an SD of 2. That score is 1 SD above the mean. A given score of 6, with a mean of 10 and an SD of 2. That score is 2 SD below the mean.

Notice the notation of the standard deviation. This, SD = 2, is different than this, 2 SD. This notation, SD = 2, tells you the value of each standard deviation unit. This notation, 2 SD, tells you how many standard deviation units a score is away from the mean.

Think of SD = 2, like using different measuring systems. For instance, an inch is this long —. But a centimeter is this long —. Both are useful, but we must decide on what unit we are using. In baking, we often communicate using cups or tablespoons.

However, advanced bakers advocate using a scale and weighing ingredients using grams. The reason is that measuring a cup of flour is not quite the same amount as the precision of using grams of flour (hint: Google the reasons). The point of this baking example is that we have to decide on what measuring unit we are going to use. In some ways, the notation $SD = 2$ just indicates that the unit we are using to measure the distance of a particular score is away from the mean; the “2” means that we are basing this distance unit according to the variable range and the variance of the scores. For a variable with a range of 1 to 10, an $SD = 2$ is based on the range and the variance or spread of scores around the variable’s mean. So, for a variable with a range of 1 to 100, $SD = 20$ is based on the range and the variance or spread of scores around the variable’s mean.

The notation of 2 SD is different because it is part of a communication of the location of a particular score. We never say “2 SD” on its own. We say, “A score is 2 SD above the mean.” We usually want to know how far a particular score is from the mean to evaluate the score’s distance. We also want to know what a score would be if it were one, two, or three standard deviations above or below the mean. Indicating how many SDs is an indication of the location of the score.

Why do we want to know about the location of a score? The location of the score is an indication of distance away from the mean. The distance away from the mean is part of the process of determining significance, and this issue is addressed in chapter eight. For now, we want to know if the score of interest is well beyond the mean. Put another way, we want to know if the score is interesting (significant, rare) because it is well beyond average (something that is typical and frequent). A particular score is significant if it generally falls two SD above/below the mean. A particular score is not significant if it generally falls within one SD above/below the mean.

What do we mean when we say the location of the score? Recall that all values, or scores, are placed on a continuum. For continuous variables, this continuum is from low to high. It is always low to high. We always discuss scores in terms of how low or high they are relative to the starting point.

When we say the location of the score, we mean where the score is located along the continuum along the X-axis of an X-Y frequency chart. Location is important when we discuss significance. Fast forward to the significance of a statistical test, when we determine significance by determining if a result rarely occurs and if it is far away from the mean. When we discuss how far a score is away from the mean, we mean the location and use the standard deviation to determine exactly how much a score is away from the mean.

Does it matter if the standard deviation value is small or large? There is no inherent value if the standard deviation is small or large. The value matters if you want a homogeneous or heterogeneous sample. High standard deviations suggest considerable variability in the sample. High variability à heterogeneous sample: The higher the SD, the greater the spread and wide distribution. Low standard deviations suggest less variability in the sample. Low variability -> homogenous sample.

Studies that describe the extent of an issue usually have heterogeneous variation. We want to see that everyone is different from each other and that there is a range of differences. In descriptive studies, such as epidemiological studies, you want a wide standard deviation indicating a heterogeneous sample. For example, discovering that there is a considerable range of people who need services, such as students who experience stress over homework. You want to see that there is a range of students who have some, more, or considerable stress over homework. In correlation studies, you want a high standard deviation, indicating a heterogeneous sample to establish patterns between variables. For example, those who suffer from micro aggressions and low social support also have mental health concerns. Yet, those who suffer from micro aggressions but have high social support have fewer mental health concerns. You cannot establish this pattern unless you have variation.

The instances where you want studies to have a small standard deviation value usually come from comparing groups, especially treatment and control groups. You do not want those groups to overlap; you want them to be distinct, so you want each group’s variation to be small. For example, to examine a treatment effect, your treatment and control groups at the pre-test likely have similar sizes in their standard deviation values. At the post-test for the treatment group, you want everyone in the treatment group to have a positive outcome. For example, you want everyone who completed the mindfulness treatment to come out more centered and better able to handle stress. You want your treatment group to be homogenous, hence a low standard deviation value.

There are other uses for evaluating the size of the standard deviation value.

You might have wondered what happens if the SD equals zero. If the variable’s $SD = 0$, there is no deviation or spread. The variable is a constant. Completely useless. This situation never happens because there usually is some variation in every variable.

- Recall that if the SD value is greater than the mean, that is a red flag that the distribution is not normal.
- If the SD value is just about the mean value, that suggests wide variability in the sample. You might consider whether the distribution is normal or not.
- Or take the value of the mean and divide it in half; if the SD is about half the mean, that also suggests wide variability.
- If the SD value is less than $\frac{1}{4}$ of the mean, that suggests little variability.

Why do we take the square root of the variance to calculate the standard deviation?

The answer to that question is mathematically complicated. Maybe one way to approach this question is to think about how we want the standard deviation units to be equal in length. Although dividing something in half would provide that result, continuing to divide it in half may not produce units that are equal in length. So, it is easier to take the square root to achieve that goal.

If you bring home a pizza, you want it all to yourself. If your roommate wants some pizza, you will take the square root of the pizza or divide it in half. Then, if your roommate brings their partner to the impromptu pizza party, you will have to figure out how to provide equal proportions. Taking the square root of the pizza would do just that.

The reason why we have different values for the standard deviation is because no matter the size of the pizza, taking the square root will give you equal proportions. However, each of those proportions differs in size depending on the size of the original pizza. However, the square root keeps all these proportions equivalent, no matter what size of pizza it is. Hence, the square root keeps all the standard deviation units equivalent, no matter the size or the number associated with the variance.

Range

The range is a simple way to describe variance. It is based on the highest and lowest scores for a variable. Simple as it is, it has its uses.

The range is good for describing the variability of continuous variables without limits. It is good for studies where you need the upper and lower limits to define the sample. Age, height, and weight are important variables for diet and eating disorder studies, and it is useful to know the lowest and highest values for each of the variables to understand the outcome of these studies. In medical studies, variables such as HIV viral loads and blood cell counts often report the lowest and highest values in their sample. Studies about incarceration have the lowest and highest number of criminal acts to contextualize the severity of their sample.

6.4.2: Z-scores

Z-scores are used to communicate the location of scores. To find out how far a score is from the mean, you need three pieces of information: the score, the mean, and the standard deviation. With a z-score, all you need is the z-score. Simpler....

If a score is 1 SD above the mean, the z-score is +1; if it is 2 SD above the mean, the Z-score is +2; if it is 3 SD, the Z-score is +3. If a score is -1 SD below the mean, the z-score is -1; -2 SD below the mean, the Z-score is -2; 3 SD, the Z-score is -3.

Z-scores transform raw scores. Z-scores are used to express (transform) raw scores in terms of standard deviations. Raw scores are those scores where you cannot understand the meaning of the score based on the score alone. $X = 20$. What does 20 represent? I have no clue; you need context. The high score, the low score, the mean, and the SD. But if I said the raw score of 20 is transformed into a z-score of +2, then you know exactly what's going on. You know that this score of 20 is two standard deviations above the mean. In other words, it's a high score, possibly a significant score.

The common use of z-scores is that they allow us to compare different distributions with different scale ranges. Z-scores are useful because different distributions from two or more variables have different means and standard deviations, different ranges, and low and high scores. There's no way to compare them directly. For example, we have many ways to measure depression. The Center for Epidemiological Studies Depression Scale is 0 to 60. The patient health questionnaire, PHQ-9, PHQ-9 score can range from 0 to 27. How do you compare a CESD 50 with a PHQ-9 of 17? Another example is Academic aptitude. The SAT score range is 400 to 1600, and the ACT score range is 1 (low) to 36 (high). How do you compare SAT 1000 with ACT 27?

The solution is to standardize them so that all the scores' distributions have a mean of 0 and an SD of 1. Using the same metric allows for easy comparison. It allows us to get a quick picture of how well above or below the mean the score is. Allows for better communication of scores when the score is standardized.

Using our depression example, for Z-scores, how do you compare a CESD 50 with a PHQ-9 of 17? CESD Z score is +3, and PhQ z-score is +2. A person has a higher score on the CESD compared to the PhQ. Academic aptitude: The SAT Z score is +3, and the ACT Z score is +2. A person has a higher score on the SAT compared to the ACT.

To calculate Z-scores, take the score of interest, subtract it from its mean, then divide it by its standard deviation. For the Self-Esteem Scale, the range is from 20 (no self-esteem) to 80 (high self-esteem). Sample $M = 63$, $SD = 5$. The raw score and individual client score are 76. The score is over 2 SD away from the mean. Self-Esteem Raw score converted to Z score: $(76 - 63)/5 = 2.60$ Z score. Just say the z score is 2.60, and you know exactly where that score is located and if it is significant.

For the Satisfaction with Life Scale, the range is from 5 (not satisfied) to 35 (satisfied). Sample $M = 20$, $SD = 4$.

The raw score and individual client score are 29. The score is over 2 SD away from the mean. Satisfaction Raw score converted to Z score: $(29 - 20)/4 = 2.25$ Z score. Just say the z score is 2.25, and you have all the information you need.

Helpful features of Z scores are:

- You quickly get information based on the z-score. If a Z score is 2.5, it means it is above the mean and in the upper 5% of scores. If a Z score is -.5, you know that it is below the mean and within 68% of scores around the mean.
- You can compare the scores from different ranges. For example, CESD and BDI have different score ranges. For the same subject, the CESD: Z score is 1.5, and the BDI: Z score is 1. The subject did worse on the CESD. For the same subject, the SAT: Z score is .6, and the ACT: Z score is .9. The subject did better on the ACT.

Z-scores, so.... They are kind of useless. Honestly, I have never used z-scores in any statistic test, reported them, or read about them in journal articles. I've only used them for outlier analysis. Any score with a z-score of above ± 3 is flagged as a possible outlier. To be an outlier, the score needs to be alone, i.e., no other scores around it. But I never report a z-score for outliers. The z-score analysis is for diagnostic purposes that are behind the scenes as I diagnose normal distributions.

6.4.3: T-scores

A T-score is a standardized score based on a score distribution with a mean of 50 and a standard deviation of 10. For example, a raw score that is 1 standard deviation above its mean would be converted to a T score of 60.

Note that the t-score is not to be confused with a statistical t-test. A t-test compares the mean scores of two groups. The t-score is a standardized score, and it has nothing to do with a t-test.

T-scores are used in psychometrics. A t-score in psychometric (psychological) testing is a specialized term that is not the same thing as a t-score that you get from a t-test. T-scores in psychometric testing are always positive, with a mean of 50. A difference of 10 (positive or negative) from the mean is a difference of one standard deviation. For example, a score of 70 is two standard deviations above the mean, while a score of 40 is one standard deviation below the mean.

A t-score is similar to a z-score — it represents the number of standard deviations from the mean. While the z-score returns values from between -3 and 3 standard deviations from the mean, the t-score has a greater value and returns results from between 0 and 100, and most scores will fall between 20 and 80. Many people prefer t-scores because the lack of negative numbers means they are easier to work with, and there is a larger range, so decimals are almost eliminated. Table 6.1 shows z-scores and their equivalent t-scores.

Table 6.1 *Equivalent Z-scores and T-scores*

	z-score	t-score
	-5	0
	-4	10
	-3	20
	-2	30
	-1	40
	0	50
	1	60
	2	70
	3	80
	4	90
	5	100

6.4.4: Percentiles

Percentiles tell you the location of the person's score along the normal distribution and how many subjects scored above and below. Percentiles are used for comparisons. For example, person A scored in the 90th percentile, and person B in the 80th percentile.

Percentiles are not to be confused with percentages. Percentages need a denominator, which can shift, making comparisons impossible. Take 50%. You have no idea if that is good or bad because you do not know the denominator. 50% could mean 2/4 or 50/100. The denominator shifts, giving 50% a different context.

Summary Tables

$z \text{ score} = (\text{raw score} - \text{mean}) / SD.$

$T \text{ score} = 10 (z\text{-score}) + 50.$

Helpful Websites

<https://www.statisticshowto.com/prob...score-formula/>

<https://web.cortland.edu/andersmd/STATS/stdscore.html>

<https://www.hellovaia.com/explanatio...on-percentile/>

This page titled [6.4: Standardized Values of Reporting Scores](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

6.5: Discussion Questions

Discussion Questions

1. Why do we need variance?
2. What are the different types of variances, and why do we separate variance into those types?
3. What are the ways we measure variance?
4. Why is standard deviation an important statistic?

This page titled [6.5: Discussion Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

CHAPTER OVERVIEW

7: Sampling, Standard Error, Central Limit Theorem

Learning Objectives

By the end of this chapter, you will be able to:

- Explain why we need samples to represent the population
- Apply the guidelines for sample sizes
- Examine the consequences of sampling bias
- Explain the need for the standard error of the mean
- Explain the purpose of the Central Limit Theorem (CLT)

Key Terms:

- Sample and the Population
- Sampling
- Sample Bias
- Standard Error of the Mean
- Central Limit Theorem

[7.1: The Purpose of Sampling](#)

[7.2: Sampling from the Population](#)

[7.3: Getting a Good Sample - Get a Random Sample](#)

[7.4: Sampling Bias or Error](#)

[7.5: Central Limit Theorem \(CLT\) and Its Implications](#)

[7.6: Summary of Reasons Why You Need a Large Sample Size](#)

[7.7: Discussion Questions](#)

This page titled [7: Sampling, Standard Error, Central Limit Theorem](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

7.1: The Purpose of Sampling

Why do we think our samples represent the population? That is the goal of statistics. We have a sample, and the results from that sample are true for the population. But how do we know that?

One of the hallmarks of statistics is estimation. We are always estimating or guessing the truth. We want to describe the variation in height in the U. S. We want to describe the variation in salaries in a given state. We want to describe the variation of standardized test scores in a given school district. We want to describe the variation of alcohol drinking across universities in the U.S. We want to describe the variation of depression among young people in the U.S.

In each case, we want to estimate the following population parameters: the population mean, the standard error of the mean, and the normal distribution. What is the average height in the U.S. (BTW, you may think that is an innocuous example, but I listened to a podcast about how average height between nations reflects each nation's population's quality of access to health care. Better health care leads to taller citizens. And America's average height declined over the years compared to other countries. Interesting. We want to know the standard deviation or the spread or range of salaries from the average salary in the U.S. We want to know the standard error of the mean, which is to what degree we have the mean of standardized test scores that accurately reflects the test scores in a given state. We want to know the average amount of alcohol consumed across universities. We want to know the average and spread of depression severity across young people in the U.S.

Every time we state we want to know something, we want to know the population parameter. We want to know the truth about something. But we cannot know the truth with certainty. All we can do is guess, which is called obtaining a sample statistic.

We use sample statistics to estimate the population parameter. Parameters are characteristics of a variable's variation that we think represent the variation of the phenomenon in the population.

Here are some questions we want to address about the sample and the population:

1. How do we know that our samples do represent the population?
2. Why does everyone say we need a large sample size?
3. Why do we want to establish the population?

Let us start with the last question. We want to establish the population because that population is the basis for determining what observations are expected or unexpected. We can establish which observations have a high or low probability of occurring. This process helps us establish what observations and findings are significant or not significant.

An important concept for establishing the population is that we need to know the baseline of an issue. One example of a baseline would be this question: What would occur if nothing happened? If we did nothing, what would be the level of depression among young men? Obtaining the population value is one way of establishing what things are like now, and then what if we did something? Would it change the population value? If we implemented a community outreach program encouraging young men to seek treatment for depression, could we change the level of depression? We establish the population.

Let us address the first question: How do we know our samples represent the population? Recall that the goal is to estimate the population parameter. I consider the population to be the truth. The truth of what is really going on involves everyone. I like to think of the population as the answer key. How do we know if we have the correct result unless we have an answer key containing the correct result? We need a way of knowing if we are right or wrong. However, you may recall that it is impossible to estimate the population.

Logistically, it is impossible to do so because you must sample everyone, everywhere, and at any time.

Conceptually, we do not have the measures to establish a population value of many psychological constructs. We do not have a good measure of depression; although the MMPI, the Beck, and the CESD are good measures, they each measure depression from a different viewpoint. We also do not have an established protocol for determining if the measures of depression are invariant across different contexts, especially gender, racial, and ethnic groups. These problems persist when we try to measure constructs such as mindfulness, racism, social desirability, violence, and attachment. These constructs are simply harder to measure because it is hard, rightfully so, to establish a single definition of the construct and how it varies.

Methodologically, it is difficult to obtain measurements from the population for issues that people generally do not disclose. It is hard to get a prevalence rate of sexual assault, PTSD, and ADHD because it is hard for people to come forth and provide data on these difficult issues. We don't have a census of certain populations, such as adolescents with eating disorders. We don't know who to ask. For these and many other reasons, it is difficult to obtain a population value for any issue.

This page titled [7.1: The Purpose of Sampling](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

7.2: Sampling from the Population

We will address the second question: Why does everyone say we need a large sample size?

We cannot obtain data from everyone in the population; we sample from the population. What is the problem with sampling? The problem with sampling is knowing if we have enough of a sample and a varied enough sample representing the population distribution. How do we know if we have a good sample or a sample that is our best guess about the population parameter?

What would be a bad sample? We would never base a conclusion on an N of 1. No one person, one observation, can represent everything. There are examples of why a sample size of one will not work. You would not get married to someone on a first date, nor would you give up on dating if one date with one potential suitor did not work out. Suffice to say, common sense would dictate that an N of 1 would not work.

Statistically speaking, a sample size of one would not work because you have no idea if the next sample you obtain will be the same or different. For continuous variables, you have no idea if the next sample will be higher or lower than the first sample. For categorical variables, you have no idea if the next sample will be the same or a different type from the first sample. You cannot establish a trend in the variable, which means you cannot predict what will happen next. Statistics is all about prediction – we use the independent variable to predict what will happen to the dependent variable. So, the first reason we need a large sample size is because we need enough data to determine what will happen next.

You would also not base a conclusion on a sample size of one when that one sample was taken at just one location. No one would believe you if you took a temperature reading by the lakefront in Chicago and stated that temperature reading is the correct temperature for Chicago. Everyone in Chicago knows the phrase ‘cooler by the lake,’ so taking a temperature reading there would not suffice. The temperature readings must be varied throughout the Chicago area for the temperature to represent Chicago.

The issue becomes challenging when you think about populations with issues that are hard to address. Trying to get a population estimate of adolescents with eating disorders is difficult because it is hard to determine if you have the full range of the severity of the eating disorder. Where and in what context could you get the full range? You could start at high schools, different grade levels, and different locations with different cultural and economic contexts that undoubtedly affect the stressors and influential factors associated with the severity of an eating disorder.

This page titled [7.2: Sampling from the Population](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

7.3: Getting a Good Sample - Get a Random Sample

Given these concerns, how do we know we got a good sample? Put differently, is our sample a spot-on estimate of the population mean? Or did we miss the bullseye?

Keep in mind that getting a good sample is not just a statistical issue, but a research design issue. The research design issue is important because the procedures used in sampling and recruiting a sample affect the validity of the sample and whether the sample is a good representation of the population. Statistics do nothing to help a sample recruitment strategy that only consists of Caucasian, upper-middle class individuals, and college students. Depending on the intent of the results, the sampling strategy needs to reflect the goals of the study and the theoretical premise of the study. That being said, some populations are simply hard to sample and recruit. Police officers with PTSD generally do not participate in studies about their mental health because of the fear of being stigmatized. LGBTQ+ individuals who are not out to their communities are difficult to recruit via sites or centers, such as LGBTQ+ agencies, where there would be a ready audience. The research method of sampling and recruitment needs to be monitored when considering populations that are simply not readily available to participate.

- What would be the start of a good sample?
- Sampling ideally has to be random. We want a random sample. Why random?

If we keep sampling, obtaining observations or data at random, meaning no pattern, no purpose, no objective, we should get results from each sample that are the same as the population. Random means getting observations higher and lower than the true population mean. This is how the observations would disperse in the distribution if we did nothing. We want nothing to affect our estimation of the population parameter.

The opposite of random sampling is convenience sampling. This sampling is not a statistical term; it is more of a research design sampling procedure. Convenience sampling is recruiting readily available participants. Examples include sampling college students in a classroom by walking into the classroom and asking for volunteers. Surveying clients waiting in a doctor's office is a form of convenience sampling. While convenient and easy to execute, it is not random sampling and runs the risk of sampling only a portion of the population rather than obtaining a sample from the full range of the population.

This page titled [7.3: Getting a Good Sample - Get a Random Sample](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

7.4: Sampling Bias or Error

However, random or convenient the sample is, we still have a problem: no matter what we do, we will never know if we got the correct sample that represents the full range of the population. We call this problem sampling bias or error inadvertently sampling just a portion of the population. That sample really is not the best representation of the population. It means that you got the wrong sample....and by wrong, we mean that your sample only pertains to one area, too high or too low, of the population. If the sample is too high or too low, that means we overshot or undershot the population mean.

What's the big deal? The consequence is that we make conclusions about the population based on that sample, but that conclusion really does not represent the population. Note that you will always make a sampling bias or error. Every sample has a sampling bias because we have no population to compare the sample with. We can never tell if we are right or wrong because we have no population or an answer key to determine if we are right or wrong.

Back to our question, "How do you know your random sample is the right sample?" The answer is you don't. There is no way to tell because of the concept of sampling bias or error. Sampling bias or error is not a bad thing. You will always make a sampling bias/error because you have no idea what the truth or what the population mean is.

7.4.1: How to Address Sampling Bias or Error

There is no way to remove sampling bias or error. But we can try to address it, and the solution is to keep sampling. A sample size of $N = 1$ will not work. So, increase the sample size by getting a collection of samples. Increasing the sample size by getting more samples is known as the sampling distribution – the distribution of sample means. The sampling distribution refers to the following: If we constructed a frequency plot of the sample means, if you put all the samples together, eventually they should look like the population distribution.

This is the second reason why we need a large sample size.

7.4.2: Standard Error of the Mean

We keep sampling. But we can't keep sampling forever and continuously.

How do we know when to stop sampling? We really do not know, but the standard error of the mean can give us an indicator of whether our sample is getting close to the population's mean or the truth. The definition of the standard error of the mean is the standard deviation of the sample mean. It is a number (value) that tells us how close our sample mean is to the true population mean. We need some sort of gauge to tell us when we can stop sampling, and the standard error is our gauge.

Why is it called an error? Because we always make mistakes when we use the sample mean to estimate the population mean. The nature of the mistake is that the sample mean is too high or too low compared to the population mean.

How do we calculate the standard error of the mean? As we get more samples and add them together, and our sample mean does not shift or change that much, we are getting closer and closer to the population's mean.

What exactly is shifting in the standard error? Every time you add a new observation, the mean shifts. How much of a shift? The shift depends on the observations themselves. The mean shifts a lot if you get highly variable observations. The mean shifts a little bit if you get observations that are close to the mean.

The mean shifts become minimal the more we keep sampling. The fewer observations, the more shifts because you get more variable responses. But the more observations you get, the shift eventually becomes less because you are getting closer to the mean. This is the third reason we need a large sample size. You always want more observations and more samples because you will get fewer shifts in the mean.

Here is a demonstration: If you are a psychology graduate student, chances are you have to travel to practicum sites. Ask each student to state how long it takes them to travel to a practicum site. Write down each response. The first response almost always does not represent the classroom's average response time. Now, ask another student, and then take the average of the first two responses. Note the average. Then, ask another student to respond and take the average of the first three responses. Then, notice the difference between the first average and the second average. Keep collecting responses until you obtain the entire classroom. With

each average, you will notice the difference between the averages decreasing to the point where there will be a minimal decrease. This minimal decrease is the behavior of the standard error as you keep collecting data.

7.4.3: Interpreting Standard Error of the Mean

We want the standard error to be small. Notice that it is called a standard *error*. In a sense, we always want to minimize errors, so we want these errors to be small.

The typical range of a standard error is 0 to ~ 1.5 . You want that value to be as close to 0 because that means that the sample mean is not that different from the population mean, or the sample does not shift that much each time you add a new observation.

Note that the standard error of the mean will never be 0 because 0 means you got the correct sample mean; there is no error. However, that outcome is impossible to achieve because every time you sample, you add information or data, and the mean will shift.

What happens is that the standard error will get near zero the more we sample, but it will never actually be zero because, conceptually, we can never get the correct, or population, mean.

What standard error values are good? A low standard error is good with values from $\sim .01$ to $.03$. This range means the sample mean is the same as the population mean. The sample mean does not shift that much each time we add more samples, otherwise known as more observations.

If the standard error is higher than $.03$, is that bad? Not necessarily, but if you have a low sample size, you might want to consider collecting more data. If you can, of course. Sometimes, we cannot collect more data due to resource limitations or the nature of the population, which is hard to sample. There are other issues to consider regarding the accuracy of your sample other than just relying on the standard error. BTW, despite its importance, you rarely see the standard error reported in journal articles. Statisticians call the standard error of the mean the “root of all statistical evil.” The reason it is evil is that there are situations where it is hard to estimate the standard error. In the above example, the standard error predictably decreased when you kept sampling. However, the sampling must occur at random. Each observation taken at random has to be independent of other observations. Independent means that taking a sample has no bearing on how the next sample is taken.

There are quite a few situations where observations are dependent on other factors. For example, better access to mental health resources is associated with increased income, so those two factors are not independent of each other. The problem occurs when you have nested data or when the participants are more alike rather than independent of each other. For example, residents on a neighborhood block are more alike in their income rather than having independently varying incomes. So, instead of sampling 10 residents, you have sampled 10 residents with similar income, amounting to just one block. It is difficult, then, to determine if you have 10 independent sources of income or one very similar source of income. When your observations are more alike than varied, the standard error is difficult to determine. Discussion of this issue is extensive, and you are encouraged to review resources about sample dependence.

Suffice to say, the standard error is rarely reported, but it is an important concept because it is how we determine we have the best sample mean.

Back to the original question – how do you know you got the correct sample? Use the standard error of the mean. If it is near zero, you got the best sample estimate possible. Is it really the correct sample? No, because the population parameter (the truth) is always unknown. At the end of the day, sample size, the standard error, and your research recruitment method are all you have to determine if you got the “right sample” or the sample that is the best representation of the population.

The goal is to use a sample to estimate the population's mean. But each time we take a sample at random, it is possible that our sample mean is not a good estimate of the population mean. That is called sampling bias/error. And if we always make a sampling bias/error because it is an inherent problem that the sample will never equate to the population, what do we do?

Solution: keep sampling. Get a sampling distribution of means. Technically speaking, if we keep sampling, eventually, we should get something that comes close to the population's mean. But how many samples should we get?

The solution is to use the standard error of the mean. The standard error of the mean tells us how much the sample mean will shift with each sample. With each sample mean, we add it to the previous samples, and if the mean doesn't shift too much, then we are probably converging on the true population parameter.

So that's it? Is this the process for determining if a sample represents the population? Nope. We need one more concept.

This page titled [7.4: Sampling Bias or Error](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

7.5: Central Limit Theorem (CLT) and Its Implications

The Central Limit Theorem is the glue that ties the sample to the population. You need the Central Limit Theorem (CLT) to state that you can claim that the sample represents the population. The central limit states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger. In theory, if you keep sampling at random, increasing the sample size, then three things happen:

1. Eventually, your sample distribution of the sample means (NOT the sample distribution) will resemble a normal distribution.
2. The standard error of the mean will decrease because as you get a bigger sample, our sample is a good representation of the normal distribution.
3. Our sample mean matches the population mean.

The CLT is about convergence. As you keep sampling, your sample mean gets closer and closer to the population's mean. As you take more samples, especially large ones, your graph of the sample means will look more like a normal distribution.

The important implication of the CLT is that you can claim your sample represents the population. It is a theoretical proposition, hence the name theorem, for a theoretical problem. The problem is, can a sample represent the population when, theoretically, it is impossible for the sample to represent the population? A theoretical problem is solved by the theoretical solution, the CLT.

What is interesting about the CLT is that no matter the shape of the population distribution, which can be skewed or normal, the distribution of the sample means will look like a normal distribution. There is a difference between the distribution of the sample means and the sample and the population distribution. The sample means can be taken anywhere from the population distribution. BTW: under the CLT, the bell curve is the sampling distribution of sample means; it is NOT the sample distribution. Putting the sample means together forms the normal distribution, not continuous sampling.

The CLT implies that you can stop sampling when you have a low standard error. The low standard error means that, according to the CLT, your sample is a good estimate of the population. With a high standard error, you need to sample more. CLT says your sample is not a good estimate of the population.

This page titled [7.5: Central Limit Theorem \(CLT\) and Its Implications](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

7.6: Summary of Reasons Why You Need a Large Sample Size

- Drawing a conclusion based on a sample size of one is inaccurate. The reason why you need a large sample size is that you have no idea if the next sample you take is going to be the same or different from the first sample. Likely, the next sample is going to be different.
- Drawing a conclusion based on a sample size of two is inaccurate because you cannot establish an upward or downward trend in the scores. You have no idea if the next or third sample you take is going to continue the trend established by the first two samples. For continuous observations, the first sample could be low; the second sample could be high, suggesting an upward trend. However, the third sample could be low, suggesting a downward trend, or high, suggesting a continuous trend. The fourth sample could be low, suggesting a downward trend, or high, suggesting a higher trend. You need a larger sample size to determine the overall trend, increasing or decreasing, of the variable.
- Type I and Type II error considerations. Low sample sizes tend to result in type I or type II errors. If you have a sample size of four observations, and all four observations confirm your hypotheses, technically, you have made a Type I error. However, the likelihood of you making the Type I error increases because it is possible that if you sample another 26 observations, those 26 observations will not confirm your hypothesis, hence the Type I error. The same is true for the Type II error. If you have a sample size of four observations, and all four observations do not confirm your hypotheses, technically, you have made a Type II error. However, the likelihood of you making the Type II error increases because it is possible that if you sample another 26 observations, those 26 observations will confirm your hypothesis, hence the Type II error. The sample size and the Type II error considerations are addressed with a power analysis, which is covered in the next chapter.
- Sampling bias is a theoretical consideration for why we need a large sample size. Every sample we take has a sampling bias; in this case, the sample, specifically the sample mean, will not be a correct representation of the population's mean. Every sample we take will be higher or lower than the population's mean. We can never know because the population's mean is unknown. If we do not know, then our best alternative is to keep sampling and hope the culmination of our samples will bring us closer to the population's mean. But to do that, we keep sampling, hence the need for a larger sample size.
- The standard error of the mean is a mathematical reason for needing a large sample size. The standard error of the mean needs to be low because that is the gauge we use to determine if we are getting closer to the population's mean. The mathematical formula for the standard error of the mean involves the sample size as the denominator. When the denominator is large and divided by a small number in the numerator, the result is a small number, less than one. The larger the sample size, the larger the denominator, and the calculation will yield a number considerably less than one.
- The Central Limit Theorem is a second theoretical consideration for why we need a large sample size. The theorem states that when sampling continues at random, the distribution of sample means will theoretically match the population parameters (normal distribution, standard error, population mean). The theorem is based on continuing to sample or a large sample size.

7.6.1: Guidelines for Sample Sizes

Notice that the preceding reasons for needing a large sample size do not specify exactly how large the sample size needs to be. They do not state the lower bound or the minimal sample size you need. There is no indication of how small the sample size must be when conducting statistical analyses becomes problematic.

There are no good guarantees regarding what the sample sizes should be. There are too many considerations, such as the distribution of the variable in question, the research question itself, and the logistics of recruiting the population under consideration. The other problem is missing data. Missing data is inevitable. Participants start surveys and then do not complete the remaining questions or skip other surveys. Participants skip questions. Participants complete the demographics and nothing else. Every time you obtain a pool of participants, it is likely that not everyone in the pool will complete all our study surveys. Even if you do set a minimum sample size, you always want to plan to survey more as a backup plan because you need the extra data to compensate for the missing data that will inevitably occur.

There are some guidelines for what sample sizes should be. These are guidelines, not rules. Anyone who states that the numbers are strict and that any sample sizes below the guidelines makes the study invalid is clearly wrong and does not pay attention to the context of the study. A problem with stating what the sample size should be as a rule occurs when the sample size is just underneath the minimum sample size. If a sample size is set at 30, what happens when the sample size is 29? If you do not have that 30th data point, does that mean your statistics are invalid? No. Statistical results do not operate on a green light/red light switchboard. The

statistics do not tip over into significance when the sample size is 30 but are not significant if the sample size is 29. However, students (and faculty) get worried about not meeting the sample size mark. The sample size does not solely make or break significance.

The same is true for a maximum sample size. Why not increase the sample size? Because a large sample size does not give you any additional information about the stability of your statistical results beyond the information you obtain from your presumably sufficient sample size. A sample size that is too large would be considered to contribute to a Type I error. The more you sample, the more you will find a result anyway, not because you found an actual result but because you sampled so many times, based on chance alone, and eventually, you will find something. We call that fishing, if you recall. The more you fish, the more you'll eventually catch, but that doesn't make you a fisherman; it makes you lucky. Students (and faculty) might consider getting a large sample size to conduct supplemental or additional analyses to find interesting effects beyond their hypotheses. But that still is not a good reason to keep sampling because, conceptually, it does not make sense according to your hypotheses, and you are just collecting data, just in case, to conduct additional analyses.

So, what are the guidelines?

- The first guideline is that the sample size should be at least 30.
- The second guideline is that the sample size should be a maximum of 300.
- The third guideline is that there should be a sample size of five per survey item.
- The power analysis is a primary guide for determining your sample size, which is the subject of the next chapter.

Where did these guidelines come from? These sample sizes are based on computer simulations of random data. For the sample size of 30, computer simulations indicate the standard error of the mean for a given sample does not change considerably when the sample size is 30. The same goes for a sample size of 300, as computer simulations indicate the coefficients of association between variables do not change considerably when the sample size is 300.

The sample size of five participants per item on a survey instrument is based on factor analysis. Despite concerns about factor analysis as the analysis of choice for examining dimensions of scales, the five participants per survey item is a guideline for determining if an item is an indicator of a scale dimension.

The sample size issue is a complex one. Yes, it cannot be ignored, but to place so much stock in a single number as the only determinant for statistical significance, and if a student (or faculty member) should proceed with a study, is not good practice. The bottom line is if someone is criticizing your study and its sample size, or preventing you from conducting your idea solely on the basis of anticipating that you will not get your desired sample size, or making you feel nervous or uncomfortable because your sample size will be inadequate, you need to consult with someone (and you are welcome to consult with me). No one should make you feel like a study cannot be done just because you cannot collect an adequate sample. Yes, the sample size needs to be considered, but no one should make you feel bad about any consideration of your sample size. Always consult with someone so that at the end of the day, you are comfortable with any decision you have about your study, your results, and the role of the sample size for interpreting your results.

Helpful Websites

<https://www.statisticshowto.com/probability-examples/>

<https://statisticsbyjim.com/basics/central-limit-theorem/>

This page titled [7.6: Summary of Reasons Why You Need a Large Sample Size](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

7.7: Discussion Questions

Discussion Questions

1. What are the differences between the standard error of the mean and the standard deviation?
2. Explain the process of how we can consider a sample as representative of the population.
3. If you do not have a large sample size, is that a problem and why or why not?
4. Complete this table on reasons to obtain a large sample size.

Questions	What happens if the sample size gets larger?	What happens if the sample size gets smaller?
Reasons		

This page titled [7.7: Discussion Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

CHAPTER OVERVIEW

8: Probability, Hypothesis Testing, Type I and Type II Errors

Learning Objectives

By the end of this chapter, you will be able to:

- Understand how to test if a result is significant
- Distinguish between Type I and Type II errors

Key Terms

- Probability
- Hypothesis

Recap: Statistical inference (AKA all about that p value)

We are in a descriptive statistics camp. So far, we have established the normal distribution, central tendency, variability, standard error, and central limit theorem. These concepts are the building blocks that will lead us to the basis for the next camp: inferential statistics.

The hallmark of inferential statistics is inferring. Inferring is similar to choosing, deciding, and making the best guess possible if we have a significant or non-significant effect.

A significant effect is something likely not happening based on chance alone. Chance alone is synonymous with randomness. In statistics, we hate random. That means nothing is happening. Anything that does happen is out of our control. Anything can happen.

In research, we want to find something that explains what we see. In psychology, we want to say that what we do makes a difference [in](#) our clients' lives. Clients do not get randomly better. Clients who see psychologists do much better than those who do something on their own. In social justice, we want to say we are doing something to create change for the better. We do not like relying on time to heal social problems. We need to do something to make the positive changes we want to see.

That something, that interesting finding, that nonrandom finding, that significant finding, is usually in the form of a pattern. Patterns are usually in these two forms.

- Finding differences among groups
 - Group A is higher / lower than Group B (your F-test, t-tests).
 - If there is no pattern, Group A is similar to Group B.
- Finding associations among variables
 - As Variable X increases, Variable Y increases / decreases (your correlations, regression)
 - If there is no pattern, we say, there is no relationship between Variables X and Y.

These are the patterns found in the conventional ubiquitous parametric statistics. These statistics are your t-tests, F-tests, correlations, and regressions. Parametric statistics means the statistical test is based on population parameters, which are based on the central limit theorem and include the normal distribution and standard error of the mean. Yes, there are other patterns based on non-parametric statistics, such as chi-square, and other patterns based on multivariate statistics. To keep it simple, we see two broad patterns using conventional parametric statistics.

[8.1: The Role of Probability](#)

[8.2: Statistical Decision - Is a Result Significant or Not Significant?](#)

[8.3: Hypotheses for Statistical Inference](#)

[8.4: How to Determine if a Result is Significant](#)

[8.5: Interpreting a Statistical Test Result](#)

[8.6: Testing a Hypothesis](#)

[8.7: How Do We Get the P Value for the Statistical Test?](#)

[8.8: Confidence Intervals](#)

[8.9: Type I and Type II Errors](#)

[8.10: The Definition of Type I and Type II Errors](#)

[8.11: Discussion Questions](#)

This page titled [8: Probability, Hypothesis Testing, Type I and Type II Errors](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

8.1: The Role of Probability

In statistics, we want to find that something is significant. Compared to what? A random result. Probability establishes what observations are random results and what observations are not random. Put another way, what observations are significant or likely not random, and which ones are not significant or likely random?

Of note is that significance is an interesting concept because if a result is significant, it has to be significant compared to ... what? Recall that a hallmark of statistics is the study of variation. Variation is based on differences. Differences are in this form: this is different than that. The “that” is important because when we say something is significant, it means that the difference between your observations and “that” is likely not a random occurrence. But what is “that?” The “that” is randomness.

However, in the field, we do not usually compare something to randomness. In psychology, we usually compare a treatment group to a control group or a group that does nothing. Or we compare it to usual treatment, such as medication. The main issue is that when something is significant, we must ask the question, “Compared to what?” In one sense, anything can be significant compared to random chance. Anything can be significant compared to doing nothing. Something might not be significant when compared to something established, such as usual treatment care. The issue is a research design issue, not a statistical issue. Research design sets up the comparison of “this compared to that.” Statistics determine if the comparison is significant or if the difference between this and that is no better than random chance. In the end, you want to be cautious when someone says, “This is significant.” You want to respond, “significant compared to what?” A result that is significant compared to random chance is really not that interesting.

It is worth noting that this problem of “compared to what” can be politically sticky. By politically sticky, I mean we can be biased if our results are significant simply by reconfiguring what we are comparing them to. In global warming, we say that our climate is warmer than before. Before what? By following the global temperature across hundreds of years, we can see a warming trend. But others might say, well, compared to the last 20 years, this warming trend is nothing but a mild increase and probably random. Consequently, in a few more years, we will see a cooling trend. Depending on the motives of interested parties, we can make anything “significant” or “not significant” depending on what we choose as our comparison frame. And no, that is not a good practice, so we must be mindful of when someone says, “It is significant,” we should always ask, “Compared to what?” so that we can be thoughtful about the validity of their statement.

What does it mean for something to be significant? What is the thought process for determining that something is a random result versus something is happening? Statistically and technically speaking, we can never be sure that our result is due to something significant. There is always a chance that our result is due to chance. When we say that something is significant, we mean that the probability of the result happening by only random chance is so small that it is more plausible that something must have led to that result because there is very little chance that the result could have occurred on its own by random chance.

A classic example is the poker hand. It is entirely possible that you obtain a full house on the first hand of poker. It is entirely possible that you obtain a full house on your second hand of poker. It is entirely possible that you obtain a full house on your third hand of poker. You can obtain three consecutive full houses in three hands of poker, all due to randomness or just plain luck. However, the probability of getting three consecutive full houses in three hands of poker on plain luck alone is very low. The opposite of chance is a pattern, or “something is going on.” The alternative explanation is that you are doing something to lead to that result because that result would likely not have happened on its own by pure chance. Emphasis on “likely.” Yes, the three consecutive full houses for each poker hand are entirely possible due to chance, but the likelihood is very low. The more plausible explanation is that you did something to lead to that outcome.

Suppose we want to test a new treatment for alcoholism. The average sobriety time after discharge from an alcoholism treatment program is three months. We test your new treatment for alcoholism, and the average sobriety time becomes one year. Is it possible that your set of clients got lucky and just happened to stay sober for one year on their own? It is always a possibility. But the probability of that happening is so low that the more plausible explanation is that your new treatment program is so good that it is likely the treatment program led to a good outcome rather than pure luck.

This issue gets politically sticky. Take the example of global warming. Is the increase in temperature spikes a random event, or is it due to something happening? As of 2024, some believe that increased temperature spikes can occur randomly throughout the history of the planet’s temperature; therefore, what we are experiencing as a temperature spike is nothing more than a random event.

This concept of intentional actions rather than random luck creating specific results is the basis of significance testing. Technically, we can never be certain that what we found is due to our treatment effect or hypothesized association between variables. Technically, anything can happen based on pure luck alone. However, we must weigh the possibility that instead of something happening by random luck, an action must have been taken to cause the outcome because the likelihood of this outcome occurring randomly is low.

Notice we say that the likelihood is low. We didn't say that the threshold for determining the probability of something happening by pure luck is too low. We did not specify how low a low probability is. This specification is the focus of the remaining chapter.

This page titled [8.1: The Role of Probability](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

8.2: Statistical Decision - Is a Result Significant or Not Significant?

The following is a framework for what significance means.

- A result is not significant if:
 - It occurs frequently and has a high probability of occurring,
 - It is common
 - It is not interesting
 - It is boring
 - It is random; this result likely occurred by random chance. Nothing happened.
- A result is significant if:
 - It does not frequently occur and has a low probability of occurring
 - It is not common
 - It is interesting
 - It is not boring
- It is not random; this result likely did not occur by chance. Something happened.

Yes, it is safe to say that we want significant results. We would like to think that if we have an interesting result, that something is going on that is responsible for the result that we observe. Usually, something is a treatment effect – for example, our treatment led to this positive clinical outcome. Or that something is an important variable that is associated with a clinical concern – for example, increases in the number of racist experiences are associated with severe depression symptoms.

There are plenty of times when we do not want significant results. Sometimes, we do not want significant results as part of a research design to establish equivalency. For example, when comparing a new treatment to an established treatment, we want the client demographics in both treatment groups to be the same. We want the number of males and females to be equal in both groups, equal in average age, income, and education levels. The reason is so one group is not imbalanced or favors one demographic that might have a role in the treatment outcome. We want everything to be the same in both groups, meaning we want to see non-significant results, or equivalency, between the two groups.

We do not want significant results when we want things to be similar. This scenario depends on the research question. Sometimes, you want conditions to be equivalent. For example, the idea that talk therapy is just as effective as medication. In this scenario, it would be good to establish that talk therapy produces outcomes similar to those of medication. In a climate where we medicate all mental health issues, it is worthwhile to demonstrate that talk therapy is an alternative to medication.

Sometimes, you want to have no relationship between variables. For example, an increase in self-esteem is not associated with a decrease in alcohol use. That might be a helpful finding because feeling better about yourself does not mean you are engaging in behaviors to decrease your alcohol use. Another example is finding no relationship between parenting styles and ADHD. That might be a helpful finding because it will relieve the blame parents experience if they think their parenting style resulted in their child having ADHD. The point is always about conceptualizing the issue. Always think conceptually to determine if non-significant results are helpful.

This page titled [8.2: Statistical Decision - Is a Result Significant or Not Significant?](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

8.3: Hypotheses for Statistical Inference

When structuring your decision-making, set up your hypotheses in the following manner:

- Null Hypothesis: H_0
 - The result is due to chance.
- Decide to accept the null hypothesis.
 - Nothing happened. All random. Nothing interesting here.
- Alternative hypothesis: H_A
 - The result is not due to chance.
- Decide to reject the null hypothesis.
- Something must have caused that result, because that result does not usually occur.

Usually, the goal is to reject the null hypothesis because we want to say we found something.

BTW, writing a hypothesis is much simpler when you state the hypothesis as a prediction. Although the following is statistically correct, it is awkward to read.

An introduction to the study could be written like the following passage.

Hypothesis one: The null hypothesis was that there would be no difference in the mean scores between the groups. The alt hypothesis was that there would be differences in the mean scores between the groups.

In the results section, then you would have to write the following, “The findings did not support the null hypothesis.”

Technically, the sentences are correct, but the presentation is awkward since the reader has to do mind flips to realize that the findings support the hypothesis. When writing hypotheses, keep them simple. An introduction to the study could be written like the following passage.

Hypothesis one was that there would be differences in the mean scores between the groups.

In the results section, then you would have to write the following, “the findings supported hypothesis one.”

This page titled [8.3: Hypotheses for Statistical Inference](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

8.4: How to Determine if a Result is Significant

Statistically, we use the alpha level to determine if a result is significant.

When you see alpha, think about probability. The alpha level is usually represented with an italicized p ; usually, you read $p < .05$. Sometimes, we use the alpha symbol α , and usually, we write $\alpha = .03$ to represent the probability associated with a statistical test. The p represents probability. Think of the alpha level as a cut-off probability. Based on the normal distribution, the probability of any outcome purely by chance alone ranges from 1% to 100%.

Based on the normal distribution, recall that results that happen with high frequency and high probability are common, boring, uninteresting, and not significant. Results that happen with low frequency and low probability are not common, interesting, or significant. We want rare results, and by rare, we mean there is a low probability that the result will occur by random chance.

How low is a low probability, or how rare does a result have to be that we are confident that it does not occur by chance?

The alpha level, the level of significance, is a percentage. A percentage of what? A percentage of the normal distribution. Recall that the normal distribution is basically an X-Y graph of the possible values of a given variable on the X-axis, and the Y-axis is the frequency count or the number of times that value occurred. Recall that for a normal distribution, the most frequent occurrence of values occurs around the mean, the highest part of the bell curve. The tails are a) low frequency occurring values and b) far away from the mean. Recall that the distribution is the dispersion of scores, and we use sums of squares, or we square the dispersion of scores, to get the area underneath the curve. The normal curve is shaped such that most of the area underneath the curve is around the mean, hence the wide and tall portion of the curve. That large area underneath the curve means a lot of observations fall in that area because these observations are frequent, close to the mean, and not that interesting. What we want for significance is something greater than the average or that has a small chance of occurring. That area underneath the curve is small, and that area is far away from the mean and short, meaning those observations do not happen that often based on random chance.

When we say alpha, we mean percentage, and the percentage of the area underneath the curve. When you read the ubiquitous, $p < .05$, what that means is that the alpha level is set at .05, which is a 5% probability that the result occurred by chance. Think “this result will only occur by chance 5% of the time. That’s a pretty low probability.” Think of the alpha level as a cutoff point that we place somewhere along the probability line. In statistical terms, we call that the region of rejection. Rejecting what? Rejecting the null hypothesis that says the results are due to chance and the result is not significant.

BTW, why 5%? Why not 10% or 1%? Who died and set the alpha at 5%? We answer that question by discussing Type I and Type II errors.

This page titled [8.4: How to Determine if a Result is Significant](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

8.5: Interpreting a Statistical Test Result

Statistical inference, or determining if a result is significant, is based on two pieces of information—first, the statistical test result, and second, the p value. For example, you will read the following numbers: t -test = 2.45. ANOVA f -test = 4.34. correlation $r = .58$. Chi-square $\chi^2 = 10.21$. Higher numbers indicate significant test results.

Statistical test results are essentially ratios. The ratio is the numerator over the denominator. The numerator represents the variability we want to see, the pattern we want to observe. These patterns are usually one of two types – group A is higher/lower than group B, or increases in variable X are related to increases/decreases in variable Y. The numerator is the pattern we want to see, and that is the true variance. The denominator is the error variance or random error. Observations can happen on a random basis. If so, then these observations are random and are part of the denominator. We always want to beat error or have true variability that is greater than error. When the true variance is greater than the error variance, the numerator is higher than the denominator. Then we have a large number, hence a large stat value and a significant result.

For comparing groups, usually, any test (t -test, ANOVA) results above 1.96 are significant, indicating that there are differences between groups. You could think that 1.96 as the true variance is twice as much as the error variance, which is good. For determining associations between variables, usually results above .20, ideally .30, indicates there are associations between variables. Anytime the statistical test result is around 0, it is not significant; there are no differences between groups and no associations between variables.

Statistical inference is based on a second piece of information, the p value. For example, t -test = 2.45, $p = .03$. ANOVA f -test = 4.34, $p = .02$. correlation $r = .58$, $p < .05$. Chi-square $\chi^2 = 10.21$, $p < .01$. Lower p values indicate significant test results. For comparing groups, a p value less than .05 is usually significant, indicating differences between groups. For determining associations between variables, usually, a p value less than .05 is significant, indicating there are associations between variables. Anytime the p value is above .05, it is not significant and there are no differences between groups and no associations between variables.

The value of the test results and the p value are inversely related. As the test result value goes up, the p value goes down. This situation is what you want. Do you look at the test result and the p value, or one or the other? Basically, you want to attend to the p value. The statistical tests can be difficult to interpret using hard, fast numbers. Statistical test values can be fickle, meaning they can change based on several factors, such as the sample size. You can de-emphasize the test result by using the p value.

This page titled [8.5: Interpreting a Statistical Test Result](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

8.6: Testing a Hypothesis

You want a plan for interpreting the statistical test numbers. For this example, the hypothesis is there will be a difference between the two groups (treatment vs. control):

- Set α at .05.
- Run the statistical test and obtain the result and p value.
- You want a high t -test value, usually above 1.96, and a low p value.
- t -test value = 3.96, $p = .02$. This p value is less than .05, so it is significant.
- Reject null, state the t -test value and the p value
- Write or state, “The t -test was $t = 3.96$ and was significant at $p < .05$.”
- If you have a low t -test value, usually below 1.96, and a high p value, then...
- For example, t -test value = 1.21, $p = .15$. This p value is greater than .05, so it is not significant.
- Retain null, state the t -test value and the p value
- Write or state, “The t -test was $t = 1.21$ and was not significant, $p = .15$.”

Some notes about writing the p value.

- According to APA style, note that when the test result is not significant, give the statistical test value and the actual p value. For example, “The t -test was $t = 1.21$ and was not significant $p = .15$.”
- If the test result is significant, there is no need to go further than the two decimal points. Write it out if the p value is from .05 to .01. Anything below .01, you can write $p < .01$.
- For example, “The t -test was $t = 3.96$ and was significant, $p = .03$.”
- Or “The t -test was $t = 10.01$ and was significant, $p < .01$.”

A note about not needing to write out the p value result beyond .01. Keep in mind that you should never write the following: $p < .001$, never say “highly significant,” “trending towards significance,” or “approaching significance.” I will explain later in this section, but for now, it is misleading to write everything mentioned in the previous sentence.

This page titled [8.6: Testing a Hypothesis](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

8.7: How Do We Get the P Value for the Statistical Test?

Note that when we say $p < .05$, that .05 is something that we, the researchers, set. The reason we set it at .05 is a different story, and we will cover that when we discuss Type I and Type II errors. But how does the actual p value get generated for a statistical test? In other words, if a statistical test, such as a t -test = 2.45, $p = .027$, how did the p value of .027 get calculated?

The answer to that question is that it is based on a t -distribution. Statistical inference depends on knowing the population mean and SD. Recall from the normal distribution discussion that for normal population distribution, all values will fall under the normal distribution curve in the form of values that are 68%, 95%, and 99% above and or below the mean.

But recall that it is a population distribution; by definition, the population is unknown. Consequently, we do not know the population's mean and standard deviation. What do we do? We do what we always do: we estimate population means from sample means. Recall that every time we estimate using a sample, our estimates vary because each sample gives us a different estimate. What do we do?

We use a t -distribution. Not a population distribution. It is a sample distribution. The t -distribution is a table of generated t -values corresponding to p values for a given sample distribution. If the sample is large, the sample means will be similar to the population mean. Based on data simulations, if the sample size is $n \geq 30$, the t -value is 1.96, and results are significant at the $p < .05$ level. If $n < 30$, the t -value is adjusted; consult a t -table. These t -tables are found in most introductory statistics textbooks in the appendix. BTW, the formula for calculating a p value is elusive, and it involves calculus. Learning this calculation is not easy and is not worth your time. I never use a t -table; your statistical software will compute the p value for you, and SPSS always provides a p value for you. You do not have to compute the p value associated with the statistical test. Let the computer program do it for you.

This page titled [8.7: How Do We Get the P Value for the Statistical Test?](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

8.8: Confidence Intervals

In addition to reporting the p value, report the confidence intervals (CI). Confidence intervals are the range where a set of values will fall. In this case, it is the range of statistical test values. Confidence intervals are important for determining the stability of a result. Because with repeated sampling, results can change. The question is, by how much could the result change, and if it does change, is the result still significant?

Confidence intervals are based on the concept of error. We accept or reject the null hypothesis based on one sample. But if we take another sample, we might get another, different test result.

What is the range of possible test results if we keep sampling? Stated differently, if we collect a different set of observations and re-conduct the statistical test, what will happen? The question is, will we get the same or a different result?

The statistical definition of a confidence interval is – Estimated range of sample test values that likely contains the true population parameter. In English, this statement means that if we keep sampling, our test result will fall into what range of possible values contains the “correct” value. I say “correct” because the population is unknown, and we will never know the correct answer.

The purpose of the CI is to corroborate the p value finding. The confidence interval gives

you a range of values from the lowest to the highest. 95% CI: 95% of the time, we will get a result that will fall within this range of values. The possible CI ranges are 68% CI, 95% CI, and 99% CI. More about these intervals later. For now, note that these intervals are similar to the normal distribution and the standard deviation.

What are you looking for if the CI contains a value of 0 that would lead us to retain the null hypothesis? If the CI range from low to high includes a 0, it means that there is no significant effect. If the CI range from low to high does not include a 0, it means that there is a significant effect, and the CI range corroborates or confirms the p value outcome, which declares that a statistical test is significant.

BTW, if the p value is significant, nearly 100% of the time, the CI does not contain the value 0. If the p value is not significant, nearly 100% of the time, the CI does contain the value 0. I have never encountered a situation where the p value was significant, but the CI contained 0 and indicated no significance. I say if the p value is significant, nearly 100% of the time, the CI will confirm that significant result because I have never seen the CI not confirm a significant p value. Like everything else, anything is possible; that is why I say “nearly” 100% of the time, but in my entire career with statistics, I’ve never seen it. As far as I am concerned, the p value indicates significance; then the CI corroborates that finding 100% of the time.

So, how good is the CI if it almost always confirms the p value result? Sometimes, I do use the CI to evaluate how wide the CI interval is. A wider interval means there is more error and more uncertainty in the estimate of the true value of the population parameter. A shorter interval means there is less error and more precision in estimating the true value of the population parameter. How wide or short does the interval have to be? It depends on how far away the upper and lower bounds are from the test result. There are no hard and fast guidelines to gauge the distance and if it is a wider or shorter interval. You should consult if you are wondering if the CI is too wide or short enough to be considered a good estimate.

In the end, why should you report confidence intervals in addition to the p value? As of 2020, the American Psychological Association (APA) wants you to report it. APA strongly recommends that, in addition to the p values, researchers provide effect sizes and confidence intervals because the all or nothing hypothesis testing does not provide the full range of possible outcomes of statistical analysis (for an explanation, see $p.$ 88, section 3.7, APA manual, 7th edition). If your statistics software program (e.g., SPSS) allows an option for a CI, select it and report it whenever possible.

This page titled [8.8: Confidence Intervals](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

8.9: Type I and Type II Errors

We covered the significance test, which answers this question: “How do we know that our observations, our findings, our desired pattern, are significant? By significant, we mean that something is going on. There is very little chance that this outcome, this observation, would occur by random chance alone. Something must be going on that resulted in this outcome. When a statistical test indicates that something is significant, that means something is going on. That “something” is usually a treatment effect or an established relationship between two variables.

Recall that a hallmark of statistics is estimating. We are estimating the truth, or the population parameters, by using what we observe, which is our sample or sample statistics. By definition, every time we make an estimate, we make a guess. And, like all our guesses, we can make the wrong guess. Why do our guesses have the potential to be wrong? Because the population is unknown, and we will never know if we are right.

Is there an answer key for knowing if you have found the special person in your life whom you will marry? Nope. The universe does not have an answer key that indicates you have found the right partner for the rest of your life. The universe does not tell you if you have found the right partner, nor will it tell you who the right partner is for you for the rest of your life. Every time you or someone else gets married, you say, “I do.” That “I do” is essentially a guess. You still have no idea if you made the right decision.

In statistics, there are two situations we want to avoid. Situation A: We say something happened, but that something actually was just a fluke, random chance. Situation B: We say nothing happened, but actually, something happened.

The terms we use to describe those two situations are Type I and Type II errors.

This page titled [8.9: Type I and Type II Errors](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

8.10: The Definition of Type I and Type II Errors

In a Type I error you reject null when it is true— “I think my finding was true and significant, but really it was due to chance.” If the alpha level α is at 5%, 5% of the time, I reject the null. There was a significant effect, but actually, I should have kept it because there was nothing there. We state, “My statistical result supports my hypothesis, but it is possible that my results were due to random chance, and I made a Type I error.” The type I error is the same as a false positive. You think you found something, but in reality, you did not.

In a Type II error – you accept null when it is false. --“I think my findings were not true and not significant, but really, something did happen.” If the Beta level β is at 95%, I fail to reject the null. There was no significant effect, but actually, I should have rejected it because there was something there. We state the following: “My statistical result did not support my hypothesis, but it is possible that my results indicated something did occur, and I made a type II error.” The type II error is the same as a false negative. You think you did not find something, but in reality, you did.

8.10.1: Examples of Type I and Type II Errors

A classic medical example is when a diagnostic test says you have a medical condition, but you really do not. The classic example is a pregnancy test. A Type I error means that you are pregnant when you really are not pregnant. A type II error means that you are not pregnant, but you really are pregnant. Back in 2020, we had the COVID pandemic. COVID-19 tests were developed quickly, yet there were concerns about Type I or Type II errors. A type I error means that you have COVID-19 when you do not have COVID-19. A type II error means that you don’t have COVID-19 when you do have COVID-19.

A psychology/forensic example would predict the risk of whether an offender will re-offend. The problem is that if you release an offender and state that the offender will not re-offend, you are making a Type II error because you say the offender will not re-offend, but there’s a chance the offender will re-offend. A psychological diagnostic example of a Type I error is assessing a child as having ADHD when, in fact, the child does not have ADHD.

Switching to humor, if you break up with your dating partner, you can use this break-up line – “I’m breaking up with you because I made a type I error. I thought you were significant, but in reality, you are not significant; you’re just an error. I made the mistake of thinking you were significant.” If someone breaks up with you, you can use this comeback line – “You are breaking up with me? You just made a Type II error. You think I am not significant, but in reality, I am significant. You just lost out on dating me because I am a significant finding.” Those two comeback lines should address all of your problems about what to say when you experience a dating breakup.

8.10.2: The Likelihood of Making a Type I or Type II Error

How likely are you to make a type I or type II error? This situation usually occurs when your statistical test value has an accompanying p value that is right on the alpha level line of .05. What do you do when you are in this situation? You have to examine statistical information, such as sample size and research design.

Let us walk through a Type I error. Situation A: $p = .04$, $n = 300$, 150 in the treatment group, 150 in the control group. Walking through the steps, first, we have to say that our test is significant because the p value for the statistical test is less than .05. Second, we have to automatically say we made a Type I error. Because technically, we found something, but there is always a chance we are wrong, and in reality, we did not find something.

In this situation, we have an outcome that is right near the .05 alpha level. So technically, we say we found something, but we made a mistake and made a Type I error. But did we really? One hint is the sample size. With an overall sample size of 300 evenly divided between both groups, indicating ample sample sizes in both groups and because sample sizes near 300 do give a stable estimate of the population, we can be confident, although never entirely certain, that our result is a real result.

Continuing with Type I. Situation B: $p = .04$, $n = 30$, 15 in the treatment group, 15 in the control group. Walking through the steps, first, we have to say that our test is significant because the p value for the statistical test is less than .05. Second, we have to automatically say we made a Type I error. Because technically, we found something, but there is always a chance we are wrong, and in reality, we did not find something.

In this situation, we have an outcome that is right near the .05 alpha level. So technically, we say we found something, but we made a mistake and made a Type I error. But did we really?

One hint is the sample size. With an overall sample size of 30 evenly divided between both groups with a sample size of 15, and because sample sizes below 30 do not give a stable estimate of the population, we cannot be confident, although never entirely certain, that our result is a real result, meaning that we did find something, but it is more likely it is a fluke or a random error result. In reality, we did not find anything because with a larger sample size, we might get more observations that actually contradict the significant finding.

Let us walk through a Type II error. Situation C: $p = .06$, $n = 300$, 150 in the treatment group, 150 in the control group. First, we have to say that our test is not significant because the p value for the statistical test is more than .05. Second, we have to automatically say we made a Type II error. Because technically, we did not find something, but there is always a chance we are wrong, and in reality, we did find something.

In this situation, we have an outcome that is right above the .05 alpha level. So technically, we say we did not find something, but we made a mistake and made a Type II error. But did we really? One hint is the sample size. With an overall sample size of 300 evenly divided between both groups, indicating ample sample sizes in both groups, and because sample sizes near 300 do give a stable estimate of the population, we can be confident, although never entirely certain, that our result is a real result.

Continuing with Type II. Situation D: $p = .06$, $n = 30$, 15 in the treatment group, 15 in the control group. Walking through the steps, first, we have to say that our test is not significant because the p value for the statistical test is more than .06. Second, we have to automatically say we made a Type II error. Because technically, we did not find something, but there is always a chance we are wrong, and in reality, we did find something.

In this situation, we have an outcome that is right above the .05 alpha level. So technically, we say we did not find something, but we made a mistake and made a type II error. But did we really? One hint is the sample size. With an overall sample size of 30 (15 per group) —and because sample sizes below 30 do not give a stable estimate of the population— we cannot be confident, though never entirely certain, that our result is a real result. This suggests that we did not find something. In reality, we might find something because with a larger sample size, we might get more observations confirming a significant finding that we will eventually find.

8.10.3: Hallmarks of Type I and II Errors

By definition, you will always make an error when you estimate. Type I and Type II errors occur every time. Whenever you say you found a significant result, you make a Type I error. Whenever you say you did not find a significant result, you make a Type II error.

You cannot make a Type I and Type II error simultaneously. You cannot make a Type II error when you have a significant result. You cannot make a Type I error when you have a non-significant result.

Type I errors often occur. We want to find and publish significant results, and when we do, we always make a Type I error, by definition. When I say we often see Type I errors, it is only because we publish significant results.

You do not see Type II errors that often because we do not publish non-significant results. The profession keeps advocating that we should also publish non-significant results, and hence see more Type II errors, because null findings are just as important as significant findings. Until our profession stops favoring significant findings for publication, you will see more Type I errors than Type II errors.

If a statistical test falls right on .05, what do you decide? Think of the .05 as the border on a sports field. In baseball, soccer, and tennis, if the ball falls right on the line, then the ball is considered “fair” or “in play,” and we keep playing. So, if a result falls right on the line, it is considered significant. What would it mean to have a p value result of .049? Is it significant? The answer is yes because rounding up .049 to .05 still means the result falls on the significance of the line. Side note: It is curious that in other sports, like basketball and football, if the player or the ball falls on the line, it is considered “out of bounds,” and the play stops. Why the line means two different outcomes across sports is a mystery to me.

You could take this problem of “on the line” even further. What happens if the p value result is .051? Do you round down and say the result is at .05? Or do you be honest and say, well, the p value is technically right above the .05 line, so technically, it is not

significant. Or if you have a result that is at .054. Like most of us who learned this math concept in high school, any value of 4 and below gets rounded down to the lower value, and any value of 5 and up gets rounded up to the higher value. Should you round down a .054 to a .05 and call the result significant? Or you can be honest and say, well, this value is above the .05 line, so technically, it is not significant. But what if we extend the decimal places to .0501? Now that 1 in the millionth place of the decimal, does that mean the result is not significant?

If you think, “This is like hair-splitting, and it is giving me a headache,” you would be correct. Notice how we place WAY TOO MUCH weight on the $p < .05$ as the “make or break it” rule for statistical significance. It is not good to make major decisions on a .05 line. Later in this chapter, we will discuss the futility of making decisions based on $p < .05$.

Is there any way to fix this error? The error is making a mistake. I guessed the significance of the study, but it was wrong. Conceptually, this error will occur every time you make a guess about the significance of the statistical test. The only way to fix Type I and II errors is to prevent a wrong guess by replicating the study.

You will never know the truth about an outcome. All you can do is be confident that you think you did everything you could to get to the truth. Because the definition of confidence is minimizing the chance that you are incorrect in your conclusion, the only way to maximize your confidence is to continue collecting observations and replicating the study until you find the same result repeatedly. If something happens over and over again, if you get the same result, no matter what you do, then you have confidence. Recall the definition of insanity is doing the same thing over and over again and expecting a different result. The definition of confidence is doing something in different ways over and over again and still getting the same result no matter what you do. The definition of confidence is the reason why we can confidently claim that the Chicago Cubs stink. No matter the circumstances, no matter who is on the team, no matter who manages the team, no matter where they play, no matter the changes in baseball rules, no matter the time period, they still stink. On the other hand, the Chicago White Sox are superior. No matter who is on the team, no matter the circumstances, no matter who manages the team, no matter where they play, no matter the time period, they still are winners. That is the definition of confidence (I am taking artistic writer’s liberty here. My book. My examples).

The concept of the Type I and Type II errors is another reason you need a large sample size AND replication. You can’t base your conclusion on one statistical test or one study. We have to replicate to see if we get the same result no matter what we do. Soapbox moment – our field and profession seem to frown upon replicating results. We seem to value original research and new contributions. There is much value in replicating results, especially results that are suspect or could use updating. Never let anyone tell you that replicating a study is not valuable. It is valuable.

It is possible to manipulate the study design to obtain a significant p value. All these scenarios lead to a Type I error. You say you found an effect, but you really did not. And you did not because you manipulated the study design to stack the chances in your favor of finding a significant effect. One method is to increase the sample size. Among statisticians this method is called “fishing.” The more you sample, the more likely you are to find an outcome. The more you cast your fishing line into the lake, the more likely you are to catch a fish eventually. Does catching a fish make you a fisherman after, say, 300 tries? Nope. It means if you keep looking, eventually you will find something. Interestingly, if you sample 100 times, according to the normal distribution, five of those observations will be significant simply due to random luck. But it does not mean you found something.

A second method is to conduct more statistical tests with lots of outcome variables. The statistical term for this is “increasing the Type I error rate” or “increasing the experimenter-wise error rate.” This method is like fishing. If you have 100 dependent variables and run 100 statistical tests, eventually, at least five of those tests will yield a significant result, simply due to chance alone. This scenario is conceptual. It is hard to believe that a) any study will have 100 separate dependent variables, and b) there is nothing about conducting a statistical test 100 times with the same sample that will alter the significance of results.

A third method is to increase the sampling, which usually leads to increasing the sampling range or getting more varied participants in the sample. The more variations, the more likely you are to find a pattern in those variations. If you have a range of variables from 1 to 10, and the sample you obtain is on the high end, such as from 6 to 9, your range is truncated, and you likely will not get a significant result because there is not much of a range to establish a pattern. If you keep sampling, you will likely find some observations that fall in the lower and upper regions of the variable. You will get observations below 6 and more observations that are 10. The more range in the sampling, the more chances there are to find an effect.

The lesson in all of this? The research design, in this case, the sampling frame and the number of dependent variables, can be engineered to find a significant effect. The lesson is to always conceptualize. By conceptualizing the relationships between your

variables, you can determine if the effect you found tracks or aligns with your expectations of how those variables should be associated with each other.

Which error is worse? Type I or II? It depends if you want to find an effect or not. When you are testing treatments, you want to be confident that you have found an effect, so you don't want to make a Type I error, that is, you only take results that are very unlikely to occur by chance, so that you can be confident that you have a real treatment effect. When making diagnoses, or you want to err on the side of caution and not make a Type II error. If you assess a child and say they do not have autism when, in fact, they do, that is a Type II error. You do not want to miss the diagnosis, so you might err on the side of caution to avoid making a Type II error. If you tested a treatment for migraine headaches and some patients did get better, but overall, your treatment showed no effect; you made a Type II error. But you might still offer the treatment because even if the treatment did not work for everyone, it might work for you, and if you suffer from migraines, you will take any chance at treatment, particularly if there are minimal risks involved with the treatment. So, there is no inherent value in whether Type I errors or Type II errors are worse. The answer is based on your conceptualization of the research issue. If we had to default to something, though, Type I errors would be more problematic. Everyone wants to find results because results lead to publication and funding for more research. We tout our results, but we are probably making Type I errors every time we think we found something significant. If anything, we are probably more prone to making Type I errors.

If you are wondering what to do to minimize Type I and Type II errors, the answer is always to replicate the results to confirm the answer. If you are on the fence about whether your result is a Type I or Type II error, then a consultation is needed. The problem with the p value is that EVERYTHING is based on the p value. We will return to this issue at the end of this section, but it is not good to base our decisions on one piece of information that is fickle. Certainly, you do not date a person simply because they make over \$80,000 per year. What if someone makes \$78,000 per year? Do we dismiss that person??? The answer is no, because there are other qualities we desire in a person, such as personality, kindness, and sense of humor. If a person has other good qualities, the \$78,000 per year is still good, even if their salary does not make the cutoff. Yes, give a person a chance to date you, even if they do not meet all of your qualification thresholds.

Turning back to psychology statistics, we have to look at information other than the p value: The sample size, the means, the standard deviations, the goal of the study, and the sampling method. All of those are considerations. How they are all put together to make a decision is discussed later. For now, keep in mind that when you are fussing over if a result is significant at $p < .05$, disengage from the p value and look at other factors to help with your decision. And always consult. It is not good practice to declare significance, and hence the value of the study results, based on one piece of information – the p value. More later.

8.10.4: Type I or II Errors and Why We Set the P value at .05

Who died and said for all time, “The p value is at .05?” What happens if you move this sacred special p value of .05 by increasing or decreasing it? Or why not just set the p value at .05 for all time and be done with it? Why tinker with it? And do you even give a rat's ass about this?

The answers to these questions are a) increasing or decreasing the p value increases or decreases the chances of making a Type I or Type II error, b) we set the p value at .05 because it is the best compromise of making a Type I or Type II error, c) you care because to guard against consequences, or to increase the range of possible positive outcomes, there are times when you prefer to make a Type I or Type II error.

BTW, there is no good way anyone can explain this concept. It involves a lot of thought, and you have to study the material to understand this concept. There is no way to only read about this issue and have that be sufficient for you to understand this concept. You must come up with your own way of thinking about this concept.

The p value set at .05 means that you are only interested in results that occur by chance 5% of the time.

You would be right to ask this question:

Q: “Why not set the p value at .01? That way, you can only get the significant results.”

A: “Because with such a strict and narrow range, you might miss out on a significant result.”

Conversely, you would be right to ask this question:

Q: “Why not set the p value at .10? That way, you can get a lot of significant results.”

A: “Because with such a loose and wide range, you will get results that are actually not significant.”

These situations illustrate what happens when you increase or decrease the p value; doing so changes how you decide if a result is significant or not significant.

Keep in mind that Type I and Type II errors are inversely related. As one error increases, the other error decreases.

To illustrate this point if alpha is set at 5%:

- If you decrease the alpha level (i.e., decrease alpha to 1%), you lower the possibility of committing a Type I error but increase the likelihood of committing a Type II error.
- If you increase the alpha level (i.e., increase alpha to 10%), you lower the possibility of committing a Type II error but increase the likelihood of committing a Type I error.

To demonstrate this, let's use a score result as our example:

- Coping skills: The range is 1 to 10; a score of 10 indicates good coping skills.
- The TX group's mean score is 8. The control group's mean score is 5.
- Set $p = .05$.
- The difference between the TX group and the control group is 3 pts.
- Suppose we test if that difference is significant. Our t-test value = 3.50, $p = .03$.
- That test result is lower than $p = .05$; therefore, the result is significant.
- Automatically, I made a Type I error. I found a significant result, but I could be wrong; the result really is not significant.

What happens when I decrease the p value?

- Decrease p value from .05 to .01.
- Let's use the same t-test value = 3.50, $p = .03$.
- Now, with $p = .01$, that same result becomes not significant.
- I have now made a Type II error. I did not find a significant result, but I could be wrong; the result is significant.
- So, decreasing the p value increases the chance of making a Type II error.

Let's take another score result.

- Coping skills: The range is 1 to 10; a score of 10 indicates good coping skills.
- The TX group's mean score is 7. The control group's mean score is 5.
- Set $p = .05$.
- The difference between the TX group and the control group is 2 pts.
- Suppose we test if that difference is significant. Our t-test value = 1.70, $p = .07$.
- That test result is higher than $p = .05$; therefore, the result is not significant.
- Automatically – I made a Type II error. I did not find a significant result, but I could be wrong; the result is really significant.

What happens when I increase the p value?

- Increase the p value from .05 to .10.
- Let's use the same t-test value = 1.70, $p = .07$.
- Now, with $p = .10$, that same result becomes significant.
- I have now made a Type I error. I found a significant result, but I could be wrong; the result really is not significant.
- So, increasing the p value increases the chance of making a Type I error.

So, we decrease the p value (.05 to .01), and we increase the Type II error. We increase the p value (.05 to .10), and we increase the Type I error.

There is no way around committing one of these errors. There is no situation where we are certain that we are not committing a Type I or a Type II error. No matter what we do to the p value, there is always uncertainty.

And that is why we set the p value at .05 because that's the best of both worlds, or put differently, that is the best value we can set to guard against making a Type I or Type II error.

8.10.5: Why Increase or Decrease the P value from .05?

If $p < .05$ is the best of both worlds, why bother adjusting it?

Q: “Why fuss with this? Just keep the p value at .05 and be done!!”

A: “Because there are situations when you want to increase or decrease the p value.”

BTW, how do you increase or decrease the p value? It's very easy. You just declare it. You say, “instead of $p < .05$, it is now $p < .01$ ” or “instead of $p < .05$, it is now $p < .10$.” There is no mathematical formula, nothing for statistical programs to do; you enter the desired p value in an “options” section of the statistical program. And why .01 or .10? Conventional guidelines. It is much easier to understand .01 or .10. You could set it to anything you want, but nothing important is added by setting it to .03 or .12.

In what scenario would you want to decrease the p value from .05 to .01? When you want a strict range and only result with a very low probability of occurring by chance. Remember, the .01 is 1%, which means you have very high standards. Conceptually, you only want the best of the best when you set your alpha level at .01.

Let's take a humorous example. Your friends say that their criteria for selecting a potential date is someone who makes \$80,000 per year. But you set your salary criteria at \$100,000 per year. You lowered your p value from .05 to .01. You are raising your standards and restricting the range of salary recruitments for your potential date. Nothing wrong with having high standards because you only want the best of the best. However, the trade-off is this: you might miss out on a great partner simply because they make \$90,000 per year. If you are okay with that trade-off, then so be it.

Let's take a medical example. You are testing medications for cancer that really need to work. You set $p = .01$ because you only want those medications that have a high probability of having a treatment effect, and you do not want any medication that does not have a treatment effect. You do not want a Type I error. You don't want false positives. You do not want medication on the market that you say works but really does not. No false hopes.... However, the tradeoff is this: you might be missing out on effective medications simply because they did not quite get the effect you wanted to see across everyone who received the medication. If you are okay with that trade-off, then so be it.

In what scenario would you want to increase the p value from .05 to .10? When you want a loose range, and you want more results. Remember, the .10 is 10%, which means you are lowering your standards. Conceptually, you only want a wider range of results when you set your alpha level at .10.

Let's take a humorous example. Your friends say that their criterion for selecting a potential date is someone who makes \$80,000 per year. But you set your salary criteria at \$50,000/year. You raised your p value from .05 to .10. You are lowering your standards and widening the range of acceptable salaries for your potential date. There is nothing wrong with lowering standards because you want a wider variety of potential partners to choose from. However, the tradeoff is this – you might be dating a partner who is not that great for you because they make \$50,000 per year. In other words, you are settling. If you are okay with that trade-off, then so be it.

Let's take a medical example. You are testing medications for migraines. Set $p = .10$ because if the medication has any chance of working, you will take it, assuming there is minimal risk to taking the medication. If you have a migraine, you will try anything that works. It may not have worked for others, but if it worked for you, then great. If it did not work for you, if it was a false positive, a Type I error, then fine. As a migraine sufferer, you move on and try the next medication because you will try anything that might work for you.

At the end of the day, unless you have a conceptual reason for shifting the p value because there is something of value regarding the need to be comfortable with making a Type I or Type II error, you are better off leaving it alone or leaving it at $p < .05$. Never shift the p value to make your results and yourself look better.

8.10.6: Here is the Kicker: Abandon P values

A valid response to all this discussion about p values is “what a lot of fuss!”

A takeaway message from the discussion about p values is that significance values can be easily manipulated. You simply increase or decrease them. You can make anything significant by increasing the p value; you can make anything not significant by decreasing the p value. It is a fickle piece of statistical information.

When you read these phrases, the messiness of making conclusions based on the p value is compounded.

- $p < .001$
- “Highly significant”
- “Trending towards significance” or “approaching significance.”

Never use any of the above phrases. The problem with all these phrases is that people think there is some value in the number itself when interpreting a significance value. The p value is simply a cutoff value. The only question you can ask when reading the p value is “Is the statistical test significant?” and the only two answers to this question are yes or no. There is no such thing as highly significant. Just because a p value is .0001 does not mean the test result is better than a statistical test result whose p value is .01. When someone says that a result is highly significant, what does that mean? Why is one statistical test with a .0001 significance level better than one with a .01 level? A more significant result means that it is not likely to occur by chance. But anything can have a low probability of occurring by chance. There is no need to directly compare the importance of the results by comparing their significance value. P values are not intended for that purpose.

Listing p values such as $*p < .05$, $**p < .01$, $***p < .001$ is a practice with no meaning. Results are significant or not significant. There is no such thing as more significance based on the value of the significance test.

The phrases “trending towards significance” or “approaching significance” are misleading. Researchers think that a $p = .06$ is “close enough,” so they might as well call them significant. Essentially, this practice is akin to a Type II error. You are saying that the results are not significant, but in fact they are, so you might as well call them significant. There simply is no basis, evidence, criteria, or consensus that a .06 statistical significance should be considered significant when $p < .01$. For that matter, what makes a .07 not significant, but a .06 significant? Is the difference of 1% enough to make a study’s results not significant? The reasoning is a stretch.

Everything discussed leads you to say that from this point forward, you hate the significance test and the p value. We simply cannot put so much emphasis on one piece of statistical information, the p value, that will make or break a study. As of this writing, there is a growing movement among statisticians to abandon the p value. If statisticians say to abandon the p value, then you ought to listen. The movement to abandon the p value does not mean ignoring the p value altogether. Statisticians are saying to never put so much weight on deciding the value of a study and its statistical results, based exclusively on the p value. Keep in mind that statisticians want us not to rely solely on the p value; they want us to use the p value as just one piece of statistical information, along with other pieces of statistical information and the research design, and to inform us as we evaluate the validity of a research study.

Helpful Websites

<https://www.winginstitute.org/does-sample-size-correlate#:~:text=Small%20sample%20size%20studies%20produce,diminished%20with%20increasing%20sample%20size.>

This page titled 8.10: The Definition of Type I and Type II Errors is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Peter Ji.

8.11: Discussion Questions

Discussion Questions

1. What are the steps to determine if a result is significant?
2. Why do we want to know if something is significant?
3. Provide your own metaphors to describe a Type I and a Type II error.
4. Why should we not put so much emphasis on p values?

This page titled [8.11: Discussion Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

CHAPTER OVERVIEW

9: Effect Sizes and Power Analysis

Learning Objectives

By the end of this chapter, you will be able to:

- Understand the value of effect sizes
- Examine how to do a Power Analysis

Key Terms

- Effect Size
- Power Analysis

[9.1: Significant \(P Value\) Versus Meaningful \(Effect Size\) Results](#)

[9.2: What Are Effect Sizes?](#)

[9.3: Statistical Power](#)

[9.4: How to Conduct a Power Analysis](#)

[9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together](#)

[9.6: Discussion Questions](#)

This page titled [9: Effect Sizes and Power Analysis](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

9.1: Significant (P Value) Versus Meaningful (Effect Size) Results

What do we do if we shouldn't emphasize the p-value so much? To address that question, let us remind ourselves of the meaning behind the p value. The p value as a significance test only indicates that something happened, or put differently, the p value answers the question, "did something happen?" The p value provides an answer to that question, and that answer is a hard "yes" or "no."

The problem is that significant results do not mean meaningful results. Even small differences can be significant, but the differences are too small to be meaningful. The term we use for meaningful results is called the effect size.

As of right now, you HATE the p value, and instead, you LOVE effect sizes.

Effect sizes describe the magnitude of the difference between two groups or the strength of the relationship between two variables. Effect sizes are a way to measure the amount of change that occurs. Effect sizes address questions such as, "how much impact did the treatment have on the outcome?" or "how important is this variable in helping me predict an outcome?" Think of effect sizes as similar to "clinical significance." While you may observe a change in your clients, is that change clinically meaningful? Some examples:

- The client had a dramatic improvement in MMPI functioning from 75 to 35 and went from a clinical problem to normal functioning. This situation is a large effect because it is quite a transformation.
- The client may have reduced drinking from 20 times a week to 15 times a week, but that is still a lot of drinking. This situation has a small effect because although something changed, there was not a change in the client's drinking to a different, manageable level.
- Adding mindfulness training increases the ability of clients to manage trauma stress from one day a week to six days a week. This situation has a large effect because it is quite a transformation.
- Adding mindfulness training increases the ability of the client to manage abstinence from alcohol from 30 days sober to 40 days sober. This situation has a small effect because although something has changed, there was not a change in the client's drinking to a different, manageable level.

This page titled [9.1: Significant \(P Value\) Versus Meaningful \(Effect Size\) Results](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

9.2: What Are Effect Sizes?

Effect sizes are expressed as standard deviation units. Recall that a standard deviation establishes a metric for the distance of a score away from the mean. When an observation is one standard deviation away from the mean, we know the location of the score and how far the score is from the mean. In this case, the score is higher than 34% of observations that scored at the mean. What is good about using a standard deviation unit is that it is “standard,” no matter the range of the variable, the standard deviation unit remains the same. Recall that one standard deviation unit encompasses 68% of the scores around the mean, two standard deviation units encompass 95% of the scores around the mean, and three standard deviation units encompass 99% of the scores around the mean. These units and their corresponding percentages around the mean stay consistent no matter what the range of the variable is.

9.2.1: Different Effect Sizes

There are two types of effect sizes. Recall that there are two types of patterns we find with statistics. Note that these are patterns associated with general linear model statistics. We will discuss this concept when we turn to statistical tests. For now, we find there are two patterns. The first pattern is where group A is higher or lower than group B. We term this pattern as ‘mean differences among groups.’ For this pattern, the effect size we want to see is how far apart the mean scores are for each group. The greater the difference between the group means, or the greater the distance between the groups, the larger the effect size.

The second pattern is where increases in variable X are associated with increases or decreases in variable Y. We term this pattern “associations.” The stronger the association between the two variables, the larger the effect size is.

The following are the common effect size types. For mean differences between groups, we use the following effect sizes.

- Cohen’s D: Cohen’s d is a statistical measure that compares the means of two groups by standardizing their findings onto a common scale. Cohen’s D is a population estimate of the effect size.
- Hedges G: Hedges g is a statistical measure that compares the means of two groups by standardizing their findings onto a common scale. Hedge g is a sample estimate of the effect size.

For associations between variables, we use the following effect sizes.

- Correlation coefficient
- Standardized beta coefficient
- Eta-squared describes the ratio of variance explained in the dependent variable by a predictor while controlling for other predictors

9.2.2: Calculating Effect Sizes

The basic formula for an effect size is the following: $\text{mean (group 1)} - \text{mean (group 2)} / \text{pooled standard deviation}$. Of note with this formula is the denominator, which is the pooled standard deviation. Dividing by the standard deviation tells you by how many units the two means are apart. It is like saying you have a lot of dough to make lots of cookies. You may have a large pile of dough, but unless you know what size cookie you will make, you do not know how many cookies you can make out of that dough. If you have a large cookie cutter, you could only get so much out of the pile of dough compared to if you have a smaller cookie cutter. The size of the cookie cutter is analogous to the standard deviation. We divide our pile of differences between the two means, then use the standard deviation as our cookie cutter to determine how many units that difference is.

With the advent of statistical programs, there is no need to calculate an effect size on your own. Most statistical programs and their associated analyses have options to click to obtain the effect size. If you see the option, click on it. If it does not have the effect size option, effect size calculations are available online.

Despite the importance of effect size, which we will discuss later, journal articles do not report the effect sizes. If there is no report of an effect size, as a rough guess of an effect size, you could subtract the difference between the two group means, take the average of the standard deviation between both groups, and do a rough division of the difference divided by the . You obtain a rough estimate of the effect size. Of note, if we really need to read about effect sizes, why then do journal articles do not report them? Well, if the authors of the article do not report it, or the journal editors do not request the authors report the effect sizes, then they are idiots. APA and statistical organizations all but require effect sizes to be reported. Oh well.

9.2.3: The Range of Effect Sizes

Effect sizes are expressed as standard deviation units. The word “standard” means that everyone knows what these sizes mean.

The range of effect sizes is usually 0 to ~1.5 or slightly above. One rough way to read this range is that 0 means 0 standard deviation units, and 1.5 means 1 ½ standard deviation units of difference.

The general guidelines for reading an effect size are as follows: Small effect size = .3, Medium = .5, Large = .8. I like to think of effect size as a proportion of the standard deviation. I use these guides: .3 would be 1/3 of a standard deviation, .5 is ½ of a standard deviation, .8 is over ¾ of a standard deviation.

I like to think of the effect size as a measurement of the size of an ingredient. A small effect size of .3 is like putting 1/3 of a cup of sugar in your coffee. A medium effect size of .5 is like putting ½ of a cup of sugar in your coffee. A large effect size of .8 is like putting more than ¾ of a cup of sugar in your coffee. And I am sure you know people like this, probably even someone like you.

I like to think of effect sizes as how transformative the effect is. Small sample sizes produce effects, but they are hard to notice. This metaphor is a bit of a stretch, but let's try it. Putting 1/3 of a cup of sugar in your coffee will have a small effect on you, or you will get through the day, but no one will notice any visible change in you. Putting 1/2 of a cup of sugar in your coffee will have a moderate effect on you, or you will get through the day with an extra spring in your step, and quite likely, a few people around you will notice the change in you. Putting 3/4 of a cup of sugar in your coffee will have a large effect on you, or you will get through the day and probably through the night, and everyone around you will notice the change in you.

9.2.4: What Effect Sizes Do You Want to See?

Small, medium, and large effect sizes are simply descriptive. Based on the above examples, it would seem like large effect sizes are valued because we want to see notable impacts of our studies. But evaluating effect sizes is based on the context of the study and within the field discipline. There is no hard and fast rule indicating that large effect sizes are good and small effect sizes are not worth noting.

Psychology as a social science is quite an open system. So many variables are out there that can influence something. There are so many confounding, alternative variables that could adversely affect your conclusions. We do struggle with how to obtain precise measurements of our variables. Many variables are social constructs. In psychology, the concept of personality is a social construct. We all know it is there, but it is a latent construct; we have to infer that there is such a thing as personality. Racism is a construct. We all know it is there, but we have different ways of conceptualizing how racism varies, so it is hard to create one measurement of racism. Even variables that we think are easy to measure are hard to measure. Counting the number of alcohol drinks a person consumes is difficult because of the multidimensionality of alcohol. One glass of beer, one glass of wine, and one shot of whiskey could all count as one drink, but the alcohol content differs, so it is hard to state that they all count as one drink. To complicate matters, having one drink per night to relax after a long day, versus one drink per hour while you are at a watch party for your favorite sports team, would probably lead to different outcomes. In this case, drinking alcohol is compounded by two other variables, time frame and context. It is hard to say that we have good measurements of variables. These factors, and others, make it difficult to find large effect sizes. It is difficult to find one or a few variables that have a notable impact on the outcome.

Any research discipline would covet large effect sizes because a large effect size implies a large, notable impact. It is akin to finding the “magic cure” or the panacea. In psychology, we certainly want to find those variables that have a large impact on our outcomes. Nothing wrong with that desire. I would offer the field of psychology, and its theories and stages of change, would propose that there are no “magic pills” or a best treatment that would cure someone’s mental health struggles. Fast, large effects wear off quite quickly. It would be akin to a fad diet. You might lose weight very quickly, but our bodies, or physical capacity, just do not adapt well to sudden, rapid shocks to our system. Large effect sizes might show immediate effects, and obviously, we must account for context, but large effect sizes might not be all that promising.

Ostensibly, a low effect size means that the effect of the independent variable on the dependent variable is weak. The weaker the effect, the less prominent the effect of that variable is on the outcome. In psychology, because there are so many variables that impact an outcome, and because of imprecise measurements of our variables, research designs in psychology cannot possibly include every variable in their study; low effect sizes are bound to occur. Getting an effect size of .3, a small effect size, is good, although there is no industry standard or consensus indicating a small effect size is good enough.

In psychology, low effect sizes might be and should be desired. The goal of psychology is to improve mental health. Change does not happen overnight. Changes happen slowly, and for good reasons. It takes time for our bodies to adapt to change, even for the better. In psychology, people resist change, even if it is for the better. It can take a long time for a change to take hold. Changes are not always noticeable at first. And we relapse. Some days we can handle everything, some days we cannot. In this context, a low effect size might reflect the realities or the difficulties of how people change. It is a slow process, but slow is probably better as we adapt to change.

Low effect sizes might align better with treatment planning. Small behavioral changes are victories. Sometimes, just getting out of bed is a victory. Sometimes, just avoiding an alcoholic drink for one night of the week is a victory. Sometimes, just reading a book for 30 minutes a week is a victory. These changes might not seem big, but they are big to us. Small changes, small effect sizes, might prove to be worthwhile over the long run for the lasting change we want to see as we recover from our mental health struggles. As they say in exercise, “progress, not perfection.”

Are medium effect sizes the best of both worlds? This author's position is to be agonistic. As always, default to your conceptualization of the issue and the variables under consideration and use a literature review to catalog similar effect sizes, as all of your conceptualization is about the value of a medium effect size for your study and its outcomes.

9.2.5: Effect Sizes and Other Considerations

Effect sizes are not directly impacted by other statistical considerations, such as the sample size or the p value. You might think they are related, but conceptually, the effect sizes are independent of other statistical considerations.

Effect size and sample size. A low sample size says nothing about the resulting effect size. Low sample sizes tend to produce unstable results, so it follows that the effect sizes would be unstable. Effect sizes do not decrease as the sample sizes increase. You might think that more sampling means more variation, which means more possible errors, but that is not the case. Do effect sizes increase if the sample size increases? That is not the case, either.

The only thing increasing the sample size will do is increase the chance of detecting an effect. Increasing the sample size is not solely responsible for a corresponding increase or decrease in the effect size itself. This issue is basically a power analysis. More on that later.

Effect sizes and the p value. P values that indicate non-significance are associated with a low, almost zero effect size. If there is no significant pattern in the variables, it is not significant, and there are no effect sizes to detect. If there is a significant p value, a lower p value does not mean it is even more significant, and the effect size does not correspondingly increase. Lower p values have no relationship to the effect sizes. We already established there is no such thing as “highly significant at $p < .001$,” consequently, there is no such thing as a high effect size when the p value is .001, vs. a moderate effect size when the p value is .01.

This page titled [9.2: What Are Effect Sizes?](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

9.3: Statistical Power

How do we know that an effect is there to detect? That question involves the concept of statistical power. The power of the statistical test is the likelihood that we will be able to detect an effect if indeed that effect does exist.

Power is yet another reason why we are so fixated on getting a large sample size. When we say that we need to collect more data or get a large sample size, we mean get enough power to detect an effect if indeed the effect does exist. In this context, power is akin to sample size. Think of power as fuel. Do we have enough power, sample size, and fuel to keep sampling so that we will find the effect that we say is there to detect?

In some ways, we could always say, “Well, why not keep sampling, observing, and collecting data until you get what you want? What do we need a power analysis for if we are to keep collecting data anyway?” We keep saying that we need to collect more data, but how much data should you collect before saying, “Stop, I’ve had enough?” The purpose of the power analysis is to address this issue: What is the minimum sample size needed to detect an effect if the effect does, in fact, exist.

9.3.1: What are the Values of Power?

Power ranges from 0 to 100%. Zero is obvious; your sample size is so low that there is no opportunity to find the effect. 100% means that your sample size is so high, you have a 100% chance that you will find something, assuming there is something to find. 100% is unlikely to be achieved. There is conceptually no situation where you have all the power you need to detect everything. That’s like saying your car has unlimited gas for fuel; you can keep going and going until you get to where you want to go.

A power value of .50 means you have a 50% chance of finding an effect. Well, that is akin to flipping a coin. Some days you might find the effect, some days you might not. Therefore, 50% is not a great number. If 100% is unrealistic, what is a good percentage? That would be 80%, which gives you just enough confidence to find something, but it is not over the top, where you must have a large sample size to find something. More about why getting a large sample size is problematic will be discussed in the later sections.

9.3.2: Power and Type II Error

Power addresses the problem of Type II error. Recall the Type II error – you get a non-significant result, but actually, your result is significant. Usually, this Type II error is due to having a low sample size. You did not look hard enough to find the effect. If you do not find something, how do you know that you did everything you could, or sampled as much as you could, to conclude that you found nothing?

Consider this example. You lost your car keys. You looked in your car and you did not find them. Conclusion: You lost your car keys. You might have made a Type II error; you think you found nothing, but in reality, your car keys are still out there. What does your roommate, bestie, or family member say to you? “Did you look hard enough? Did you look everywhere?” You keep looking, and eventually you find your car keys in your pants pocket in your bedroom. You concluded you found nothing, and by default, you made a Type II error. But you did not find anything because you did not keep looking for your keys, or you did not sample more places to find your keys. Then you found your keys. Your original effort was underpowered or insufficient to find your keys.

Consider this example. You come home from one date, which was bad, and you conclude that you will be alone for the rest of your life. You might have made a Type II error. What does your roommate, bestie, or family member say to you? “Keep going on dates because eventually you will find someone.” You keep going on dates, and eventually, after the third try (because the third time is the charm), you find your soulmate. You concluded you found no one, and by default, you made a Type II error. But the reason you did not find anything is that you did not keep going on more dates. The more dates you go on, the more likely you are to find your soul mate.

In psychology, we investigate treatment effects. Sometimes our treatment does not show an effect. But maybe if we had more participants, the effect would emerge. That said, the power analysis will tell us how many participants we need to recruit into the study to find the effect we hope to find.

The power analysis will “rule out” a Type II error. “Rule out” is in quotes because the purpose of the power analysis is not to rule out a Type II error. The power analysis will tell you the sample size you need to find the effect. If you did not find a significant effect, and you did sample enough participants, you can reasonably conclude that you did not find a significant effect. In this case,

the likelihood of having made a Type II error goes down. Consider the car keys example. If you looked everywhere, your bedroom, your desk, every clothes pocket, then yes, your keys are definitely missing. Consider the dating example, if you go on, say, 300 dates, and still cannot find your soul mate, then yes, you will not find your soul mate. If that is the case, then, well, work on yourself, lower your expectations for a soul mate, or re-examine your life. In psychology, if you sample 300 participants and still cannot find your treatment effect, your treatment does not work, and you can be certain that it does not after testing it with 300 participants.

9.3.3: Why not Get Large Sample Sizes Anyway?

The power analysis tells you the minimum sample size you need to find the effect if, in fact, the effect does exist. Well, we are going to sample a lot anyway, so why not just keep sampling as much as you can? Why not always get large sample sizes, like $N = 1000$, so you have all the chances to find an effect?

A couple of reasons against large sampling. Recruiting many participants costs a lot of resources. Recruiting and collecting participant data requires time, personnel, money, and sleep (we all need rest). And if you are a graduate student, you do not have four things in your life: time, friends, money, and sleep.

Most clinical populations are not that large. Clinical populations have a low base rate and are rare in number. For example, people with rare medical diseases. In clinical psychology, it is not so much that there is a low base rate, but due to stigma or awareness, some clinical populations are hard to find. Middle Eastern women who experience domestic violence in their marriages are hard to recruit because of the cultural stigma of seeking help. Adult males who engage in self-harm behaviors might not readily participate in research because of the stigma of being an adult male who engages in self-harm behaviors. Cubs fans who are intelligent are hard to find because what intelligent person becomes a Cubs fan? Intelligent people become White Sox fans. The lesson is that sometimes it is difficult to get the sample size we want, given the characteristics of the issue and population at hand.

Large sample sizes usually result in Type I errors. You can keep sampling, and eventually you will find something, but you must decide if you did find something because it is a real effect or if you simply got lucky because you kept looking.

Getting a large sample size is prohibitive for these reasons. Power analysis is meant to guide you in being efficient in your sampling strategy.

Even if you did get a number for your sample size from the power analysis, it behooves you to still sample extra participants. Every research study has missing data and missing responses from surveys. People who drop out midway through a battery of surveys or drop out during the experimental treatment. Having more participants and more observations will allow you to make up for the missing data. How that works is discussed in later chapters.

9.3.4: What Helps Get More Power?

Power seems synonymous with sample size. But there are other ways to get more power or find the effect you want to see. Put differently, how else can we give ourselves the best chance to find the effect?

The first way is to increase the sample size. I have already discussed this at length.

The second way is to increase the alpha level, say from .05 to .10. By increasing the alpha level, you are giving yourself more chances to find something. Increasing the alpha level is akin to widening the region of rejection under the normal distribution. It is akin to lowering your standards. You want to find an effect? Lower your standards, and you will find someone who will meet your lower standards. But that does not mean you found something; you just lowered the bar and said that anything will serve as an effect. I am not a fan of this option because increasing the alpha level increases power but also increases the chances of making a Type I error. Essentially, you are saying, “anything can be significant.”

The third way is to increase the desired effect size. If the true population means are very different, the difference is very apparent. It will not take much to find this difference because it is easy to find. The statistical test will have high power because it will not take many observations or data to find the effect. Let us say you want to find out if height is related to making basketball free throws. Are taller players more likely to make basketball free throws? . You sample 100 6-foot players and compare their free throw percentage with 100 6-foot-5-inch players, and you find that height is not related to free throw percentage. But that difference of 5 inches, especially at that height, probably won't make much of a difference. What can you do? You can compare the 6-foot players to, let's say, 4-foot preschoolers. Lo and behold, the 6-foot players make more free throws than the 4-foot preschoolers. And you

only sampled 10 6-foot players compared to 10 4-foot preschoolers. You found your effect: height is related to free-throw shooting. But the only reason you found the effect is because you stacked the deck in your favor. You artificially increased the actual effect – 4-foot preschoolers vs. 6-foot players. So, I am not a fan of this method because it seems manipulative.

I am also not a fan of increasing the desired effect size because it is hard to know the effect size in advance. What kind of effect, or impact, will there be for mindfulness on alcohol abuse? What kind of effect, or impact, will there be for social support and the likelihood of coming out as LGBTQ? These effects are hard to determine in advance. The best way to estimate the effect size is to do a literature review and catalog what kind of effects past studies found. That will give you a guide, but it is difficult to say the effect sizes you find in the literature will be based on the variables you are interested in.

So, of the three options to increase power, a) increase sample size, b) increase alpha level, c) increase the effect size, the best option is to increase the sample size. BTW, notice that all the options for increasing power are to increase something. If you find that question on the EPPP, if one of the answers involves an increase, that might be a good clue for answering the question.

9.3.5: Small Effect Sizes are Hard to Detect

If the difference is very small, it will take a lot (large sample size, more observations) to find this difference. This is the umpteenth reason why we need a large sample size. Effects are difficult to detect in small sample sizes, so you need more power, or larger sample sizes, to detect effects.

Consider these examples. Usain Bolt blew away the field in the 100-meter race in the 2008 Olympics. The gap between him and the second-place finisher was so large that it was convincing that Usain Bolt was the fastest. There was no need to make Usain Bolt redo that race and prove himself once again as the fastest runner. He did it anyway at the next Olympics and had the same result; case closed. No need for a larger sample

But Michael Phelps won his relay race by .01 of a second. That difference was so small that it is difficult to tell if he just got lucky in that race or if something happened. The difference was so small, it could have been random luck. Or there is something about Michael Phelps' swimming technique that put him in a position where he could outstretch his opponents and win the race by a fingertip. The only way to determine if it was luck or a real effect is to run the race again. You need a larger sample size to determine if an outcome was luck, hence unstable, because no one is that lucky twice, or if something is going on, such as great swimming technique. His 11 gold medals prove that he has the skill and speed to settle this debate.

9.3.6: Power Does not Always Mean Obtaining a Large Sample Size

We keep talking about power in terms of sample size. In fact, we seem to be obsessed with sample size and getting more of it. However, statistical tests do not rise and fall with sample sizes. There are other types of observations.

One observation is over time. You do not need to get more participants; sometimes, you can just observe the same person over time. We will discuss the benefits of this observation approach when we discuss longitudinal analysis. However, power can increase not just with an increase in sample size but also by adding more observations over time.

Another way to increase power is to use reliable measures. If the measures have little measurement error, then you have better precision, hence confidence that the number you generate is the actual number. Error kills off statistical significance. Any error enlarges the overall error term, and as we learned, the more error, the more the denominator increases, and any statistical test with a large error in the denominator is bound to be not significant. Better precision means higher reliability of measurement and a better chance of detecting an effect.

This page titled [9.3: Statistical Power](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

9.4: How to Conduct a Power Analysis

To conduct a power analysis, you first need a power analysis program. The most used freeware program is the Gpower program. The latest version of SPSS also has a power analysis. There are other online power analysis programs. I have no idea about the differences between them, and I am agnostic as to which is better.

To conduct a power analysis, you need three things: the p value, the effect size, and the desired power. You also need to know what statistical test you want to conduct and the design considerations of the statistical test.

The p value is the significance level, and the power analysis determines what kind of effect you want to detect at what significance level. Unless you have a conceptual reason to change the p value, the p value is set at the ubiquitous $p < .05$.

The effect size is the size of the effect you want to detect. Effect sizes range from .01 to above 1. You determine the effect size based on your conceptualization of the issue and past literature reviews of effect sizes from previous studies. A default response would be .4, between a small and moderate effect size. Hopefully, your effect will be sizable enough to be detected, but not so small that it will take a larger sample size to detect.

The desired power is the likelihood you want to see an effect. As mentioned earlier, .80 or 80% is a good default value for power.

All those values are entered into the power analysis.

This page titled [9.4: How to Conduct a Power Analysis](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together

Let us pause and do a process check on our inventory of questions and possible responses when making decisions about statistical test results. You should consider your effect size as a complementary piece of information with the p value. Here is a suggested order of questions and a list of simple answers when interpreting statistical test results. Remember, the KISS method – “keep it simple, silly.”

- Does the statistical test result indicate significance? Your response choices are yes or no
- If the answer is no, then there is no need to complete the remaining steps.
- If the answer to the question is yes, then you ask what the pattern is.
- There are only two patterns.
- Pattern one is - finding differences among groups.
- The statistical test for this pattern is usually a t-test or an F-test.
- The possible pattern is that Group A is higher / lower than Group B
- Pattern two is - finding associations among variables.
- The statistical tests are correlations or regressions.
- As Variable X increases, Variable Y increases / decreases
- If significant, are these findings or these patterns expected or unexpected, given the context of the research question? Your response choices are yes or no.
- If significant, what is the effect size? Your response choices are low, medium, and high.
- If not significant, there is no pattern
- If there is no pattern, Group A is like Group B, or
- If there is no pattern, we say there is no relationship between Variables X and Y.
- And there is no effect size. Is there no finding of a pattern that is expected or unexpected, given the context of the research question?

After you go through those decisions, the complementary question is, are these statistical results valid? That question entails examining statistical information and research design.

For the statistical information, you could consider these factors.

- Sample size. Is the sample size too low or too high? A sample size that is too low might be unstable in its estimates of statistical significance. A sample size that is too high, coupled with a significant effect, might yield a Type I error. Both scenarios cast doubt on the stability of the statistically significant finding.
- Sampling error. Are the means too high or too low? Sampling errors always occur, resulting in a sample mean that is too high or too low compared to the population mean. There is no way to gauge this issue because there is no population answer key about what the population mean should be. When you interpret the means, do they seem too high or too low compared to your theoretical understanding of what the means should be?
- The quality of the distribution of the variables. Normal distribution increases confidence in the results. Skewed distributions do not. However, some variables are meant to have skewed distributions. The question is whether the skewness is severe enough to warrant a different statistical analysis, such as a non-parametric analysis. More on that later.

Helpful Websites

<https://meera.snre.umich.edu/power-a...fect-size.html>

This page titled [9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

9.6: Discussion Questions

Discussion Questions

1. Create a chart or decision tree and connect these concepts together: sample size, p value, effect size, and power.
 2. Why should we care more about effect sizes rather than p values?
-

This page titled [9.6: Discussion Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

CHAPTER OVERVIEW

10: Introduction to Inferential Statistics

Learning Objectives

By the end of this chapter, you will be able to:

- Understand the goal of inferential statistics
- Describe the general linear model

Key Terms

- Inferential
- Patterns
- Models
- General Linear Model

Recap

In chapters one through four, we were in statistical concepts camp. In that camp, we learned the foundations, the concepts, the “why,” the hallmarks of statistics, the process of analyzing numbers, their variation, and how they represent the phenomenon we observe. Then, in chapters five through nine, we were in descriptive statistics camp. The purpose of descriptive statistics is to describe the numbers and how they represent the variation we see in the phenomenon we see. Describing these numbers means sorting, organizing, and simplifying the numbers in ways that tell us a story about the meaning of the variation.

In chapters ten and twelve, we venture into inferential statistics. One issue to highlight from the outset is the issue of causality. Statistics cannot determine causality. All statistics can do is determine if something is going on, or if something is a random occurrence. Only a research design can be used to determine if one variable causes the other variable to change. Never say that a variable causes another variable, or that a result is due to something. People freak out when you tell them that your statistical test indicated that something caused something else. People seem to insist that saying this is akin to committing a sin. While this is an overreaction, saying there is a causal relationship based solely on statistical analysis is inaccurate. But there’s no need to freak out over it. Just say, “something is associated with something else,” and everything will be fine.

[10.1: The Meaning of Inferential](#)

[10.2: Patterns](#)

[10.3: General Linear Model](#)

[10.4: Discussion Questions](#)

This page titled [10: Introduction to Inferential Statistics](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

10.1: The Meaning of Inferential

In inferential statistics, we make an inference, otherwise known as a guess or an estimate. We are guessing if there is a pattern in our results. We then determine if the patterns in our results are random or if something is going on that is responsible for the results we are seeing.

This page titled [10.1: The Meaning of Inferential](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

10.2: Patterns

There are many patterns. There are many statistical tests. We will start with the patterns obtained from the ubiquitous statistical tests in this book. The patterns we are observing always involve two variables. These two variables are the IV and the DV. Examples:

- Is gender (IV) associated with differences in help-seeking attitudes (DV)?
- Is teacher type (regular or special education) (IV) associated with better or worse ratings of autism behavior displayed by children (DV)?
- Does attending more counseling sessions lead to better outcomes?
- Does better social support mean better acceptance of one's LGBTQ+ identity?

When you have an IV and a DV, you are always looking for ways to describe how they are related or associated with each other. These associations or relationships among the variables are your patterns, or the results of your statistical tests. To recap, there are two basic patterns in statistics.

- As variable X increases, variable Y increases or decreases.
- Group A is higher or lower than Group B.

This page titled [10.2: Patterns](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

10.3: General Linear Model

These patterns or these results are generated by the statistical tests we will address in this book. Statistical tests are derived from statistical models. A model is how we describe the pattern among the variables. Think of a model as an equation of variables and the process of assembling the variables together to produce an outcome. Models are good for prediction. If we include these variables and describe the relationships between them, we can predict an outcome.

You can think of a statistical model like a recipe. A recipe has your ingredients; these are your independent variables. How you put the variables together is similar to how you describe the relationships among the variables. In baking, you put the dry ingredients together and the wet ingredients together. A recipe has an outcome, your finished product. In baking, you make cookies. The relationships among the ingredients, or your independent variables, will be associated with the outcome. One association is that the more sugar you put into your cookie, the sweeter the cookie. The more baking powder you put into your cookie, the more the cookie will rise. The more baking soda you put into your cookie, the more the cookie will spread (I learned this baking association – powder puffs, soda spreads). Another association is that if you use white sugar, your cookie will rise more and if you use brown sugar, your cookie will be denser. If you use regular flour, your cookie will be chewy and dense, and if you use cake flour, your cookie will be softer and lighter. We select our ingredients, put them together, and create an outcome. That is one metaphor you can use to learn how statistical models work and how statistical tests are used to predict an outcome.

These tests are based on the general linear model, which is based on frequency statistics, based on the central limit theorem. There are many ways to describe the general linear model, but for a starting point, think of the general linear model describing relationships between variables using straight lines. The linear model is one of the following scenarios: a) as variable X increases, variable Y increases, or b) as variable X increases, Variable Y decreases, or c) group A is higher than group B, or d) group A is lower than group B.

In psychology, here is an example of scenario

- A – as your client attends more counseling sessions, your client's positive well-being increases; scenario
- B – as your client attends more counseling sessions, your client's symptom severity decreases; scenario
- C – mindfulness treatment has more positive well-being compared to medication only; scenario
- D – mindfulness treatment has less symptom distress compared to medication only.

Of note, when describing linear relationships, it is best to always think of the variable X as increasing. This practice is good for describing relationships between several variables. When describing relationships, you want to keep these descriptions consistent, so it is easy to read. By stating that variable X is increasing across all the relationships, the reader does not have to do engage in cognitive mind flips to understand which variables are increasing or decreasing.

You are right to wonder if there are any situations where you can state, “as variable X *decreases*, variable Y decreases?” Technically speaking, that description is the same as this statement – “as variable X increases, variable Y increases.” Low scores on variable X correspond to low scores on variable Y. There are situations where that description fits the relationship between the variables. The less minutes you spend exercising, your health decreases. The less minutes you spend on homework, your test scores decrease. If your intent is to provide results to support those statements, then phrasing the relationship in terms of both variable X and Y decreasing could make sense. There is nothing wrong with those statements. Conventionally, we are used to reading relationships between variables as, when variable X increases, variable Y increases, or decreases, because it is easier to have the first variable increase, and then determine if variable Y increases or decreases. Phrasing the relationships between the variables takes practice and knowledge of the concepts you are trying to describe. While technically you can phrase the relationship between variables X and Y anyway you want, for ease of reading and keeping the relationships consistent, it is simpler to describe your relationships by starting off with stating that variable X is increasing.

The general linear model means that you want to think about your variables and the linear relationships they have with each other as you embark on your statistical analyses. These linear relationships are what you are looking for in terms of patterns or results.

There are other statistical models, and there are other ways of describing relationships between variables. The general linear model uses frequency statistics. It means that we count the number of times we see an observation and determine its significance based on how rare or infrequent it is. Other models, such as Bayesian models, use probability models. Briefly, these models determine the

likelihood that something is going to happen. There are other statistical patterns, such as latent variable models, where we determine if there is a latent or hidden trait that is responsible for the variations in observed variables. The point is that there are many models that we use to describe the variations in the observed variables, and how these variables relate to each other.

The implication is that if the general linear model does not seem like a good fit for describing the variations in your variables and the relationships between the variables, then there are other models that suit your purposes. The rat's ass part of this discussion is this – you want to get comfortable with a statistical model, in this case the general linear model, so you can organize your understanding of how results from statistical tests can help you understand the phenomenon you are trying to study.

You might think that these patterns are too simple. Psychology is much more complicated than the general linear model and its two patterns – as X increases, Y increases or decreases, or Group A is higher or lower than Group B. In one sense, you are right. Psychology is complex and intricate, and relationships are complicated. Advanced thinking, advanced models, advanced equations, advanced theories, and advanced discussions are needed to understand these relationships. The general linear model may not be well suited to understand these complex situations. The model does have limitations, such as a) if the data do not conform to a normal distribution, we cannot use statistics based on the general linear model, and b) if the patterns between variables do not follow a linear model, then the results obtained from statistical models do not apply. In these cases, we cannot use the linear model.

However, to learn this craft of statistics, we want to start at a simple place. We need to understand the basics of statistics and start with the general linear model before advancing. Advanced statistics do use statistics based on the general linear model. For example, the concept of correlation permeates the process of understanding a latent variable model, such as factor analysis. It is not like you will leave basic statistics behind once you venture into advanced statistics.

Before writing off basic statistics and the general linear model as too simplistic for the issues you want to investigate, consider Albert Einstein's statement: "If you can't explain it simply, you don't understand it well enough." The art of statistics and research is to understand complicated phenomena in simple terms. The K.I.S.S method applies here (keep it simple, stupid). Racism, the LGBTQ+ coming out process, PTSD, autism, treatment resistance, and mental health stigma in the Muslim religion—all are complicated issues that deserve simple explanations. Mind you, simple models might not fit these issues, but with simple analyses, we can begin the process of understanding these complicated issues and then move on to complex models.

Beware of people who try to sound, talk, and act in a complex and complicated way. Intelligence is not about going off on possibilities, or about spiraling and going down rabbit holes, by getting off target ("what if..."), or thinking about different contexts ("it depends on..."). Intelligence is not about criticizing ("your sample size is too small..."; "you did not include this variable..."; "your sample is not representative and generalizable"). Intelligence is not about intimidating others based on what you think you know. The sign of intelligence is to be simple. We want to have good bedside manners when discussing statistics and research. We want others to understand us. The art of intelligence is to be succinct. We want economical, elegant explanations for complex phenomena. We also don't want to make things too complicated for ourselves. Recovery in mental health is a complicated process. But in some ways, the recovery process is simple – pay attention to your mental health, value your mental health, build your physical and mental and emotional strength to do things that are in your best interest, and avoid the things you know are not in your best interests.

Keep it simple, in life, and in statistics.

Helpful Websites

<https://conjointly.com/kb/general-linear-model/>

<https://statstinking21.github.io/statstinking21-core-site/the-general-linear-model.html>

This page titled [10.3: General Linear Model](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

10.4: Discussion Questions

Discussion Questions

1. What does inferential mean?
2. What are the two main types of patterns we find in statistics?
3. What does the general linear model describe between variables?
4. Are there other models besides the general linear model, and why?

This page titled [10.4: Discussion Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

CHAPTER OVERVIEW

11: T-tests

Learning Objectives

By the end of this chapter, you will be able to:

- Know when to use a t-test
- Understand the design setup for a t-test
- Interpret a t-test result
- Read a journal article that uses t-tests

Key Terms

- Assumptions of a t-test
- T-test as a ratio

And here is your first inferential statistics test: the independent t-test. The purpose of the independent t-test is to test if the means of two groups are the same or different. You are trying to detect if one group is higher or lower than the other, or if the two groups are similar.

[11.1: When to Use the Independent T-test](#)

[11.2: Design Set Up](#)

[11.3: Commentary](#)

[11.4: T-test Assumptions](#)

[11.5: Reading T-tests](#)

[11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology \(EPPP\) Licensing Exam](#)

[11.7: Reading Articles That Involve a T-test](#)

[11.8: Reading T-tests from Journal Articles](#)

[11.9: Summary](#)

[11.10: Discussion Questions](#)

This page titled [11: T-tests](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

11.1: When to Use the Independent T-test

On the surface, we use the t-test to compare mean scores for two groups. The common application is determining if males and females have different mean scores. You could compare if clients who drop out of treatment versus clients who stay in treatment have different mean scores. You could compare if one classroom of students compared to another has different mean scores.

BTW: the t-test came from [studies about beer](#). Who says statisticians cannot party?

Always look for ways to visually depict statistics. We are visual learners. Visualization helps us to get our ideas presented rather than having to keep everything in our heads. Our heads are too cluttered to remember everything.

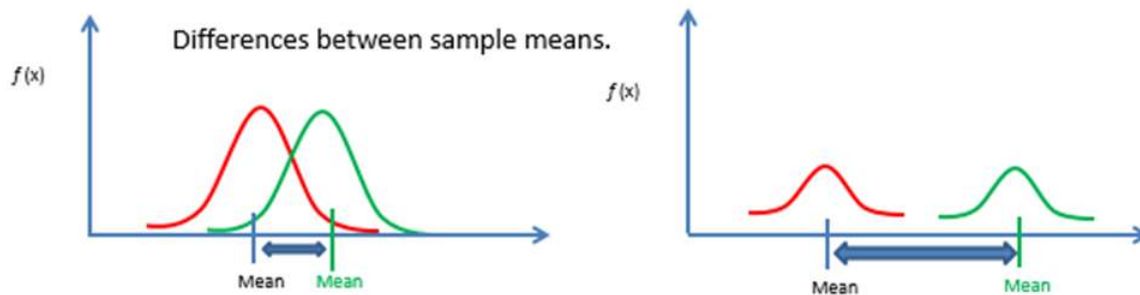


Figure 1: *Differences Between Sample Means.*

In Figure 1, the figure on the left depicts two groups. Their means are close together, and their distribution, or spread of scores, overlaps. You do not need me to tell you that these two groups are likely not that different. Because of the overlap, the two groups look like they could be from the same group.

The figure on the right depicts two groups, but their means are spread apart, and their distributions do not overlap. You do not need me to tell you that these look like two different groups. They look like they are different because there is no overlap, and it looks like they are clearly separate from each other.

The problem is that while we can describe how these groups are different heuristically, that description will not fly because we cannot use verbal descriptions to say groups are different. It is best that we have a number, or a statistical value, that indicates that these groups are different. That value for this situation is the t-test value.

This page titled [11.1: When to Use the Independent T-test](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

11.2: Design Set Up

Setting up a design consists of three elements: identifying your independent and dependent variables and formatting your hypothesis.

11.2.1: Independent Variable and Dependent Variable

The independent variable consists of two dichotomous groups. The independent variable is nominal or categorical, which are synonyms. There can only be two groups for that nominal/categorical variable. Examples are males v. females, Black versus White, treatment completers versus treatment dropouts.

The dependent variable is continuous. Recall that continuous variables are the ordinal, interval, and ratio variables. There must be a low to high range for these continuous variables.

Examples are student ratings of teacher effectiveness, level of income, number of sessions attended, ratings of helpfulness of therapy sessions, and number of alcoholic drinks consumed.

11.2.2: Hypotheses

Null hypothesis: The two groups have the same mean score.

Alternative hypothesis: one group has a higher/lower mean score than the other group.

And that's it!

This page titled [11.2: Design Set Up](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

11.3: Commentary

The commentary sections for each analysis are intended to highlight nuances or myths about each statistical test. For t-tests, most of the commentary is about how the t-test seems like a simple statistical test, when in fact, it is an essential test for simplifying how we organize our process of comparing groups.

11.3.1: “Looks Too Simple”

You would not be wrong if you thought, “This t-test looks simple, and kind of useless. All you can do is compare two groups and compare their means. What’s the big deal about this test? Where’s the complicated statistical analysis?”

Statistics sound and look complicated and advanced. However, we always strive to KISS or “keep it simple, silly.” The simplest way to compare anything is in pairs, or two at a time. In fact, that is the only way we can compare anything. We always compare in pairs. We compare treatment to controls, we compare males to females, we compare good kids to bad kids, we compare bullies to survivors.

The problem becomes when we have multiple groups. We compare treatment to control to usual care, we compare males to females to non-binary identified, we compare good kids to bad kids to somewhat bad kids, we compare bullies to survivors to bystanders. Comparing all of these groups at once is like trying to determine who wins “rock, paper, scissors” when there are three people playing. If all three throw down rock, paper, or scissors, you have no idea who won that round. The problem in that scenario is that you do not know where to start the comparison. Do you start with the rock, breaking scissors? What about the paper covering the rock first, which negates the rock breaking the scissors? And if you play Rock, Paper, Scissors, Lizard, Spock, you are in a world of hurt (<https://youtu.be/x5Q6-wMx-K8?si=-MDNe9-kc2IvFY9N>).

If you compare treatment to control to usual care all at once, you have a conundrum. Where do you start? Where is the comparison? The only way to make this comparison is to keep it simple, and that can only happen by comparing pairs. Treatment vs. control, control vs. usual care, usual care vs. Treatment. Comparing pairs and organizing the pairs so all combinations of pairs are accounted for is the process of making sense out of several groups at once. Comparing the pairs helps you rank the pairs. Treatment did better than control, control did better than usual care, and treatment did better than usual care. Arranging the comparisons allows you to see that treatment does work, treatment is better than the usual care clients receive, and the control group, that does nothing, will not improve.

In the next analysis on ANOVA, you will find that for any analysis where multiple groups are used, the follow-up tests involve a version of the t-test. When you compare multiple groups, or multiple independent variables, or multiple dependent variables, the method of making sense of the multiple comparisons is to keep it simple by breaking the comparisons into pairs. Making statistics complicated and advanced is not a good goal. Always simplify and keep your analyses in order and organized. Comparing pairs is really the only way to make sense of the complication.

And that’s why the t-test is an important tool in any statistical test. So don’t knock it.

11.3.2: Dichotomous Groups? That's Too Simple

You would not be wrong to say that there are many groups for a given variable. Race is more than two groups. Living in a binary world is not the greatest because it ignores continuous spectrums. Gender is not a binary term; there is more than just males and females. Sexuality is not a binary term; there is more than just heterosexual and gay. You are correct. The t-test functions as a test for two groups. For better and often for worse, we reduce and collapse complex and continuous experiences into a two-group, dichotomous format. You would not be wrong if you asked, “Is that okay?”

As you are starting to see, an ongoing theme throughout the instruction of statistics is that you must always think conceptually. In this case, the question is – is the dichotomous grouping sufficient for our research question and analysis? For example, you are analyzing if a new treatment for anxiety is better than the usual treatment for anxiety, but you are concerned that gender should be considered part of the analysis. For better or for worse, there are differences in how males and females experience anxiety. There are differences in how cisgender males, self-identified males, transgender males, and non-binary males experience anxiety. Does collapsing all variations of being male into Male vs. Female alter the analysis of a treatment for anxiety? Unless you have a theoretically compelling reason and enough previous studies with evidence, you do not have a strong basis for including every

version of the male spectrum identity into the analysis of the treatment effect. For the purpose of this example, collapsing different male identity spectrums into a single “male” category would be sufficient.

Splitting the male identity into several categories means you will be adding additional variables into your analysis. While comprehensive, you are using up your sample size for every variable you use. When you use up your sample size, you are using up your power, or your predictive power to detect an effect in your outcome variable. This issue will be addressed when we discuss degrees of freedom as part of the chapter on correlations. But for now, throwing all the groups and variables into an analysis, while apparently comprehensive, considerably reduces your power, or the likelihood that the statistical test will find an effect.

The bottom line is that you should not throw every possible group into an analysis unless you have a conceptually compelling reason to do so. By conceptually compelling, I mean there is a theoretical reason why all groups need to be included: they are important predictors of an outcome. If there is no reason, then keep it simple and use dichotomous groups.

11.3.3: Forming Dichotomous Groups

Forming dichotomous groups means collapsing multiple groups into two groups. Race has multiple groups. These groups can be combined into two groups. For example, you could take Black, Hispanic, Asian, and White and collapse them into two groups, such as White vs. Minority group. Here again, remember to keep it conceptual. If, for some reason, it makes sense to collapse into White vs. Minority, you are doing so because the comparison is meaningful. If you are interested in minority access to mental health compared to White populations, then the collapse is meaningful because there is something in common about minorities and their access to mental health services.

You can form groups by using a continuous variable and applying a cutoff score. For example, you can have two groups: clinical or subclinical groups based on the MMPI scores. You would be correct in saying, “Wait a second, aren’t these groups supposed to be dichotomous, categorical, meaning no level of low to high?” SPSS does not make any decisions about the variable being a categorical dichotomous variable. You can enter a categorical dichotomous variable, and the two groups can be low or high. This situation often occurs. We divide people into clinical or non-clinical depression. We divide people into “above the cutoff” and “below the cutoff” for admissions scores. We sort people into “good prognosis” or “bad prognosis” after discharge. We can reconfigure a continuous variable into a low-to-high group for any purpose. The t-test will not be rendered invalid.

Keep in mind that collapsing a continuous variable into two groups means losing precision information. Collapsing a continuous variable into two groups essentially means you are forming an ordinal, rank variable. The ordinal, rank variable is basically “high” vs “low,” and, if you recall the characteristics of a rank variable, it does not matter how high or low you are. You could be barely in the high group, or the top person in the high group. You are both considered in the high group.

When would this matter? Think of the reconfiguration of the continuous variable into an ordinal / rank variable as if having more of something really matters for your research question. For example, in determining alcoholism, it does matter if you drink, say, 10 beers, 12 beers or 25 beers in one night. Once you reach that threshold, the effects of alcohol are enough to determine alcoholism. However, if you are interested in alcohol poisoning, then yes, you do need to have a precise record of the number of drinks consumed. Although at that point, your blood alcohol level is a more precise record than relying on a person’s recall of exactly how many drinks they consumed to get blood alcohol poisoning. The point is that the configuration of your outcome variable as a continuous or an ordinal/rank variable depends on what information you need to answer your research question.

11.3.4: The Outcome Variable is Continuous

Technically, the outcome variable for a t-test is a continuous variable. This means the variable could be ordinal, interval, or ratio. However, interpreting the outcome variable is easier when the variable is an interval or ratio. Interval or ratio variables have equal intervals between the numbers. Two is greater than one. With equal intervals, you can say that 2.5 is greater than 1.5.

Using an ordinal variable as the outcome variable in a t-test makes little sense because we do not normally compare partial ranks. Getting a mean score for an ordinal variable makes little sense. If they are ranked first and the other second, the ranking is one and two. There is no such thing as a 1.5 rank. Ranks are whole numbers; they are not decimals. T-tests test the means of two groups. Saying that a group has a mean rank of 2.5 does not make much sense because there is no such thing as a 2.5 ranking. A person is either ranked two or three, not 2.5. That is like saying the average family has 2.3 children. You either have two or three children. You do not have two children, and 1/3 of a child (although I’m sure some families feel that way, unfortunately).

11.3.5: Why the Term “Independent” T-test?

Often, we say “t-test” and do not bother with saying “independent t-test.” In most cases, they mean the same thing. The word “independent” refers to having groups that are independent of each other, or where the members of each group are mutually exclusive, which means that a member of one group cannot be considered a member of the other group.

Another way of putting this is that the participants cannot be part of both groups. To state the obvious, a participant cannot be in both the male and female groups. Yes, assuming a binary gender world, a participant must be in one or the other group.

This situation may be obvious, but there are situations where it is not clear if someone is in one group and not in the other group. We like groups because they are clear, but most issues, diagnoses, and experiences are continuous, blended, and multi-dimensional. Assigning clients to alcohol abuse only, one drug abuse only, or poly drug use only, may seem clear and definitive, but it is likely in the messy context of alcohol and drug use, that clients clearly do not fit in just one of those groups, given the vicissitudes of their reported alcohol and drug use. Assigning clients to emotional abuse, physical abuse, and sexual abuse may seem clear in definition, but given the messy context of abuse, it is likely that clients experienced multiple types of abuse at different times during their lifetime. These are unfortunate situations, but depending on your conceptualization of your research question, it behooves you to consider whether your groups are independent or mutually exclusive. Consider your literature review, your theory, and consultation to be certain that you have independent groups.

Does the aforementioned discussion matter? Issues are messy and continuous, so no matter how we sort people into groups, there will be issues about the clarity of the assignment to groups. The t-test will run, and you will get a result. If the outcome shows no difference between the two groups, perhaps the messy assignment of people into the groups might account for the non-difference. If the outcome is what you wanted, and there is a difference between the two groups, perhaps the messy assignment did not matter, and there is a fundamental difference between the two groups.

Does it matter to say an “independent t-test” to test the means between two groups? Probably not. There are other t-tests, though, and the primary ones you want to be aware of are the paired t-test and the one-sample t-test, which will be discussed at the end of this chapter. It is necessary to say you are conducting a paired t-test or a one-sample t-test because those t-tests do serve specific purposes. However, for comparing two groups, the t-test is just fine; there is no need to add the extra “independent” word.

11.3.6: Uses of T-tests

On the surface, we use t-tests to compare the means of two groups. In which situations do we use this t-test?

We could use a t-test to compare a treatment and a control group, but that would be ill-advised. Making a definitive statement about treatment effects requires more than just comparing groups and one outcome variable. Suffice to say, you would need at least a pre-test and a post-test, and covariates, such as demographics, and a randomized controlled trial, at a minimum, to detect a treatment effect. Using only a t-test to test the effects of the treatment is ill-advised.

T-tests are used at the outset of studies where you want to determine if the demographics of samples are equivalent. In these situations, you do not want your t-test to be significant because you want to determine that the samples are similar. Hence, you are not looking for differences. You could use a t-test to determine if males and females have a similar mean age, similar income, and similar number of years of education. Doing so would allow you to determine that the males and females are similar in all demographic respects, except for being male or female. Making the males and females equivalent would help your overall goal. For example, if you wanted to determine if there are differences between males and females in terms of the average number of sessions of treatment they complete, you could first determine if they were equivalent demographically, such as in age, income, or years of education, so you could rule out the possibility that age, income, or years of education could have any effect on the number of sessions they complete. After establishing that the demographics are equivalent, you could rule out those demographics as possible confounds or alternative explanations and then proceed with comparing males and females on the number of sessions of treatment they complete.

A t-test is used for scale validation. If you want to see if males and females have different scores on a scale, a t-test is sufficient. If you want to see if in-patient and out-patient clients have different scores on a scale, a t-test is sufficient.

The best way to determine if you need a t-test is to consider your research question and your research design. A series of t-tests can be sufficient for your needs. Although the test seems simple, simple is good when detecting effects.

This page titled [11.3: Commentary](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

11.4: T-test Assumptions

All statistical tests have assumptions. Assumptions in the context of statistical tests are characteristics or requirements that need to be met for the test results to be valid or correct. There is only one result from a statistical test: significant or not significant. By valid or correct, we mean that the results are likely a good estimate of whether or not your finding is significant. If the assumptions are not met, then it is possible that your finding of significant or not significant might not be a good estimate. If you conducted a different test or did the test again with a different sample, you might get a different result. If the assumptions are not met, then it's best to get a consultation because there are considerations before you make a decision to proceed with your test.

11.4.1: Homogeneity of Variance

The first assumption is homogeneity of variance between the two groups. This assumption means that both groups have the same variance, or spread of scores, and the spread of scores is a normal distribution. The variances, or the normal distribution, of both groups should be the same.

Figure 1 presents what the homogeneity of variance between the two groups should look like. The one on the left is homogeneity. The one on the right is heterogeneity. Does this matter? The one on the left has two clearly different groups because their spread of scores does not overlap. The one on the right has two groups with different variances. The first group has a normal distribution, but the second group has a wider distribution than the first group. The problem is that you cannot tell if the second group has a mean score that makes them different from the first group, or if the second group had a wider variation, or spread of scores, that made them different from the first group. Does this situation matter? Groups are supposed to be mutually exclusive or fundamentally different from each other. If the spread of scores of one group overlaps with the other, the groups may not be so different from each other.

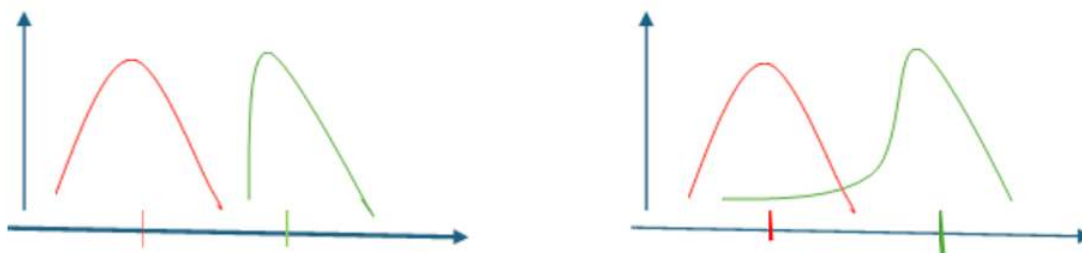


Figure 1: Homogeneity of Variance Between Two Groups

Check this assumption by using Levene's test for Equality of Variance. If this test is not significant, the variances are homogenous and equal. Therefore, report the t-test as is. That means you can trust the result. If this test IS significant, the variances are NOT homogenous, NOT equal. Therefore, DO NOT report the t-test as is. That means you CANNOT accept the results as accurate. Then you report the adjusted t-test as part of the t-test results.

Double check – if both the as is t-test and the adjusted t-test give the same result, meaning that both t-tests say that the result is significant or not significant, report the “s is t-test because it's simpler. But, if the as is t-test and the adjusted t-test give DIFFERENT results, meaning that one t-test says that the result is significant or not significant, BUT the other t-test gives the OPPOSITE result, then CALL ME. Because something is going on.

11.4.2: Robust Test

It is important to know that the t-test is considered a robust test. Robust means you will get the same answer, whether the test is significant or not significant, even if the assumption is violated. If the homogeneity of variance assumption is not met, the t-test will likely give you the same result if the assumption were met. The preceding discussion matters if, at the end of the day, the t-test will give us an answer no matter what. The preceding discussion does matter because there might be something else going on besides the homogeneity of variance. Think of this assumption as a red flag, but it is not the only red flag that will make or break your statistical results.

11.4.3: Sample Size

Based on guidelines for a normal distribution, having 30 participants per group is ideal for obtaining a stable result. Simulation studies determined that when a sample size is $n > 20$ for each group, and if both groups are of an equal size, the t-test should produce robust, stable statistical results.

Of note is that the sample sizes for both groups should be equal. How equal is equal? If group A has 40, and group B has 50, that is equal enough. The problem arises when one group, particularly the group of interest, has a sample size that is below 20 and is not equal to the other group. If one group is male elementary grade boys with behavioral problems, and the second group is elementary grade females with behavioral problems, it is quite likely that the boys will have an n of 30, while the girls will have an n of 11. This imbalance likely occurs because, on average, there are more young boys with behavioral problems than young girls. Is this a problem? From a population standpoint, this imbalance is expected. From a statistical standpoint, it can be a problem because it is hard to establish a mean score for a group of just 11. What to do? You could recruit more participants, in this case, more young girls with behavioral problems. Or you may have to scrap the t-test and use another statistical test. If so, consult with a statistician.

This page titled [11.4: T-test Assumptions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

11.5: Reading T-tests

Reading statistical analyses in journal articles follows the same format for all statistical tests. Review your research question and hypotheses, identify your independent and dependent variables, and read the statistical results by following the same structure as stated in the next section.

11.5.1: Determining Significance

When you read a t-test in a journal article or read a t-test in your SPSS statistical printout, you will get something that looks like this: $t = 3.56$, $p = .03$. Read this output as, “the t-test value is 3.56, and it is significant at the $p < .05$, which means there is a mean difference between the two groups”.

Here are the steps:

A t-test is significant if the t-test value is greater than the threshold value, and the p value is less than your alpha level. Usually, the t-test value threshold is 1.96, $N \geq 30$ per group, and p is set at .05.

Here are examples:

T-test = 1.03, $p = .38$. The test is not significant. The t-test is below 1.96, p greater than .05, the test is not significant. When it is not significant, the mean difference between the two groups is not large enough to be significant. The two groups are similar; nothing is going on here.

T-test = 4.39, $p < .01$. The test is significant. The t-test is above 1.96, p less than .05, the test is significant. When it is significant, you can claim that the mean difference between the two groups is large enough to be significant. Something is going on. There is something about the two groups that is associated with the difference in the mean scores.

Note that the sign of the t-test is not important. You will read t-tests that say 4.39 and -4.39. The sign is just directional. If the t-test is positive, it means that the first group is higher than the second group. If the t-test is negative, it means that the first group is lower than the second group. There is no inherent value regarding the sign of the t-test. It is just a descriptive tag that signals the first group is higher or lower than the other group.

11.5.2: Statistics as Ratios – T-tests

Keep in mind that statistical tests are ratios. The ratios are the true variance/error variance. The value you obtain in a statistical test is the outcome of a ratio that compares the true variance to the error variance. If you are looking for the mean difference between two groups, what you want to see is the truth, or that the groups are different from each other, and that truth is bigger than the error, or that the groups are not truly different from each other.

Back to basic math. You get a large number when you divide a large number by a small one. You get a small number when you divide a small number by a large one. When you divide a number by the same number, you get one.

The value of the t-test, $t = 4.39$, is the outcome of true variance/error variance. If the value of the t-test is positive, that means the true variance is greater than the error variance. But we do not want the true variance to be just barely greater than the error variance. If the t-test is 1.96, almost 2.00, then the true variance is twice as large as the error variance. That is good because it clearly states that the true variance is larger than the error variance. That is why statistical tests are basically ratios. Recall that the hallmark of statistics is comparison. We are always comparing something to something. In this case, we are comparing two groups; we want to see if the difference, the truth, between the two groups, is larger than the error, or lack of difference, between the two groups.

When interpreting t-tests, as with most statistical tests, as the value of the statistical test goes up, the p value goes down. That is why sometimes you read only the p value in reports, because the p value is all you need. You do not get any more information if you have the t-value accompanying the p value. Actually, it is better to just read the p value and ignore the t-value. The t-test value is affected by several factors, including the overall sample size, the sample size in each group, the difference in the sample sizes between the two groups, and the distribution of the scores for each group. Determining the true variance and the error variance is not a straightforward way to calculate it. While we fuss over the t-test value, you are better off just interpreting the p value. Recall that the p value only indicates significance, and in determining significance, we are interested in just one question – is it significant

or not? The only answers to that question are yes or no. There is no such thing as more or less significant. The t -test value itself does not really add any information about the significance of the test beyond the p value.

Of note, given what we previously learned about not placing too much stock in the p value, you might wonder why you read about t -tests with different levels of significance, such as $p < .05$, $p < .01$, or $p < .001$. The $p < .05$ and $p < .01$ could be expected if researchers adjusted for Type I or II errors. However, the researchers should report that they are using two different p values for that purpose. There is absolutely no value in having a t -test be significant at the $p < .001$ level and stating that the t -test is highly significant. There is no such thing. Do not pay attention to these variations when you see t -tests reported as having p values of $.05$ and $.01$. They all mean the same thing: something is significant or something is not significant. Nothing more.

11.5.3: Patterns in T-test Results

It is rare for a t -test to be conducted as a single test without it being used as part of a series of t -tests. You might examine an age difference between males and females, but that single t -test is part of an overall battery of statistical tests that examine differences between males and females for several variables. Usually, we conduct a series of t -tests, and when we do so, we look for consistency in the t -test results. Is one group consistently higher than the other across t -tests?

This situation commonly occurs when comparing two groups across a series of variables because you are interested in whether the two groups are similar or different. Suppose you are examining military veterans with PTSD and comparing them with first responders with PTSD. You conduct a series of t -tests with one group of military veterans with PTSD and the other group of first responders with PTSD. Military veterans are similar to first responders in outcome variables such as depression, anxiety, and a decrease in social support. But the military veterans have higher scores, indicating more severity, on scales for PTSD, anger management, and alcoholism, compared to first responders. In this case, this pattern of t -test results might make sense. While both groups have PTSD and will demonstrate depression, anxiety, and decrease in social support, perhaps the military experiences associated with PTSD might yield differences that are relevant to the military, such as anger. Because of the culture of alcohol use among the military, perhaps they will engage in more alcohol use compared to first responders. Here, the series of t -tests yields important information because of the pattern among the collective t -test results, rather than just one t -test result at a time.

This page titled [11.5: Reading T-tests](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology (EPPP) Licensing Exam

- Student t-test for a single sample
 - One IV: single group, compared to a normative population mean, DV is a continuous variable
 - $Df = n-1$, n is the sample size
- Student t-test for independent samples
 - One IV: two independent groups, DV is a continuous variable
 - $Df = n-2$, n is the sample size
- Student t-test for correlated samples (aka “paired t-test”)
 - One IV: one group, two time points (the groups are “correlated” with each other), DV is continuous
 - $Df = n-1$, n is the sample size (only following one group over time, hence it's $n-1$, NOT $n-2$).
- Point Biserial correlation
 - One IV; true dichotomy (gender male/female). DV is continuous
 - Note that a true dichotomy is one where the two groups have naturally formed. So, in our binary world, there are only two genders, male and female. There are only two groups of handedness, left and right.
- Biserial correlation
 - One IV; artificial dichotomy (clinical vs. non-clinical). DV is continuous.
 - In contrast to the point biserial correlation, the biserial correlation has an artificial dichotomy. An artificial dichotomy is one where you, the researcher, set the groups. So, for clinical vs. non-clinical, you set the cutoff score that separates clients into clinical vs. not clinical. For accelerated vs. not accelerated students, you set the cutoff score for determining which students should, or should not be, in the accelerated learning track. You set the cutoff score for hired vs. not hired to determine who is hired and who is not. The point is that you set the point where there are two groups, as opposed to the point of biserial correlation, where the groups are already formed for you by natural means or other sorting criteria.
 - Note that nothing bad will happen if you conduct a point biserial correlation instead of a biserial correlation. You likely would obtain the same answer if you conducted a point biserial correlation and a biserial correlation.

This page titled [11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology \(EPPP\) Licensing Exam](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

11.7: Reading Articles That Involve a T-test

We turn to how to read articles that use a t-test. The first few points are common for any statistical test.

11.7.1: First – Identify your Research Question

What's the goal of the study? What's the study trying to contribute to the field? The research question is not always straightforward. The research question is about what the researchers are curious about. Identifying and understanding the research question is about understanding the objective of the study. Keep in mind that the research question should be grounded in theoretical considerations. There should be an explanation for why the authors wanted to examine the two groups.

It is not a good idea to say the rationale for conducting the study is to develop better treatment plans, or because there is a gap in the literature. All clinical explorations of an issue will have implications for treatment planning. The rationale is weak because there is nothing unique about developing new treatment plans. A gap in the literature is something that has not yet been studied. However, there are always gaps in the literature because not everything has been studied. Studying something because there is a gap in the literature is the antithesis of the "Mt. Everest" argument. You don't study something simply because it was not there.

You need the conceptualization part of the research question to understand why the results turned out as they did or did not.

11.7.2: Second - What are the Hypotheses?

You want a tally of the hypotheses to organize your understanding of the statistics. You should identify the two groups in question and the outcome variables. Usually, there are several within a given study. Keep your hypothesis in this framework: "Group A is higher or lower than Group B." Keeping your hypothesis consistent helps with organization.

11.7.3: Third - Logistics

The logistics include the details of how you interpret the numbers. First, review the scoring protocol for each measure in the method section. What do high and low scores mean? Do high or low scores mean something good or something bad? Are there any thresholds or critical values you need to be aware of? These notes will help you interpret what the means indicate. Usually, the mean scores, when compared to the range, only indicate one of three outcomes: the scores mean that the group has a low, medium, or high amount of the variable in question.

What are the two groups being compared? Any notes about how they were formed? What is the overall sample size, and what is the sample size for each group? Any oddities about the sample sizes or demographic characteristics of each group? These issues could be considered confounds when evaluating the validity of the results.

11.7.4: Fourth - Review the T-test Results

Report the means and the t-test results as significant or not significant. Look for patterns in the effects. Do the findings confirm the hypotheses?

When reviewing the results, follow these steps :

1. Read the t-test and the p value. Does the p value indicate significance? The only answers are yes and no.
2. If no then your statement is: here is no mean difference between the two groups. And nothing more needs to be stated. Remember, do not state, almost significant or approaching significance. There are no such statements.
3. If yes, then your statement is: there is a mean difference between the two groups. And nothing more needs to be stated. Remember, do not state, somewhat significant or highly significant. There are no such statements.
4. If yes, then state the pattern. For t-tests, there are only two patterns. Group A is higher than group B, or group A is lower than group B. Keep your groups consistent according to each hypothesis.
5. If yes, then determine if the pattern you observed is expected or unexpected. By expected, we mean the pattern confirms the hypothesis. By unexpected, we mean the pattern was a surprise; the hypothesis did not predict that pattern to happen. Was it expected that Group A would be higher or lower than Group B?

Then, examine the effect size. Most of the time, for t-tests, they are not reported. However, a crude way of estimating an effect size is to take the standard deviations of both groups, average them, and then divide the actual difference in the means by the combined,

or pooled, average standard deviation.

There are only three answers to the question, What is the effect size? The answers are low, medium, and high. We usually do not see high-effect sizes. We usually see low or medium effect sizes. Are the effect sizes expected or unexpected? Do we want to see that much difference between the two groups?

If there are multiple t-tests and multiple tests, pool the results together and evaluate if the results are consistent and aligned with the research question. Do the results say something about the phenomenon under investigation?

Please note the following when reviewing statistics in articles.

There is no need for criticism of the method or the overall study. Stay with the numbers. Criticism of the study only distracts you from interpreting the numbers. Every study has its critics. Yes, some studies are worse than others. Until you decide the validity of the study is in question, put aside the criticism and stick with interpreting the numbers and the statistical results.

There is no need to act, think, or state things in a sophisticated manner. Keep it simple. Sounding academic or scholarly makes people confused. What this means is “if you hear hooves, it’s a horse, not a zebra.” This mantra is from medicine. This mantra is meant to communicate to doctors that you are likely not looking at an unusual “one-in-a-million,” case. The case is simple. You have a common cold, nothing more, no rare infectious disease. The same is true for statistics. In interpreting the statistics, it’s a horse. You will likely not encounter any strange anomaly, advanced interpretation, or anything hidden in the numbers.

Read everything straightforwardly. There are no magical, hidden, or steps meant to trip you up in reading statistics. “No rabbit hole spiral thinking” – This is my mantra. Do not engage in “what ifs” such as looking at demographics and trying to say, “Well, what if there were more of this, or more of that?” Or trying to add extra variables or contexts to explain the results. This kind of thinking can distract and lead you down the rabbit hole of Alice in Wonderland and take you away from the variables in question. Stay in the lane of the variables included in the statistical test.

So, now that I’ve told you what NOT to do when interpreting statistical tests, what should you do?

11.7.5: Conclusion

Always think CONCEPTUALLY, which is your explanation of why we are looking at these variables in the first place. Conceptually means how these variables are connected. What you want to do is think about why these variables would be connected, or what these variables represent.

This page titled [11.7: Reading Articles That Involve a T-test](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

11.8: Reading T-tests from Journal Articles

Let us refer to this example from a study done by Conrad, S., Queenan, R., Brown, L., & Tolou-Shams, M. (2017). The research question for this study was: Are there gender differences in psychiatric and risk behaviors for non-incarcerated youth? The research question was based on the disturbing trend of more young girls entering the juvenile system. The researchers base this trend on increased ease of access to marijuana and (unfortunately) sexual-risk behaviors. The researchers wanted to know if the young girls entering the system were similar to or different from boys, who, unfortunately, are more likely than girls to be in juvenile systems. The researchers wanted to know if the girls were similar to boys in terms of their risk behaviors, or if girls were engaging in risk behaviors specific to their gender.

Their hypothesis was: are girls similar, lower, or higher levels of at-risk behavior or psychiatric symptoms than boys?

Table 1 presents their t-test findings. Try to look for t-tests displayed in table format whenever possible. They are much easier to read in table form. Note in Table 1 that the t-tests are combined with chi-square tests. This combination is typical. T-tests only examine two groups, and the dependent variable is continuous. The chi-square tests examine groups, and the dependent variable is categorical. To be economical in presentation, researchers combined the results of the t-test and chi-square tests.

Table 1: Descriptive Statistics: Mean, Standard Deviation, Percentage, Chi-Square, and T-test By Gender

	Total (N=60) Mean (SD) or % (N)	Male (N=42) Mean (SD) or % (N)	Female (N=18) Mean (SD) or % ()	t or χ^2
Demographics				
Age	15.54 (1.30)	15.62 (1.34)	15.40 (1.24)	.53
Race (Caucasian)	73% (44)	69% (29)	80% (13)	.61
Ethnicity (non-Latino)	79% (47)	83% (34)	80% (13)	.56
Mental Health				
YSR Total Symptoms (t -score)	55.69 (10.11)	53.95 (10.39)	60.29 (7.93)	-2.06
YSR Internalizing Symptoms (t -score)	50.29 (8.82)	48.70 (8.92)	54.50 (8.65)	-2.14
YSR Externalizing Symptoms (t -score)	60.69 (10.88)	59.62 (11.73)	63.50 (7.80)	-1.14
Childhood Trauma Questionnaire				
Total Trauma Score	48.93 (9.44)	43.70 (9.35)	49.13 (9.53)	-1.83
Sexual Abuse	6.12 (2.58)	5 (0)	7.9 (5.59)	-3.28
Emotional Abuse	8.35 (3.51)	7.71 (3.51)	8.80 (3.51)	-1.08
Physical Abuse	7.33 (2.69)	6.13 (3.60)	6.26 (1.79)	.19
Adolescent Risk Behavior				
Substance Use				
Ever Use Marijuana (yes)	88% (53)	87% (36)	93% (17)	.13
Used Marijuana Past 30 Days (yes)	51% (31)	54% (22)	53% (9)	.01
Ever Smoked Cigarettes (yes)	70% (42)	67% (28)	86% (14)	2.16
Cigarettes Past 30 Days	59% (35)	54% (22)	80% (13)	1.64
Ever Used Alcohol (yes)	75% (46)	75% (31)	86% (15)	.95
Used Alcohol Past 30 Days (yes)	31% (19)	26% (10)	53% (9)	.19
Ever Used Club Drugs	12% (7)	10% (4)	20% (3)	1.34

	Total (N=60) Mean (SD) or % (N)	Male (N=42) Mean (SD) or % (N)	Female (N=18) Mean (SD) or % ()	t or χ^2
Ever Used Cocaine	3% (2)	3% (1)	6% (1)	.59
Ever Used Heroin	0% (0)	0% (0)	0% (0)	-
Every Used OTC/Rx Medications	20% (12)	16% (6)	33% (6)	2.61
Sexual Behavior				
Lifetime Sexual Behavior (yes)	71% (43)	69% (27)	90% (16)	5.85
Condom Use Last Sex (yes)	50% (22)	54% (14)	46% (8)	2.50
Age 1st Sexual Intercourse	14.40 (.95)	14.60 (.88)	13.30 (1.03)	2.97
# of Partners Past 90 Days	1.79 (1.35)	2.35 (2.11)	1 (0.60)	2.14
Been/Gotten Pregnant (yes)	7% (4)	5% (2)	13% (2)	.64
Substance Use and Sexual Behavior				
Any Drug Use Last Sex	36% (15)	34% (9)	40% (6)	.01
Partner Used Any Drug Last Sex	36% (16)	31% (8)	50% (8)	1.80
Alcohol Used Last Sex	5% (3)	5% (2)	6% (1)	.01
Partner Used Alcohol Last Sex	7% (4)	5% (2)	13% (2)	.62
Marijuana Used Last Sex	14% (6)	15% (4)	13% (2)	.25
Partner Used Marijuana Last Sex	16% (8)	12% (3)	28% (5)	1.75

* $p \leq .05$, ** $p \leq .01$.

To read statistics tables, find your starting point. In this case, as with most tables, start in the upper left-hand corner. Identify your independent variables and dependent variables. The first dependent variable is age; the independent variable is gender, and the two levels are male and female, in the table column headers. Look for the means. For age, the mean age for males is 15.62, and for females is 15.40. Then find your t-value, which is to the right. It is .53. Then find the accompanying p value, which is .60. So, the t-value is not significant, indicating the males and females are similar in age. This finding is expected. We are examining youth, and we are unlikely to find a difference in age. Does this finding matter? It is not relevant to the research question, but it does demonstrate that males and females are equal in age. As a possible confound, this finding is good because if age is the same in both groups, then we can be certain that differences between the males and females are due to gender differences rather than an age difference. In this case, we want to find non-significant results to establish equivalency between the two groups, in this case, boys and girls.

Let us examine the t-test for the mental health variables. The YSR Total Symptoms variable has a mean of 53.95 for boys and 60.29 for females. The accompanying t-value to the right is -2.06, and its p value is .04. The negative t-value simply means the boys have a lower mean score than the girls. The t-test is significant, and the pattern is that girls have higher total mental health symptoms than boys. Of note, yes, the t-test is negative, so technically the boys have a lower total score than girls, but it is just intuitively easier to say that girls have higher total mental health symptoms than boys. This finding is expected according to the researchers' hypotheses, which is that girls might have more mental health symptoms than boys.

To interpret this finding, we need to know what these scores mean. Although their method section does not state the scoring protocol, the table does indicate that the scores are t-scores. Recall that t-scores have a mean of 50 and an SD of 11. The mean for boys is 53.95, which is right above the mean; the girls have a mean of 60.29, which means they are a standard deviation above the mean. By

itself, it seems boys have average amounts of symptoms, while girls have a level of symptoms that might border on the sub-clinical range. A good idea would be to review material about the YSR to understand these t-scores, but for now, our guess is as good as any.

Let us examine the effect size by examining the standard deviations for both boys and girls. Recall that effect sizes are based on standard deviations. Do this rough estimate of an effect size because no effect size is reported in the table. The mean difference between boys and girls is about 6 (60.29 – 53.95). The standard deviation for boys is 11.39, while for girls it is 7.93. A rough pooled standard deviation is about nine $((11.39 + 7.93) / 2 = 9.43)$. A mean difference of 6 divided by 9 is roughly .67 or 2/3 of a standard deviation. This value is a high effect size according to our rough effect size value guidelines (.3 = small, .5 = medium, .7 = high effect sizes). This difference between boys and girls is basically a high effect size, which means this effect is noticeable.

Putting it all together, it does seem that girls have higher mental health issues than boys in the context of pre-juvenile delinquency populations. This finding is conceptually expected because mental health issues among girls are (unfortunately) an ongoing trend. It makes sense that girls with mental health problems are probably at risk of experiencing the juvenile delinquency system.

Remember, we want to find patterns among the t-test results. We examine the accompanying t-tests for the other two YSR Mental Health Scales. In brief, the t-test for the Internalizing Symptoms scale is -2.14, $p = .03$, and significant; the pattern is that girls have higher Internalizing Symptoms than boys. The t-test for the Externalizing Symptoms scale is -1.14, $p = .26$, and not significant, and the pattern is that girls and boys have similar levels of Externalizing Symptoms.

The t-test result for the Internalizing Symptoms scale is expected because girls tend to have internalizing symptoms, such as depression or anxiety. The overall high YSR score for the girls could be the result of girls having a higher level of depression or anxiety than most girls, and boys. This result makes sense. Interestingly, the Externalizing Symptoms scale did not have a significant result for girls and boys. The girls have almost the same level of high externalizing symptoms as the boys. Both boys and girls have scores of 59.62 and 63.50, respectively. The t-test is not significant, but this seems to be an interesting finding. Generally, girls have lower levels of externalizing behaviors, such as aggressiveness, compared to boys. But if girls have similar levels of externalizing behaviors as boys, perhaps this combination of high internalizing symptoms, plus the high levels of externalizing behaviors at similar levels as boys, is a bad combination for girls to end up in the juvenile system. In concert with each other, the t-test results of the internalizing symptoms and the externalizing symptoms say something about the profile of girls who are at risk of entering the juvenile delinquency system. The review of these t-test results together is the process of using statistics to discover something new about an issue.

Are there any concerns about these statistics? Well, the sample size does seem like it should be noted. The overall sample size is 60, and the males have 42, but the female sample size looks low at 18. The concern is the robustness of the results. Is it possible that the results could change with more females? It is always possible. Let us use a conceptual approach. Finding and recruiting girls with potentially juvenile delinquent behavior is not easy. The method section states that the population is court-involved, non-incarcerated boys and girls. Generally, more boys than girls are in this population. Remember that youth need parental consent to participate in studies, so obtaining parental consent is a difficult process. Then, the youth must agree to participate and complete all the measures. Those requests are a lot for the youth in this context. In some ways, the researchers should be commended for recruiting 18 girls from this context. Should they recruit more? Of course, the answer is yes. Can they? Well, recruitment within a court system, which is not easy to get permission to do, takes a lot of resources and time. You have to ask yourself, could I have recruited that many participants with my own resources and time? Quite likely, the answer is no, so you have to back off the researchers and not impose a “you should have....” statement on the researchers.

The sample size is small. Small sample size means that a study is underpowered, and underpowered studies may be an issue for a Type II error. Recall that power is the ability to detect an effect if, in fact, the effect does exist. One aspect of an underpowered study is a low sample size. Low sample sizes do not have enough opportunity to detect the effect you want to see. A Type II error is saying a result is not significant when, in fact, it is significant. Is this scenario occurring in this study?

Well, the t-test result for the YSR Externalizing Symptoms is -1.14, $p = .26$. Reporting the p value here is helpful. The result is clearly above the .05 threshold, so it is doubtful that shifts in the p value to below .05 will occur. The difference between the two groups is 4, boys = 59.62, girls = 63.50, and this difference seems small, and the actual scores just hover around the 60 t-test score, which means both groups are at the same level of externalizing symptoms pathology. The t-test result says that boys and girls are not different in Externalizing Symptoms. Although the sample size is low, suggesting an underpowered study and a possible Type II error, several factors suggest otherwise. The actual p value, the small difference between the groups, and the similarity in their t-test scores point in the same direction. Conceptually, the non-significant result also makes sense. Therefore, it is possible that a Type II error did not occur, and the sample size has minimal impact on this result.

How about the YSR Total Symptoms score? The t-test is -2.06, and the p value is .04, which is right below the .05 threshold. This situation is a possible Type I error. The Type I error is saying you have a significant result when, in fact, you do not have a significant result. Given the low sample size for the girls, is it possible that with a larger sample size, the mean score might shift, specifically shift lower, resulting in a smaller mean difference, rendering the t-test result as not significant?

That is quite the scenario, and unlikely. Recall the population. These are girls in court but not incarcerated. They likely experience mental health symptoms, and it is equally unlikely that they would not be experiencing some level of depression, anxiety or other internalizing symptoms. While the statistical situation might be ripe for a Type I error, there is likely no Type I error. Remember, one statistical test does not make or break a study and recall that our approach is not to criticize or look for anomalies, exceptions, or problems. In concert with the rest of the t-test results, it does appear that the difference between girls and boys on the YSR Total Symptom variable is a valid result. Yes, another study replicating these results is desirable. Until then, we keep our original conclusions and explanations.

And that is how we read statistics, in this case, t-test results, in a journal article.

This page titled [11.8: Reading T-tests from Journal Articles](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

11.9: Summary

T-tests examine the mean differences between two groups. Although this statistical test may seem simple for most purposes, when used as a series of tests, we use t-tests for making comparisons; they are a ubiquitous part of statistical analysis.

References

Conrad, S. M., Queenan, R., Brown, L. K., & Tolou-Shams, M. (2017). Psychiatric symptoms, substance use, trauma, and sexual risk: A brief report of gender differences in marijuana-using juvenile offenders. *Journal of Child & Adolescent Substance Abuse*, 26(6), 433-436. <https://doi.org.11.1080/1067828X.2017.1322017>

Helpful Websites

<https://www.jmp.com/en/statistics-kn...-portal/t-test>

<https://www.bmj.com/about-bmj/resour...-one/7-t-tests>

This page titled [11.9: Summary](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

11.10: Discussion Questions

Discussion Questions

1. Describe the design setup for a t-test
2. What are the patterns we find for a t-test?
3. What are the assumptions of a t-test?
4. What are the steps for reading journal articles that use t-tests?

This page titled [11.10: Discussion Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

CHAPTER OVERVIEW

12: Analysis of Variance – ANOVA

Learning Objectives

By the end of this chapter, you will be able to:

- Know when to use an ANOVA
- Understand the design setup for an ANOVA
- Know when to use an ANOVA instead of multiple t-tests
- Interpret an ANOVA result
- Read a journal article that uses ANOVAs

Key Terms

- Assumptions of an ANOVA
- ANOVA as a ratio
- Post-hoc tests
- Experimenter-wise error rate

On to the ANOVA, or Analysis of Variance. Analysis of Variance is a term that describes a family of analyses. Statistics is about analyzing and understanding variance; suffice to say, it is probably accurate to say that Analysis of Variance is what statistics is all about.

To begin, the purpose of the ANOVA is to analyze the variance of three or more groups on a single dependent variable. You are trying to detect if one group is higher or lower than the other two groups, or if the three or more groups are similar? The ANOVA contrasts with the independent t-test, which examines two groups only on a single dependent variable. Keep in mind that the two groups versus one group is not the best way to describe the difference between the ANOVA and the independent t-test. A better way to describe the difference is that the ANOVA examines variance, while the t-test examines mean scores. Because the ANOVA examines variance, it is a flexible statistical test that can be used for additional purposes, which will be described at the end of this chapter.

[12.1: When to Use the ANOVA](#)

[12.2: The Concept Behind Analysis of Variance](#)

[12.3: - Interpreting the F-test and Post-Hoc Comparisons](#)

[12.4: Commentary](#)

[12.5: Reading Articles – ANOVA](#)

[12.6: Summary](#)

[12.7: Discussion Questions](#)

This page titled [12: Analysis of Variance – ANOVA](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

12.1: When to Use the ANOVA

The Analysis of Variance is an umbrella term for a host of analyses. The first use of ANOVA is to compare the means of three or more groups. The common application is to determine if race is associated with mean scores on a dependent variable. We compare White, Black, Hispanic, and Asian populations on a dependent variable, such as depression.

Note that using ANOVA to compare three or more groups is only one way we use ANOVA. ANOVA has many uses, but for the purpose of simplicity, we will start with using ANOVA to compare three or more groups.

12.1.1: Hypotheses

Null hypothesis: The three or more groups have the same mean score

Alternative hypothesis: one group has a higher/lower mean score than the other two or more groups.

12.1.2: Design Set Up

The independent variable consists of three or more groups. The independent variable is a nominal or categorical variable, which is the same term. There can be three or more groups for that nominal/categorical variable. Examples are for the variable race, which we use: White, Black, Hispanic, and Asian. For the variable treatment, we use the new treatment, control, and usual care treatment. For the variable first responders, we use firefighters, paramedics, and police officers.

You can use a continuous variable and divide it into three or more groups. You could take scores on a depression scale and divide them into non-clinical depression, sub-clinical depression, and clinical depression. You could take scores on an alcoholism scale and divide them into social drinking, binge drinking, and heavy drinking. Essentially, you are creating ordinal variables, which, as you recall, are rankings. You could ask yourself if there is a better way to analyze ordinal ranking variables, because dividing the continuous variable into three ranked groups does not seem to fit the ANOVA framework. You would be right, and yes, you could use a rank Spearman correlation. In general, using the F-test for this situation would not dramatically change the results if you used the Spearman rank correlation. Using both tests will give you unique information, but they likely will give you the same result: whether the relationship between the IV and DV is significant or not. For now, using ANOVA to analyze an independent variable that is scaled as an ordinal variable would be sufficient.

The dependent variable is continuous. Recall that continuous variables are the ordinal, interval, and ratio variables. There must be a low to high level for these continuous variables. Similar to the t-test, examples are student ratings of teacher effectiveness, level of income, number of sessions attended, ratings of helpfulness of therapy sessions, and number of alcoholic drinks consumed.

Note that you can use an ordinal variable as your outcome variable, but interpreting the ordinal variable is difficult. It is difficult to use an ANOVA to state that there are changes in ordinal ranking when ordinal ranks are based on an all-or-nothing premise. You are in first place, or you are not in first place. There is no such interpretation where you can say that you have a higher average first-place ranking compared to the second-place ranking. It does not make a lot of intuitive sense. You are better off using logistic regression.

This page titled [12.1: When to Use the ANOVA](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

12.2: The Concept Behind Analysis of Variance

Analysis of Variance is based on the partitioning of the variance. Partitioning means dividing, and as we have learned, we always partition variance into two types, true and error variance. This section examines how this process works for an ANOVA.

12.2.1: How Does it Look?

Figure 1 depicts an ANOVA.

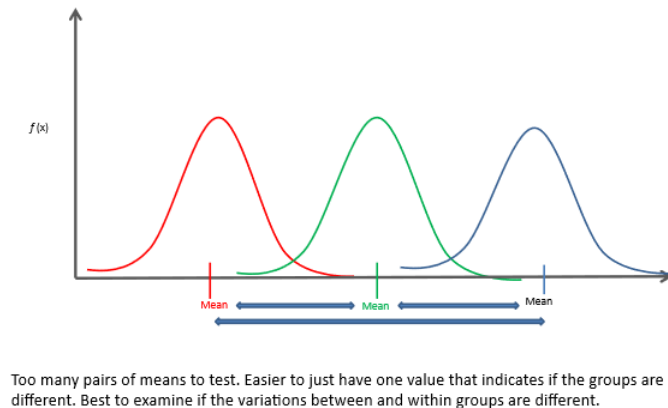


Figure 1: Graph Depicting an ANOVA

12.2.2: Why Call it Analysis of Variance?

With these three groups, we could run three t-tests. One t-test between the red and green groups, the green and blue groups, and the red and blue groups. But that is too many t-tests. It is simpler to analyze the groups all at once and get one number to determine if there are differences among the three groups. To analyze this situation, we will further discuss the pros and cons of running three t-tests versus one overall statistical test, in this case, the ANOVA.

Recall how statistics are basically comparing something to something else. In statistics, we compare the observations we want to see and determine if the pattern of observations we see differs from error or random observations. In statistics, when we compare, we use ratios. In ANOVA, we compare variances, specifically between-group variances with within-group variances.

- Within groups variation: variation of scores around each group's mean (sometimes called error).
- Between groups variation: variation among the group means.
- You want: between groups variation > within groups variation.

To illustrate, in Figure 2, the top figure represents a situation where the group means are not different from each other. You do not need me to tell you the groups look similar. Their means are close together, and their score distribution spreads are overlapping, to the point where these groups look like just one big group. The bottom figure represents a situation where the group means are different from each other. You do not need me to tell you the groups do not look similar. Their means are far apart, and their score distribution spreads do not overlap to the point where these look like three separate groups.

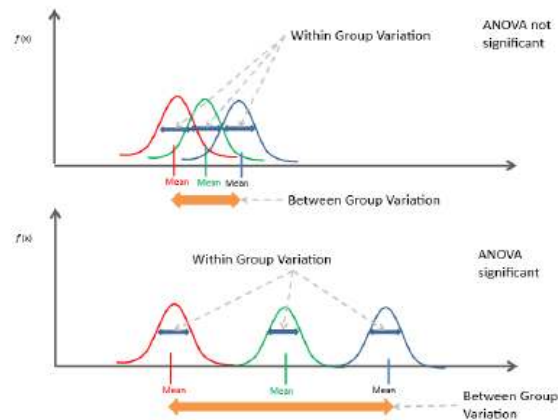


Figure 2: Difference in Between-Group and Within-Group Variation

As you know, we cannot rely on narrative to tell others that we have three separate groups. We need a number to represent the observation that we have three groups overlapping with each other to the point where they look like one big group, or that we have three separate groups. That number is the F-test. The “F” stands for Fischer, who devised the formula for the ANOVA analysis (interesting how statisticians like to name statistics formulas after themselves). We say we are conducting an ANOVA analysis; the F-test is the statistical test that produces the statistical value as a result of the ANOVA analysis. Basically, when someone states or writes about the F-test, they conduct an ANOVA.

The ratio for an F-test is between-subjects variance / within-subjects variance. The blue double arrows are the variance “within” each group. Notice how the blue double arrow goes from one end of the group’s distribution to the other end. We pool that variance together to see how much variance is within each group. The one long orange double arrow is the variance that is between each group. In contrast to the blue double arrow, the orange double arrow is specifically between the group means. We want to see if the distance between the group means is greater than the distance between the spread of scores, or the distribution, when we combine the groups. Recall for ratios, if the numerator is greater than the denominator, we get a positive number, if the denominator is greater than the numerator, we get a number less than one, and if the numerator and denominator are equal, we get one.

For the F-test, we want to see large positive numbers. The larger the number, the better. The ANOVA or the F-test is significant if the F-test value is greater than the threshold value and the p -value is less than your alpha level. Usually, the F-test value threshold is 1.96, $N \geq 30$ per group, p - is set at .05. Recalling the 1.96 value, just like the t-test, means the between-subject variance is twice as large as the within-subject variance. For the following situations:

- F-test = 1.03, p - = .38. test not significant.
- F-test = 4.39, p - < .01. test is significant.

This page titled [12.2: The Concept Behind Analysis of Variance](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

12.3: - Interpreting the F-test and Post-Hoc Comparisons

All the F-test will tell you is that something is happening. But it will not tell you exactly what is happening. If you have a significant F-test, you know there are differences between the groups, but do not know which groups are responsible for the significant F-test.

At this point, we introduce post-hoc comparisons. Post-hoc means after the fact. We only do the post-hoc tests after we have a significant F-test, or for any significant overall statistical result. Statistical tests such as the F-test are overall tests; they indicate that somewhere among all of your groups or variables, there is a significant effect. Consider post-hoc tests as specific tests that determine where the significance is. For an ANOVA, the post-hoc comparisons indicate where the effect is, or which groups are different. If the F-test is not significant, there is no need to do the post-hoc comparisons.

The purpose of the post-hoc is to determine which two groups are significant. You can only compare two groups' means at a time. When you compare two groups at a time, you use a post-hoc test, which, in this case, is actually a version of the t-test. Recall that a t-test examines the means between two groups. Hence, we use a t-test here to compare the means among all the groups, one pair of groups at a time.

Why do we use post-hoc t-tests? Why not use a series of t-tests? We use the post-hoc t-test instead of the regular, or independent, t-test because we need the post-hoc test to adjust for the Type I error, or account for the experimenter-wise error rate.

Recall that a Type I error is when you find a significant effect, but in reality, you find a null effect or nothing. In what situations would you encounter a Type I error? One situation is what we call “fishing. Recall that in fishing, you continually cast your fishing line into the lake, over and over again. If you keep fishing, eventually you will catch something. But catching something after repeated tries does not mean that you can claim to be a fisherman, which is a significant finding, because in reality, you really are not a fisherman, and catching the fish was just a random lucky effect.

Where is the “fishing” in this situation? Conducting multiple t-tests is the same as casting your fishing line into the lake over and over again. Eventually, you will find a significant effect because you kept running t-tests after t-tests. Something will happen eventually. But that does not mean you found a significant effect. You got lucky because you kept running test after test. It is like dating. If you keep asking someone out on a date, eventually they will say yes (at least in theory, I have never found that to be the case, sigh).

Instead of just multiple t-tests and the basic independent t-tests, statisticians devised a series of post-hoc t-tests to adjust the p -value for the experimenter-wise error, or the increased likelihood of committing a Type I error due to conducting several t-tests. The common ones you will see associated with an F-test are Tukey, Sidak, and Bonferroni. These t-tests were devised to handle anomalies across the groups, such as unequal sample sizes or different distributions across the groups. The mathematical differences among the t-tests are beyond the scope of this textbook. One way to use these t-tests is to select them as your post-hoc tests, and if all three post-hoc tests give you the same answer, then you have confidence in your results. If one says significance and the other two do not, or two say significance, and one does not, then consult with a statistician, because something is off.

Once you get a significant F-test, all you can say is that something is significant. Then, you proceed to interpret the post-hoc tests to determine which group is significantly higher or lower than the other.

Sometimes, not all groups might be different; sometimes, only one group is causing the F-test to be significant. It is possible to have a significant F-test, without any significant post-hoc tests. There are no clear, consistent reasons for these scenarios. If these situations happen, consult a statistician.

This page titled [12.3: - Interpreting the F-test and Post-Hoc Comparisons](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

12.4: Commentary

The commentary for the ANOVA is mostly about how there are different terms for the ANOVA.

12.4.1: ANOVA Terminology

Be mindful of the terminology when discussing the ANOVA and the F-test research design. There are different terms that mean the same thing. Sometimes we call it a turntable or a record player, pancake or flapjack, coat or jacket. There is no discernible difference between the two terms; in certain situations, people prefer one term over another. Even the terms ANOVA and F-test are interchangeable. “Running an ANOVA” or “Conducting an F-test” are the same thing.

Different terminology is used when describing the variables and their groups. Recall that the design setup for the ANOVA is the independent variable, which is nominal or categorical, and consists of three or more groups, and one dependent variable. The groups are part of an overall variable. If we want to examine race, we have one variable called “race,” and within that variable of race, we have four groups: White vs. Black vs. Hispanic/Latino vs. Asian. Sometimes we say we are analyzing groups, and what we really refer to is one variable with multiple groups. For example, we can say we are analyzing medication, mindfulness, and a control group. These are different treatment groups, and the overall variable is treatment. Sometimes we say groups, other times we say categories. They mean the same thing. We can have a variable called diagnosis, and there are three categories for that diagnosis: depression, anxiety, and substance use. The issue to note here is that while we say we are examining three or more groups, we are really saying that one nominal or categorical variable is divided into three or more groups, or categories.

Sometimes we say the independent variable has three groups or levels. It’s the same thing. We say “race” as our independent variable, and it has four levels: White vs. Black vs. Hispanic/Latino vs. Asian. It is the same thing as groups.

12.4.2: More than One IV and Terminology

There are many ways to use the ANOVA. Recall that ANOVA is not a statistical analysis with one function, in this case, to examine the variance among three or more groups. ANOVA analyzes variance, and there are many sources of variance. By sources of variances, I mean ways to vary, and everything can vary with two or more variables, not just groups, over time, and with additional continuous variables.

Here we will introduce the idea that ANOVA can be used with two or more independent variables (IV), not just two or more groups. You can use two variables, race and gender. Or you can use treatment and gender. In these examples, the groups are nominal, categorical variables. Each variable can have two or more groups. So, for race we can have Black, White, Hispanic, Asian, for gender we can have male, female, non-binary, for treatment we can have treatment, control, and usual care.

The terms statisticians use when we have two or more IVs are factor and way. What is the difference between these terms? I do not know, and it does not make any difference as far as I am concerned.

The term factor is used in the phrases in Factorial Analysis of Variance, two-factor ANOVA, or an ANOVA with two factors. The factor refers to an independent variable that is nominal/categorical. When we say two factors, we are indicating that we have two independent variables, and the variables can have two or more groups or categories. For example, a two-factor ANOVA means we have gender and race, or treatment and gender. Sometimes we say one-way ANOVA or two-way ANOVA. The one-way usually refers to one variable with multiple groups, such as race: White vs. Black vs. Hispanic/Latino Vs. Asian. Two-way usually refers to two variables, each with multiple groups, such as gender: male vs. female, and Race: White vs. Black vs. Hispanic/Latino Vs. Asian.

Researchers use different terms because that is just how they were taught. It is just different terminology. There is no standard way of describing something. If someone corrects you about using factor or way, then tell them “Bite me,” because you are not making any violations or errors if you casually use one term of the other.

So why learn these terms? For the darn EPPP exam. These exams will test you with story problems, and the problem might disguise the independent variables and the groups. You have to read closely to determine what the IV is and how many categories/groups/levels there are.

12.4.3: Factorial Analysis of Variance

Now that we have introduced the idea that ANOVAs can have more than one categorical/nominal IV, we can introduce the idea of how we usually analyze two or more IVs. When using this design, typically we are interested in interaction effects. Interactions are like combining the effects of two variables. Suppose you are interested in examining the effects of gender and race on an outcome variable. You have gender: male, female, and race: White, Black, Hispanic, and Asian. If you combine all of them, you have the following: male White, female White, male Black, female Black, male Hispanic, female Hispanic, male Asian, female Asian. For now, that is all we will say about interactions. And ANOVAs can have two or more independent variables.

12.4.4: Analysis of Covariates (ANCOVA)

ANOVA is useful for analyzing covariates. What is a covariate? Up to now, you have learned about independent variables and dependent variables. Covariates are additional independent variables but are not your independent variable of interest. Huh?

When we use covariates as part of an ANOVA, we call it ANCOVA, or analysis of covariance. Covariates are variables that are conceptually related to the DV, or useful for predicting the DV. Think of covariates as something you have to include when predicting the DV, because leaving out those covariates means missing important information. However, covariates are not your independent variable, which means the updated definition of an independent variable is not just something that predicts the outcome but something that is of interest when predicting the outcome.

The difference between covariates and independent variables is that you want to include a covariate because you are interested in how this additional variable explains the DV and then see if your IV still predicts any variation in the DV after accounting for the covariate.

For example, the IV is treatment – mindfulness treatment vs. control. The DV is depression severity. The covariate is gender. The reason why we include gender as a covariate is that it is known that males and females differ in the level of depression. Males and females also differ in their level of acceptance of mindfulness as a treatment option. We should include gender as a covariate because it likely impacts how the participants respond to the treatment, and how it affects the outcome of depression, but the variable of interest is treatment, and the question is if mindfulness treatment lowers depression, after we account for how the covariate, gender, affects both the treatment and the outcome of depression.

Covariates are usually demographic variables. We want to account for known demographics and their impact on the outcome. The usual demographics are gender, age, race, income, and level of education. All these demographics usually have an impact on psychological outcomes.

Covariates can also be known as psychological processes that undoubtedly need to be acknowledged as impacting the independent and dependent variables. Continuing with our example, let us use the covariate level of stress. The idea is that people get depressed based on their stress level. Mindfulness can help with depression, but only after we account for how much depression is associated with stress. People who are more stressed may not be able to fully engage in a mindfulness treatment, which is what the mindfulness treatment is supposed to address in the first place. We should account for their stress level first, then see if the mindfulness treatment works. We need to account for the stress level so we can isolate the effect of just mindfulness treatment on depression.

Accounting for covariates helps. Including covariates is a good idea because the covariate usually represents what we already know as impacting the outcome. We know that gender and stress are associated with depression. However, the independent variable is the one of interest. In this example, the research question is, how much more can mindfulness reduce depression after we account for the effects of gender and stress? Conceptually, it helps to add covariates because we cannot talk about the outcome, in this case depression, without acknowledging the known effects of covariates, such as gender and stress, on the outcome.

From a statistical viewpoint, adding covariates reduces the unexplained variance in the DV. Recall that the goal of statistics is to explain variation. We want to know why something varies. We try to explain 100% of the variation in the DV by using the IV. But the IV does not explain everything. If we use just the IV, we will explain some of the variance, and whatever variance is left over is the error variance. In this case, the error variance is unexplained variance, not necessarily random error. We simply do not know what else could explain the variation in the DV. We include a covariate and say, “This percentage of the variability in the DV can be explained by the covariate.” We can say that this percentage of variability in the DV is explained by the covariate, and this percentage of variability is explained by the IV. We can compare how much variance is explained by the covariate and the IV.

Hopefully, the percentage of variance explained by the IV is greater than the percentage of variance explained by the covariate. That is what we really want to know.

Of note, do not use covariates haphazardly; choose covariates based on a theoretical premise. The above example makes it seem like we should include as many covariates as possible. However, including too many covariates can reduce your degrees of freedom, depending on your sample size. Always in statistics, do not just throw in all the variables into your analysis. That is sloppy and poor planning.

But this section's main message is to include covariates whenever possible, which means, try to use an ANCOVA whenever possible.

Exactly how do you conduct an ANCOVA? Well, in SPSS, the option to include covariates is already incorporated into an ANOVA. There really is no such thing as a separate selection of ANOVA vs. ANCOVA. When you operate a statistics program, you do not see an option to select ANOVA or select ANCOVA. The option to include covariates is built into the ANOVA stats analysis. You don't need to select an ANCOVA versus an ANOVA. The ANOVA already includes the covariate option.

From a research design perspective, you will find that including covariates helps establish the equivalency of the group research design or to rule out potential confounds. For an equivalent group research design, we want to establish the equivalency of two groups. This is a research method issue, so for now, the best research design is when everything is the same, except for the unique factor between the two groups. Continuing with our example, if we want to examine the effect of mindfulness versus no mindfulness on depression, we want to establish that the mindfulness group is equivalent to the no mindfulness group in every way, except for the mindfulness. We include gender, age, race, income, and level of education to ensure that we have equal numbers of these demographics. We never have equivalent amounts, so we include these demographics as covariates so we can account for the variance in the demographics, leaving the variance accounted for by the mindfulness group. So, we “filter out” the covariates to establish a semblance of equivalency between the two groups. The same idea applies to ruling out confounds. In brief, confounds are factors that are alternative explanations for the outcome. Demographics could be confounded. So, we include the demographics as covariates to rule out the possibility that the demographics could be responsible for the outcome. Continuing with our example, by including gender, age, race, income, and level of education, we rule out the possibility that younger, female, higher income, and higher education have something to do with reducing depression, rather than the mindfulness treatment. The lesson here is that we use covariates to establish the equivalency between treatment and control groups, and we use covariates as a method to rule out confounds that could negate treatment results. , so be on the lookout for them.

12.4.5: Trend Analysis

ANOVA and the post-hoc tests are used for what is called trend analysis. Trend analysis is used to examine trends among the groups. There are two types of trends, linear and curvilinear. We are looking for an increase in the mean scores for each group. When does this situation occur? For example, we can examine education budgets, which show an increase when we sort the states from lowest to highest education budgets. We can examine crime rates as we sort the neighborhood crime rate from lowest to highest. The states and the neighborhoods are the groups, the education budgets and crime rates are the outcome variables, and we want to see if they increase in a linear or a curvilinear fashion across the groups.

These linear or curvilinear increases or decreases across the groups are best illustrated when we turn to longitudinal analyses. For me, I hate referring to these analyses as trend analyses because (a) the group order is arbitrary, and (b) trend makes me think of a trend over time, not necessarily by groups. For now, if there is a licensing exam question that asks you what the analysis of choice is to analyze changes in mean scores across groups, the answer is likely a trend analysis, and the analysis is an ANOVA or an F-test, with post-hoc tests.

12.4.6: T-test and ANOVA

The t-test examines two groups only; the ANOVA analyzes one independent variable that is nominal/categorical with several groups, or more than one independent variable with two or more groups.

What happens if you run an ANOVA with two groups? Will your statistical software explode? Nope. You can run an ANOVA with one independent variable and two groups. You will get an F-test, and if you take the square root of the F-test, you will have a value equivalent to the t-test if you conduct a t-test with the two groups. Or, if you run a t-test and square the t-test value, you will get a

value equivalent to the F-test value. Basically, a t-test squared is equivalent to an F-test with one independent variable and two groups.

Does this matter? Nope. Do you get more information using ANOVA to examine mean differences between two groups? Nope. Is it a huge error to use an ANOVA when you should have used a t-test? Nope. All you need to know is that the t-test value squared is equivalent to the F-test, which is useless information, but it might be information for a potential licensing exam question. And that is all.

12.4.7: BTW.... I hate MANOVA

The ANOVA can only examine one dependent or outcome variable. If you have more than one dependent or outcome variable, the analysis is called a MANOVA, which stands for Multiple Analysis of Variance. The term multiple in the MANOVA refers to multiple outcomes or dependent variables. The word “multiple” refers to multiple independent or predictor variables. Go figure.

if you had multiple dependent variables, you could conduct several ANOVAs, one for each dependent variable, and call it a day. So why don't we? Here is your EPPP alert – Q: Why do you use MANOVA? A: You do not conduct multiple ANOVAs because you want to avoid the experimenter error-wise rate. Recall the problem with conducting too many t-tests, or too many statistical tests in general. Running multiple ANOVAs is “fishing; when you continue to test you will eventually find an effect, but it’s likely due to random chance, which is a Type I error. The experimenter error-wise rate is the same as an inflated Type I error. You increase the chances of finding something, but it’s a random effect because you run several ANOVA tests. If we have multiple dependent variables, we use a MANOVA to avoid the experimenter-wise error rate, which is the answer to that licensing exam question.

The second reason I am addressing this issue now is this – I hate MANOVA. Why do I hate MANOVA? Because everyone automatically thinks that you have multiple DVs you HAVE to use a MANOVA. But every study has multiple outcome variables. If you think about it, every issue and phenomenon can have multiple outcomes because everything is multi-dimensional. Academic achievement is multi-dimensional; it can be divided by subject matter, GPA, and standardized test scores. Alcohol consumption is multi-dimensional; it can be divided by the number of drinks, frequency of drinking, time frame during which the drinking occurred, and attitudes about drinking. There are always multiple outcomes, which does not mean we conduct MANOVAs for everything, and it would mean there are no situations where you would ever conduct an ANOVA for a single outcome variable. In the end, what will happen is that SPSS MANOVA will provide you with an overall multivariate effect, but then, similar to a post-hoc t-test, the analysis will break down the multivariate effect with several ANOVA. In essence, you conduct multiple ANOVAs anyway to determine which groups had what effect on which dependent variable.

Why would you use a MANOVA? You pick multiple outcome variables because they are conceptually related. Conceptually related, they need to have something in common, and that commonality is based on your explanation about why the combination, or the collection of the outcome variables, is necessary for you to understand the issue at hand. There needs to be a reason you are interested in the multivariate effect, or why are all those outcome variables related? There needs to be a theoretical reason. If you don’t know how those are related theoretically, do not conduct a MANOVA.

An example of related dependent variables in psychology could be self-harm. There are multiple reasons why individuals engage in self-harm, such as seeking help from others, inflicting punishment on themselves when others wrong them, feeling something other than negative feelings, and inflicting punishment on themselves for feeling negative about themselves. These reasons can be related in terms of the functions of self-harm. A MANOVA might be appropriate to conduct here to see if group effects, such as treatment vs. control groups, have multiple effects across all these reasons for using self-harm.

Instead of using MANOVA, my preference would be to use structural equation modeling. Other statisticians likely have other preferences. The point is this – if someone tells you that you should be using a MANOVA when you have multiple outcome variables, tell them “bite me” and call Dr. Ji, because automatically using a MANOVA just because you have multiple outcome variables indicates to me that they do not fully understand the concept of a MANOVA.

This page titled [12.4: Commentary](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

12.5: Reading Articles – ANOVA

We revisit the steps for reading journal articles. First, always identify your research question. Guiding questions include: What's the goal of the study? What's the study trying to contribute to the field? What are the hypotheses?

Second, identify the logistics of the statistics. Guiding questions include: What is the scoring protocol, what do high and low scores mean? What groups are being compared?

Third, review the ANOVA results. Highlight relevant F-test results. Report the means and the F-test results. If there is a significant F-test, look for the accompanying post-hoc tests to determine which groups are higher or lower than the other. Do the findings confirm the hypotheses? Note any possible Type I or II errors, any effect sizes worth noting, and any sample size/power issues.

Fourth, the conclusion. Were the authors successful in accomplishing their goal? Guiding questions: Comment on what it means that the hypothesis was supported. It is also worthwhile to comment on which hypotheses were not supported. Did that help us understand the issue? Did the findings tell you something new about the issue under investigation?

We'll use the article by Erermis, S., et al. (2004) as an example: First question: What is the purpose of the study? For this study, the researchers wanted to know if obesity was associated with mental health difficulties. Apparently, the connection between obesity and mental health difficulties is weak or inconsistent, and there is no evidence that obesity is related to psychological difficulties that are not at the level of a mental health diagnosis.

Second question: The three groups are clinical obesity, non-clinical obesity, and normal weight controls. Clinical obesity adolescents were sampled from an outpatient clinic for pediatric endocrinology. The non-clinically obese and normal-weight adolescents were sampled from a high school. Their variables were the scales on the Child Behavior Checklist (CBCL), and higher scores indicate more levels of the scale's construct. Table 1 presents their findings.

Table 1: Mean scores for clinical and non-clinical obese groups and non-obese group on the Child Behavior Checklist (CBCL)

CBCL subscales	Clinical obese group (n=30)	Nonclinical obese group (n=30)	Normal weight controls (n=30)	F	P	Comparison between groups
Activity	42.5+-9.1	43.4+-7.7	46.9+-8.1	2.31	0.104	NS
School competence	43.2+-12.1	43.3+-4.5	45.0+-5.9	0.44	0.644	NS
Total competence	41.8+-13.9	41.5+-6.9	45.1+-6.9	1.2	0.304	NS
Aggressiveness	59.2+-8.4	55.1+-6.4	54.7+-6.1	3.69	0.028	1 > 2,3
Delinquent behavior	57.1+-7.5	52.8+-3.8	54.3+-6.6	3.67	0.029	1 > 2
Anxiety-depression	65.1+-10.8	60.0+-8.2	56.8+-7.2	6.66	0.002	1 > 2,3
Attention	61.4+-9.1	57.5+-6.9	56.8+-6.9	3.11	0.049	1 > 3
Social problems	64.9+-8.0	60.3+-6.4	54.7+-6.4	15.65	0.0001	1 > 2,3; 2 > 3
Somatic problems	62.5+-10.7	61.4+-8.6	53.5+-4.5	10.32	0.0001	1 > 3; 2 > 3
Thought problems	59.1+-7.9	57.1+-7.7	57.0+-6.8	0.74	0.477	NS
Social withdrawal	62.8+-9.2	55.5+-6.3	54.1+-6.6	11.59	0.0001	1 > 2,3
Externalizing behavior	57.9+-9.4	52.6+-7.8	51.4+-9.8	4.34	0.016	1 > 2,3
Internalizing behavior	65.8+-9.6	59.8+-8.6	53.0+-9.6	14.25	0.0001	1 > 2,3; 2 > 3

Third question: What are the results? As with all statistical tables, start in the upper left corner, work your way left to right, and up and down. Start with the first variable, “Activity.” The clinical obese group has a mean of 42.5, the non-clinical obese group has a mean of 43.4, and the normal weight controls have a mean of 46.9. The F-test is 2.31, and the *p*-value is .10. This result means there is no significant difference between the three groups in their activity level.

So, is this finding expected? Actually, it isn’t. Let us walk through this result. If I did not know any better, I would think that the clinically obese group would have a lower activity level than the normal weight control group, possibly also lower than the non-clinically obese group. It turns out that all three groups are similar in activity level. Yes, the clinical and non-clinical obese groups are slightly lower in activity level than the normal weight control group, but the difference is not significant. This finding negates a persistent myth: obese adolescents are lazy and do not do much, which contributes to their obesity. This finding does not support this persistent myth, which is probably both good and bad news. It is good news because this finding dispels the idea that obese youth are lazy, and a lack of activity contributes to their obesity. It might be bad news because increasing the activity of obese youth may not result in weight loss.

Let us take a significant result, for the variable aggressiveness. The clinical obese group has a mean of 59.2, the non-clinical obese group has a mean of 55.1, and the normal weight controls have a mean of 54.7. The F-test is 3.69, and the *p*-value is .028. This result means there is a significant difference between the three groups in their aggressiveness level. The F-test is significant, and all it says is that there is something going on in the mean scores among the groups. It does not say exactly what is going on.

For that information, we turn to the post-hoc tests. The researchers, as do most researchers in journal publications, list the post-hoc tests as comparisons among groups. For example, the post hoc test could be noted as $1 > 2 > 3$, which means group 1 has a higher mean than group 2, and group 2 has a higher mean than group 3. In this case, the post hoc test was noted as $1 > 2, 3$, which means that group 1, the clinical obese group, has a higher mean and more aggressiveness than both groups 2 and 3, the non-clinical obese group and the normal weight control group, respectively.

The effect size was not reported, but as a rough estimate, the difference between the clinical obese group and the non-clinical obese group, and the normal weight control group is about four points ($59.2 - 55.1 = 4$, $59.2 - 54.7 = 4$), and the pooled average standard deviation for the three groups is about seven (clinical obese group, $SD = 8.4$; non-clinical obese group, $SD = 6.4$; normal weight control group, $SD = 6.1$, average $SD \sim 7$), so the effect size is about $4/7$ about .6, which is a moderate effect size. The moderate effect size means that the difference between the clinical obese group and the non-clinical obese group and the normal weight controls is noticeable.

What does this result mean? Always think conceptually. First, it seems unexpected. There is no immediate idea why clinically obese adolescents would be more aggressive than normal weight controls. To think conceptually, always look for an explanation, and for explanations, try to think of a perspective, a framework, or a theory. For this result, from a social skill or bullying framework, perhaps clinically obese adolescents are being bullied by others for their obesity. It is possible that clinically obese adolescents need to be aggressive to defend themselves. This scenario is possible, but if that were true, we would see the non-clinically obese adolescents have similar levels of aggressiveness with the clinically obese adolescents, and certainly more than normal weight controls. But that result did not happen here. Clinically obese adolescents have more aggressiveness than non-clinically obese adolescents. For this unexpected result, let us try a different framework. Maybe a medical framework would help. The clinically obese adolescents were recruited from a pediatric endocrine clinic. Endocrinology addresses hormones. Maybe the hormones in clinically obese adolescents contribute to their increased aggression. Admittedly, I do not know the connection between obesity and aggressiveness for adolescents, but it is one possibility given that the clinically obese adolescent population was recruited from a medical context, specifically a clinic.

A quick review of the remaining results suggests the clinically obese adolescent group does have higher psychopathology than the non-clinical and normal weight controls. This result is consistent and suggests a pattern, which is that clinically obese adolescents have more psychopathology. It does raise the question if the pathology is related to being obese, but if so, we would have thought that both the clinical and non-clinical obese adolescent groups would be higher than the normal weight controls. An explanation may elude us for now, but the results do indicate that we need to pay attention to the emotional, social, and mental health needs of clinically obese adolescents.

This page titled [12.5: Reading Articles – ANOVA](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

12.6: Summary

The ANOVA has many uses. It is primarily used to compare group differences among three or more groups; it has other uses. We compare groups all the time: treatment vs. other treatments and control, different races, different genders and sexual orientation identities, different political parties, clinical, sub-clinical or non-clinical populations, in-patient, outpatient or residential client facilities. Yes, groups are limiting in a binary world, and putting individuals in a group does miss their individuality. However, group comparisons are inevitable, and when there is a conceptual basis for the group comparison, they can yield interesting results.

References

Erermis, S., Cetin, N., Tamar, M., Bukusoglu, N., Akdeniz, F., & Goksen, D. (2004). Is obesity a risk factor for psychopathology among adolescents? *Pediatrics International*, 46(3), 296–301. <https://doi.org/10.1111/j.1442-200x.2004.01882>.

Helpful Websites

<https://www.jmp.com/en/statistics-kn.../one-way-anova>

<https://datatab.net/tutorial/anova>

This page titled [12.6: Summary](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

12.7: Discussion Questions

Discussion Questions

1. Aside from the obvious, what is the difference between a t-test and an ANOVA?
2. Why conduct an ANOVA versus multiple t-tests?
3. What are post-hoc tests, and why do we use them instead of multiple t-tests?
4. What are the different terminologies when we discuss t-tests?

This page titled [12.7: Discussion Questions](#) is shared under a [CC BY 4.0](#) license and was authored, remixed, and/or curated by [Peter Ji](#).

Index

A

accessible education

[LibreTexts Information](#)

active_link

[Licensing](#)

actual effects

[2.1: Statistics Involves Comparison and Context](#)

adaptive textbooks

[LibreTexts Information](#)

addiction

[4.2: Categorical Distributions](#)

ADHD

[8.2: Statistical Decision - Is a Result Significant or Not Significant?](#)

advantages

[1: What is a Number?](#)

alcoholism

[1.3: Variations Lead to Numbers](#)

alpha

[8.10: The Definition of Type I and Type II Errors](#)

alpha level

[8.4: How to Determine if a Result is Significant](#)

alternative hypothesis

[11.2: Design Set Up](#)

alternativeHypothesis

[8.3: Hypotheses for Statistical Inference](#)

[12.1: When to Use the ANOVA](#)

analysis

[1.1: How to Start with Statistics](#)

[1.9: Seeing Numbers – What I Want You to Learn](#)

[10.1: The Meaning of Inferential](#)

[11.3: Commentary](#)

Analysis of Variance

[12: Analysis of Variance – ANOVA](#)

AnalysisOfVariance

[12.1: When to Use the ANOVA](#)

anomalies

[4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables](#)

anova

[2.4: Two Broad Categories of the Process of Using Statistics](#)

[3.6: Making Decisions About the Variables in Your Research Study](#)

[3.7: Lessons Learned About Variables](#)

[4: Distributions](#)

[4.8: Evaluating a Continuous Distribution](#)

[8.5: Interpreting a Statistical Test Result](#)

[12: Analysis of Variance – ANOVA](#)

[12.1: When to Use the ANOVA](#)

[12.2: The Concept Behind Analysis of Variance](#)

[12.3: - Interpreting the F-test and Post-Hoc](#)

[Comparisons](#)

[12.4: Commentary](#)

[12.5: Reading Articles – ANOVA](#)

[12.6: Summary](#)

[12.7: Discussion Questions](#)

anxiety

[1.2: A "Geek Out" Explanation of Variation](#)

[4.2: Categorical Distributions](#)

APAstyle

[8.6: Testing a Hypothesis](#)

application

[0.6: Description of Intended Audience](#)

area under the curve

[6.2: How Do You Determine Variance?](#)

assessment

[1.11: Discussion Questions](#)

assumptions

[11.4: T-test Assumptions](#)

[11.10: Discussion Questions](#)

attachment

[1.1: How to Start with Statistics](#)

audience

[0: Introduction](#)

autism behavior

[10.2: Patterns](#)

B

back_matter

[Licensing](#)

Bayesianmodels

[10.3: General Linear Model](#)

behavioral problems

[11.4: T-test Assumptions](#)

bellCurve

[7.5: Central Limit Theorem \(CLT\) and Its Implications](#)

beta

[8.10: The Definition of Type I and Type II Errors](#)

betweenGroups

[12.2: The Concept Behind Analysis of Variance](#)

bias

[1.4: The Battle of Words Versus Numbers](#)

biserial correlation

[11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology \(EPPP\)](#)

[Licensing Exam](#)

Bonferroni

[12.3: - Interpreting the F-test and Post-Hoc Comparisons](#)

C

cancer

[1.3: Variations Lead to Numbers](#)

categorical

[3: Variable Types](#)

[3.5: Nuances in Determining if a Variable Should be a Certain Type](#)

[11.2: Design Set Up](#)

categorical data

[3.4: Continuous Variables- Variables That Vary by Level](#)

categorical distribution

[4.1: Distributions – A Picture of Variation](#)

[4.4: Continuous Distributions](#)

[4.9: Discussion Questions](#)

categorical distributions

[4.2: Categorical Distributions](#)

categorical variables

[4.2.3: Evaluating the Plot](#)

[4.9: Discussion Questions](#)

[7.2: Sampling from the Population](#)

categoricaldistributions

[4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables](#)

category

[3.3: Categorical Variables- Variables That Vary by Type](#)

central

[8: Probability, Hypothesis Testing, Type I and Type II Errors](#)

central limit theorem

[7: Sampling, Standard Error, Central Limit Theorem](#)

[7.6: Summary of Reasons Why You Need a Large Sample Size](#)

central tendency

[5.1: What is Central Tendency? Why Do We Use It?](#)

[5.2: Three Types of Central Tendency and Why We Need Them](#)

[6.1: What is Variance?](#)

central tendency measures

[5.3: The Mean, Median, and Mode in Normal and Skewed Distributions](#)

central_tendency

[5: Measures of Central Tendency](#)

[5.4: Discussion Questions](#)

CentralLimitTheorem

[7.5: Central Limit Theorem \(CLT\) and Its Implications](#)

chisquare

[8.5: Interpreting a Statistical Test Result](#)

classifications

[Front Matter](#)

clients

[11.1: When to Use the Independent T-test](#)

clinical

[3.4: Continuous Variables- Variables That Vary by Level](#)

clinical outcomes

[8.2: Statistical Decision - Is a Result Significant or Not Significant?](#)

clinical practice

[0.4: Scientist-Practitioner Model; Scholar-Practitioner Model](#)

clinical studies

[12.6: Summary](#)

clinicaldata

[0.3: Evidence-Based Practice](#)

clinicalsignificance

[9.1: Significant \(P Value\) Versus Meaningful \(Effect Size\) Results](#)

CLT

[7.5: Central Limit Theorem \(CLT\) and Its Implications](#)

collaboration

[0.7: Approach to Learning Statistics](#)

collaborative learning

[LibreTexts Information](#)

communication

[1.4: The Battle of Words Versus Numbers](#)

[1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words](#)

[1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift](#)

comparison

- 1.1: How to Start with Statistics
- 1.4: The Battle of Words Versus Numbers
- 1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words
- 1.9: Seeing Numbers – What I Want You to Learn
- 1.10: Connection Between Understanding How Things Vary and What You are Trying to Do
- 2: What are Statistics?
- 2.1: Statistics Involves Comparison and Context
- 4.2.3: Evaluating the Plot
- 4.6: Normal Distribution for Continuous Variables

comparisons

- 11.3: Commentary

confidence intervals

- 8.8: Confidence Intervals

confounds

- 4.2.3: Evaluating the Plot

consequences of sampling bias

- 7: Sampling, Standard Error, Central Limit Theorem

constructs

- 1: What is a Number?
- 1.2: A "Geek Out" Explanation of Variation
- 1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words
- 1.11: Discussion Questions

context

- 1.2: A "Geek Out" Explanation of Variation
- 3.4: Continuous Variables- Variables That Vary by Level

continuous

- 3: Variable Types
- 3.5: Nuances in Determining if a Variable Should be a Certain Type

continuous data

- 3.4: Continuous Variables- Variables That Vary by Level

continuous distribution

- 4.4: Continuous Distributions
- 4.9: Discussion Questions

continuous variable

- 11.3: Commentary
- 11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology (EPPP) Licensing Exam

continuous variables

- 4.4: Continuous Distributions
- 4.5: Evaluating the Quality of Frequency Distributions
- 4.9: Discussion Questions
- 7.2: Sampling from the Population
- 11.2: Design Set Up

continuousVariables

- 12.1: When to Use the ANOVA

control

- 8.6: Testing a Hypothesis
- 11.3: Commentary

control group

- 8.1: The Role of Probability

convenience sampling

- 7.3: Getting a Good Sample - Get a Random Sample

convergence

- 7.5: Central Limit Theorem (CLT) and Its Implications

coping skills

- 3.2: How to Organize Variation With Variables

correlated samples

- 11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology (EPPP) Licensing Exam

correlation

- 2.4: Two Broad Categories of the Process of Using Statistics
- 3.7: Lessons Learned About Variables
- 4: Distributions
- 8.5: Interpreting a Statistical Test Result
- 9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together

correlation coefficient

- 9.2: What Are Effect Sizes?

correlation studies

- 6.4: Standardized Values of Reporting Scores

correlations

- 2.2: Statistics is About Finding Patterns
- 4.8: Evaluating a Continuous Distribution

counseling sessions

- 10.2: Patterns

curriculum

- 0.6: Description of Intended Audience

curriculum development

- LibreTexts Information

D

data

- 1.7: Statistics Do Something with Numbers
- 2.2: Statistics is About Finding Patterns
- 3.5: Nuances in Determining if a Variable Should be a Certain Type

data analysis

- 0.4: Scientist-Practitioner Model; Scholar-Practitioner Model
- 1.3: Variations Lead to Numbers
- 1.10: Connection Between Understanding How Things Vary and What You are Trying to Do
- 2.3: Statistics as Estimates, or Making Guesses
- 3.3: Categorical Variables- Variables That Vary by Type
- 3.4: Continuous Variables- Variables That Vary by Level
- 4.1: Distributions – A Picture of Variation
- 4.2.3: Evaluating the Plot
- 4.7: Skewed Distribution- The Opposite of Normal Distribution
- 4.9: Discussion Questions
- 5.1: What is Central Tendency? Why Do We Use It?
- 5.2: Three Types of Central Tendency and Why We Need Them
- 7.6: Summary of Reasons Why You Need a Large Sample Size
- 7.7: Discussion Questions
- 9.6: Discussion Questions

data collection

- 3.6: Making Decisions About the Variables in Your Research Study
- 6.3: Partitioning of the Variance
- 7.1: The Purpose of Sampling
- 7.3: Getting a Good Sample - Get a Random Sample
- 9.3: Statistical Power

data evaluation

- 4.5: Evaluating the Quality of Frequency Distributions

data interpretation

- 3.1: Variables - Not Always Obvious
- 11.5: Reading T-tests

data representation

- 5.2: Three Types of Central Tendency and Why We Need Them

data_analysis

- 0.5: Social Justice
- 2.4: Two Broad Categories of the Process of Using Statistics
- 5: Measures of Central Tendency
- 5.4: Discussion Questions

dataanalysis

- 0.3: Evidence-Based Practice
- 4.8: Evaluating a Continuous Distribution
- 5.3: The Mean, Median, and Mode in Normal and Skewed Distributions
- 8.5: Interpreting a Statistical Test Result

datainterpretation

- 9.1: Significant (P Value) Versus Meaningful (Effect Size) Results
- 11.7: Reading Articles That Involve a T-test

dataset

- 1.7: Statistics Do Something with Numbers
- 5.1: What is Central Tendency? Why Do We Use It?

dating

- 1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift

decision tree

- 3.8: Discussion Questions

decisionmaking

- 8.3: Hypotheses for Statistical Inference

demographic

- 3.5: Nuances in Determining if a Variable Should be a Certain Type

demographic groups

- 4.2.3: Evaluating the Plot

demographic variables

- 4.2: Categorical Distributions

demographics

- 1.3: Variations Lead to Numbers
- 3.2: How to Organize Variation With Variables
- 3.3: Categorical Variables- Variables That Vary by Type
- 3.7: Lessons Learned About Variables
- 8.2: Statistical Decision - Is a Result Significant or Not Significant?
- 12.4: Commentary
- 12.6: Summary

demographicvariables

- 4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables

dependent

- 3: Variable Types
- 3.6: Making Decisions About the Variables in Your Research Study

dependent variable

- 11.2: Design Set Up
- 12: Analysis of Variance – ANOVA

dependent variables

- 3.2: How to Organize Variation With Variables
- 3.7: Lessons Learned About Variables
- 3.8: Discussion Questions

dependentvariable

- 12.1: When to Use the ANOVA
- 12.4: Commentary

depression

- 1.1: How to Start with Statistics
- 1.10: Connection Between Understanding How Things Vary and What You are Trying to Do
- 1.11: Discussion Questions

descriptive

- 8: Probability, Hypothesis Testing, Type I and Type II Errors
- 10: Introduction to Inferential Statistics

descriptive statistics

- 2: What are Statistics?
- 4: Distributions
- 4.1: Distributions – A Picture of Variation
- 5.1: What is Central Tendency? Why Do We Use It?
- 5.3: The Mean, Median, and Mode in Normal and Skewed Distributions

descriptive studies

- 6.4: Standardized Values of Reporting Scores

descriptive_statistics

- 2.4: Two Broad Categories of the Process of Using Statistics

design

- 11.2: Design Set Up

designsetup

- 11.10: Discussion Questions

desired power

- 9.4: How to Conduct a Power Analysis

detailed_breakdown

- Licensing

deviation scores

- 6.2: How Do You Determine Variance?

dialog

- Front Matter

dichotomy

- 11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology (EPPP) Licensing Exam

discourse

- 1.4: The Battle of Words Versus Numbers

discrimination

- 0.5: Social Justice

dissertation

- 0.1: A Perspective on Statistics
- 0.2: Psychology Profession

distribution

- 8: Probability, Hypothesis Testing, Type I and Type II Errors

distribution types

- 4.7: Skewed Distribution- The Opposite of Normal Distribution

distributions

- 4.1: Distributions – A Picture of Variation
- 11.1: When to Use the Independent T-test

document

- Licensing

druguse

- 0.3: Evidence-Based Practice

DV

- 10.2: Patterns

E

eating disorders

- 2.1: Statistics Involves Comparison and Context
- 7.2: Sampling from the Population

Education

- 0.1: A Perspective on Statistics
- 0.2: Psychology Profession
- 0.6: Description of Intended Audience
- 0.7: Approach to Learning Statistics
- 1.4: The Battle of Words Versus Numbers
- 4.7: Skewed Distribution- The Opposite of Normal Distribution

effect size

- 9: Effect Sizes and Power Analysis
- 9.2: What Are Effect Sizes?
- 9.3: Statistical Power
- 9.4: How to Conduct a Power Analysis
- 9.6: Discussion Questions
- 12.5: Reading Articles – ANOVA

effect sizes

- 8.8: Confidence Intervals

effectsize

- 9.1: Significant (P Value) Versus Meaningful (Effect Size) Results
- 9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together
- 11.7: Reading Articles That Involve a T-test

elementary education

- 4.7: Skewed Distribution- The Opposite of Normal Distribution

emotional intelligence

- 0.7: Approach to Learning Statistics

engagement

- 0.1: A Perspective on Statistics
- 0.2: Psychology Profession

entities

- 1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words

environmental issues

- 1.3: Variations Lead to Numbers

error variance

- 6: Measures of Variability
- 6.3: Partitioning of the Variance

errorVariance

- 12.2: The Concept Behind Analysis of Variance

estimate

- 10.1: The Meaning of Inferential

estimation

- 2: What are Statistics?
- 2.3: Statistics as Estimates, or Making Guesses
- 7.1: The Purpose of Sampling

ethnic groups

- 2.1: Statistics Involves Comparison and Context

evaluating distributions

- 4.9: Discussion Questions

evaluation

- 0.5: Social Justice
- 4.2.3: Evaluating the Plot

evidence

- 0.3: Evidence-Based Practice

evidencebasedresearch

- 0.3: Evidence-Based Practice

examination preparation

- 11: T-tests

experimental design

- 3.2: How to Organize Variation With Variables

externalizing symptoms

- 11.8: Reading T-tests from Journal Articles

F

factors pros cons

- 3.8: Discussion Questions

faculty

- 0.6: Description of Intended Audience

falsenegative

- 8.10: The Definition of Type I and Type II Errors

falsepositive

- 8.10: The Definition of Type I and Type II Errors

females

- 11.1: When to Use the Independent T-test

findings

- 8.3: Hypotheses for Statistical Inference

framework

- 0.5: Social Justice
- 2.2: Statistics is About Finding Patterns

frequency

- 1.9: Seeing Numbers – What I Want You to Learn
- 3.4: Continuous Variables- Variables That Vary by Level

- 4.4: Continuous Distributions

frequency count

- 4.1: Distributions – A Picture of Variation

frequency counts

- 4.2: Categorical Distributions

frequency distribution

- 4.1: Distributions – A Picture of Variation
- 4.5: Evaluating the Quality of Frequency Distributions

frequencystatistics

- 10.3: General Linear Model

ftest

- 9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together
- 12.2: The Concept Behind Analysis of Variance
- 12.4: Commentary

G

gender

- 1.2: A "Geek Out" Explanation of Variation
- 3.3: Categorical Variables- Variables That Vary by Type
- 10.2: Patterns
- 11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology (EPPP) Licensing Exam

gender differences

- 11.8: Reading T-tests from Journal Articles

genderdifferences

- 11.9: Summary

general

- 10: Introduction to Inferential Statistics

general linear model

- 10.4: Discussion Questions

generallinearmodel

- 10.3: General Linear Model

Gpower

- 9.4: How to Conduct a Power Analysis

graduate programs

- 0.6: Description of Intended Audience

graphical representation

- 4.5: Evaluating the Quality of Frequency Distributions

group

- 3.3: Categorical Variables- Variables That Vary by Type

group comparison

- 2.2: Statistics is About Finding Patterns
- 12: Analysis of Variance – ANOVA
- 12.6: Summary

groupcomparison

- 8.5: Interpreting a Statistical Test Result

groupcomparisons

- 11.9: Summary

groups

- 11.1: When to Use the Independent T-test
- 11.4: T-test Assumptions

Guide

- Front Matter
- 0: Introduction

guidelines for sample sizes

- 7: Sampling, Standard Error, Central Limit Theorem

H

harmreduction

- 0.3: Evidence-Based Practice

health

1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift

health care

- 1.3: Variations Lead to Numbers

histogram

- 4.5: Evaluating the Quality of Frequency Distributions

homework

- 1.11: Discussion Questions

hypotheses

- 11.7: Reading Articles That Involve a T-test
- 12.1: When to Use the ANOVA
- 12.5: Reading Articles – ANOVA

hypothesis

- 8: Probability, Hypothesis Testing, Type I and Type II Errors
- 8.3: Hypotheses for Statistical Inference
- 8.6: Testing a Hypothesis
- 11.2: Design Set Up

Hypothesis Testing

- 8.4: How to Determine if a Result is Significant
- 8.11: Discussion Questions
- 9.3: Statistical Power
- 9.6: Discussion Questions
- 11: T-tests
- 11.5: Reading T-tests
- 12: Analysis of Variance – ANOVA
- 12.3: - Interpreting the F-test and Post-Hoc Comparisons
- 12.7: Discussion Questions

I

IDF3

- Front Matter

importance of statistics

- 6.5: Discussion Questions

inclusivity

- 3.3: Categorical Variables- Variables That Vary by Type

independent

- 3: Variable Types
- 3.6: Making Decisions About the Variables in Your Research Study

independent observations

- 4.4: Continuous Distributions

independent samples

- 11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology (EPPP) Licensing Exam

independent variable

- 11.2: Design Set Up

independent variables

- 3.2: How to Organize Variation With Variables
- 3.7: Lessons Learned About Variables
- 3.8: Discussion Questions

independentvariable

- 12.1: When to Use the ANOVA
- 12.4: Commentary

inference

- 8: Probability, Hypothesis Testing, Type I and Type II Errors

- 10.1: The Meaning of Inferential

inferential

- 10: Introduction to Inferential Statistics
- 10.4: Discussion Questions

inferential statistics

- 2: What are Statistics?
- 4: Distributions
- 10.1: The Meaning of Inferential
- 11: T-tests

inferential_statistics

- 2.4: Two Broad Categories of the Process of Using Statistics

inferentialstatistics

- 12.7: Discussion Questions

intelligence

- 1.1: How to Start with Statistics

intensity

- 1.9: Seeing Numbers – What I Want You to Learn

internalizing symptoms

- 11.8: Reading T-tests from Journal Articles

interpretation

- 8.11: Discussion Questions

interval

- 3.5: Nuances in Determining if a Variable Should be a Certain Type

interval variables

- 4.4: Continuous Distributions

interventionstudy

- 12.4: Commentary

IV

- 10.2: Patterns

J

journal articles

- 11: T-tests
- 12.5: Reading Articles – ANOVA

journalarticles

- 11.10: Discussion Questions

juvenile delinquency

- 11.8: Reading T-tests from Journal Articles

juveniles

- 11.9: Summary

K

key terms

- 9: Effect Sizes and Power Analysis

L

large effect size

- 9.2: What Are Effect Sizes?

latentvariables

- 10.3: General Linear Model

learning

- 0: Introduction
- 0.7: Approach to Learning Statistics
- 1: What is a Number?
- 2: What are Statistics?
- 3: Variable Types
- 8: Probability, Hypothesis Testing, Type I and Type II Errors
- 10: Introduction to Inferential Statistics

learning objectives

- 1.11: Discussion Questions
- 6: Measures of Variability
- 9: Effect Sizes and Power Analysis

LearningPath

- Front Matter

Levene's test

- 11.4: T-test Assumptions

LibreTexts

- LibreTexts Information

licensing

- Licensing

linear

- 10: Introduction to Inferential Statistics

linearrelationships

- 10.3: General Linear Model

literature review

- 3.6: Making Decisions About the Variables in Your Research Study

M

males

- 11.1: When to Use the Independent T-test

mean

- 5: Measures of Central Tendency
- 5.1: What is Central Tendency? Why Do We Use It?
- 5.2: Three Types of Central Tendency and Why We Need Them
- 5.4: Discussion Questions
- 6.1: What is Variance?
- 6.2: How Do You Determine Variance?
- 7.1: The Purpose of Sampling
- 7.7: Discussion Questions

mean comparison

- 11.3: Commentary

meancentral tendency

- 5.3: The Mean, Median, and Mode in Normal and Skewed Distributions

meandifferences

- 11.9: Summary

meaningful results

- 9: Effect Sizes and Power Analysis

meaningfulresults

- 9.1: Significant (P Value) Versus Meaningful (Effect Size) Results

means comparison

- 11: T-tests

meanscores

- 11.1: When to Use the Independent T-test

Measurement

- 3.1: Variables - Not Always Obvious
- 3.4: Continuous Variables- Variables That Vary by Level

measurement errors

- 6.3: Partitioning of the Variance

measurement tactics

- 3.8: Discussion Questions

measures of dispersion

- 5.1: What is Central Tendency? Why Do We Use It?

measures_of_central_tendency

- 2.4: Two Broad Categories of the Process of Using Statistics

measuring variance

- 6.5: Discussion Questions

median

- 5: Measures of Central Tendency
- 5.2: Three Types of Central Tendency and Why We Need Them
- 5.4: Discussion Questions

mediacentral tendency

- 5.3: The Mean, Median, and Mode in Normal and Skewed Distributions

medium effect size

- 9.2: What Are Effect Sizes?

mental health

- 3.1: Variables - Not Always Obvious
- 7.1: The Purpose of Sampling
- 8.2: Statistical Decision - Is a Result Significant or Not Significant?
- 11.8: Reading T-tests from Journal Articles
- 12.5: Reading Articles – ANOVA
- 12.6: Summary

mental health outcomes

- 4.2.3: Evaluating the Plot

mentalhealth

- 9.1: Significant (P Value) Versus Meaningful (Effect Size) Results

message

- Front Matter

MindTouch

- Front Matter

mode

- 5: Measures of Central Tendency
- 5.2: Three Types of Central Tendency and Why We Need Them
- 5.4: Discussion Questions

modecentral tendency

- 5.3: The Mean, Median, and Mode in Normal and Skewed Distributions

model

- 10: Introduction to Inferential Statistics

models

- 10.4: Discussion Questions

multiple comparisons

- 12.3: - Interpreting the F-test and Post-Hoc Comparisons

N

negative skew

- 4.7: Skewed Distribution- The Opposite of Normal Distribution

negative_skew

- 5: Measures of Central Tendency

nominal

- 11.2: Design Set Up

nominal variables

- 4.2: Categorical Distributions

non_parametric_tests

- 4.6: Normal Distribution for Continuous Variables

normal

- 8: Probability, Hypothesis Testing, Type I and Type II Errors

normal distribution

- 4: Distributions
- 4.4: Continuous Distributions
- 4.5: Evaluating the Quality of Frequency Distributions
- 6.2: How Do You Determine Variance?
- 7.1: The Purpose of Sampling
- 8.4: How to Determine if a Result is Significant
- 8.7: How Do We Get the P Value for the Statistical Test?
- 8.8: Confidence Intervals
- 11.4: T-test Assumptions

normal_distribution

- 4.6: Normal Distribution for Continuous Variables

normaldistribution

- 4.8: Evaluating a Continuous Distribution
- 5.3: The Mean, Median, and Mode in Normal and Skewed Distributions
- 7.5: Central Limit Theorem (CLT) and Its Implications

normative population mean

- 11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology (EPPP) Licensing Exam

null hypothesis

- 8.8: Confidence Intervals
- 11.2: Design Set Up

nullHypothesis

- 8.3: Hypotheses for Statistical Inference
- 8.6: Testing a Hypothesis
- 8.10: The Definition of Type I and Type II Errors
- 12.1: When to Use the ANOVA

numbers

- 1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift
- 1.9: Seeing Numbers – What I Want You to Learn

O

obesity

- 12.5: Reading Articles – ANOVA

objectives

- 1: What is a Number?
- 3: Variable Types
- 8: Probability, Hypothesis Testing, Type I and Type II Errors
- 10: Introduction to Inferential Statistics

observation

- 1.2: A "Geek Out" Explanation of Variation
- 1.9: Seeing Numbers – What I Want You to Learn
- 1.10: Connection Between Understanding How Things Vary and What You are Trying to Do
- 3.1: Variables - Not Always Obvious

observations

- 1: What is a Number?
- 1.7: Statistics Do Something with Numbers
- 5.1: What is Central Tendency? Why Do We Use It?

OER

- LibreTexts Information

open education

- LibreTexts Information

operational definitions

- 3.7: Lessons Learned About Variables

ordinal

- 3.5: Nuances in Determining if a Variable Should be a Certain Type

ordinal variables

- 4.4: Continuous Distributions

outcome

- 3.2: How to Organize Variation With Variables

outcomes

- 2.3: Statistics as Estimates, or Making Guesses
- 10.3: General Linear Model

outlier

- 5.4: Discussion Questions

outliers

- 4.6: Normal Distribution for Continuous Variables
- 4.8: Evaluating a Continuous Distribution
- 5: Measures of Central Tendency
- 5.2: Three Types of Central Tendency and Why We Need Them
- 5.3: The Mean, Median, and Mode in Normal and Skewed Distributions

Overview

- Front Matter

P

p value

- 8.8: Confidence Intervals
- 9.4: How to Conduct a Power Analysis
- 9.6: Discussion Questions

p values

- 8.11: Discussion Questions

pain

- 1.10: Connection Between Understanding How Things Vary and What You are Trying to Do

parametric_tests

- 4.6: Normal Distribution for Continuous Variables

parsimony

- 1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift

partitioning

- 12.2: The Concept Behind Analysis of Variance

partitioning variance

- 3.8: Discussion Questions

patterns

- 1.3: Variations Lead to Numbers
- 2.2: Statistics is About Finding Patterns
- 10: Introduction to Inferential Statistics
- 10.1: The Meaning of Inferential
- 10.2: Patterns
- 10.4: Discussion Questions
- 11.10: Discussion Questions

percentages

- 6.2: How Do You Determine Variance?

percentiles

- 6.4: Standardized Values of Reporting Scores

performance

- 6.1: What is Variance?

personality

- 1.1: How to Start with Statistics

perspective

- 1.4: The Battle of Words Versus Numbers

point biserial correlation

- 11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology (EPPP) Licensing Exam

population

- 7: Sampling, Standard Error, Central Limit Theorem
- 7.1: The Purpose of Sampling
- 7.2: Sampling from the Population
- 7.5: Central Limit Theorem (CLT) and Its Implications
- 7.7: Discussion Questions

population characteristics

- 4.2.3: Evaluating the Plot

population mean

- 7.3: Getting a Good Sample - Get a Random Sample
- 7.4: Sampling Bias or Error
- 8.7: How Do We Get the P Value for the Statistical Test?

population parameter

- 2: What are Statistics?
- 2.3: Statistics as Estimates, or Making Guesses
- 8.8: Confidence Intervals

population parameters

- 8.9: Type I and Type II Errors

populationparameters

- 4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables

Populations

- 0.5: Social Justice

positive skew

- 4.7: Skewed Distribution- The Opposite of Normal Distribution

positive_skew

- 5: Measures of Central Tendency

posthoc tests

- 12.7: Discussion Questions

power

- 9.6: Discussion Questions

power analysis

- 7.6: Summary of Reasons Why You Need a Large Sample Size
- 9: Effect Sizes and Power Analysis
- 9.3: Statistical Power
- 9.4: How to Conduct a Power Analysis

practice

- 0: Introduction

practitioners

- 0.3: Evidence-Based Practice

prediction

- 1.7: Statistics Do Something with Numbers
- 1.9: Seeing Numbers – What I Want You to Learn
- 7.2: Sampling from the Population
- 8.3: Hypotheses for Statistical Inference
- 10.3: General Linear Model

predictions

- 2.2: Statistics is About Finding Patterns

probability

- 8.1: The Role of Probability
- 8.4: How to Determine if a Result is Significant

probability distribution

- 4.5: Evaluating the Quality of Frequency Distributions

professional development

- 0.4: Scientist-Practitioner Model; Scholar-Practitioner Model
- 0.7: Approach to Learning Statistics

professional pressure

- 0.2: Psychology Profession

prose

- 1.4: The Battle of Words Versus Numbers

psychiatric issues

- 11.9: Summary

psychological assessment

- 3.6: Making Decisions About the Variables in Your Research Study

psychological measurement

- 3.4: Continuous Variables- Variables That Vary by Level

psychological phenomena

- 2.3: Statistics as Estimates, or Making Guesses

psychological research

- 9.3: Statistical Power
- 12.6: Summary

psychological outcomes

- 12.4: Commentary

psychology

- 0: Introduction
- 0.1: A Perspective on Statistics
- 0.2: Psychology Profession
- 0.3: Evidence-Based Practice
- 0.4: Scientist-Practitioner Model; Scholar-Practitioner Model
- 0.5: Social Justice
- 0.6: Description of Intended Audience
- 0.7: Approach to Learning Statistics
- 1: What is a Number?
- 1.4: The Battle of Words Versus Numbers
- 3.1: Variables - Not Always Obvious
- 6.3: Partitioning of the Variance
- 8.2: Statistical Decision - Is a Result Significant or Not Significant?
- 9.2: What Are Effect Sizes?

psychology statistics

- 11: T-tests

psychometrics

- 6.4: Standardized Values of Reporting Scores

PTSD

- 4.2: Categorical Distributions
- 7.3: Getting a Good Sample - Get a Random Sample

pvalue

- 8.5: Interpreting a Statistical Test Result
- 8.6: Testing a Hypothesis
- 8.10: The Definition of Type I and Type II Errors
- 9.1: Significant (P Value) Versus Meaningful (Effect Size) Results

Q

qualitative

- 3.5: Nuances in Determining if a Variable Should be a Certain Type

qualitative analysis

- 0.1: A Perspective on Statistics
- 0.2: Psychology Profession

quantify

- 1.11: Discussion Questions

quantifying

- 1: What is a Number?

quantitative

- 1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words
- 3.5: Nuances in Determining if a Variable Should be a Certain Type

quantitative analysis

- 0.1: A Perspective on Statistics
- 0.2: Psychology Profession

quantitative research

- 1.10: Connection Between Understanding How Things Vary and What You are Trying to Do

questions

- 1.11: Discussion Questions

R

race

- 1.2: A "Geek Out" Explanation of Variation
- 3.3: Categorical Variables- Variables That Vary by Type

racism

- 1.1: How to Start with Statistics
- 1.10: Connection Between Understanding How Things Vary and What You are Trying to Do
- 1.11: Discussion Questions

random

- 10.1: The Meaning of Inferential

random effects

- 2.1: Statistics Involves Comparison and Context
- 6.3: Partitioning of the Variance

random sampling

- 7.3: Getting a Good Sample - Get a Random Sample
- 7.4: Sampling Bias or Error

range

- 6: Measures of Variability
- 6.1: What is Variance?

ratio

- 3.5: Nuances in Determining if a Variable Should be a Certain Type

ratio variables

- 4.4: Continuous Distributions

ratios

- 2.1: Statistics Involves Comparison and Context

reading

- 12.5: Reading Articles – ANOVA

recruitment strategy

- 7.3: Getting a Good Sample - Get a Random Sample

recruitment procedure

- 4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables

regression

- 3.6: Making Decisions About the Variables in Your Research Study
- 3.7: Lessons Learned About Variables
- 4: Distributions
- 9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together

regression analysis

- 4.2: Categorical Distributions

relationship

- 8.9: Type I and Type II Errors

relationships

- 2.2: Statistics is About Finding Patterns
- 10.2: Patterns

representation

- 7.3: Getting a Good Sample - Get a Random Sample

representative

- 7.7: Discussion Questions

research

- 0.4: Scientist-Practitioner Model; Scholar-Practitioner Model
- 0.5: Social Justice
- 0.6: Description of Intended Audience
- 0.7: Approach to Learning Statistics
- 1: What is a Number?
- 1.2: A "Geek Out" Explanation of Variation
- 1.3: Variations Lead to Numbers
- 1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift
- 3: Variable Types
- 3.2: How to Organize Variation With Variables
- 3.6: Making Decisions About the Variables in Your Research Study
- 4.1: Distributions – A Picture of Variation
- 7.7: Discussion Questions
- 8.3: Hypotheses for Statistical Inference
- 9.6: Discussion Questions
- 10: Introduction to Inferential Statistics
- 11.9: Summary
- 12.5: Reading Articles – ANOVA

research accuracy

7.4: Sampling Bias or Error

research design

1.10: Connection Between Understanding How Things Vary and What You are Trying to Do
2.3: Statistics as Estimates, or Making Guesses
6.3: Partitioning of the Variance
7.3: Getting a Good Sample - Get a Random Sample
8.1: The Role of Probability
8.2: Statistical Decision - Is a Result Significant or Not Significant?
9.2: What Are Effect Sizes?
9.3: Statistical Power
11.3: Commentary

research method

1.7: Statistics Do Something with Numbers

research methodology

3.1: Variables - Not Always Obvious
7.6: Summary of Reasons Why You Need a Large Sample Size
8.11: Discussion Questions
11.2: Design Set Up
11.5: Reading T-tests

research methods

9.4: How to Conduct a Power Analysis
11.8: Reading T-tests from Journal Articles
12.6: Summary

research question

3.3: Categorical Variables- Variables That Vary by Type
3.7: Lessons Learned About Variables
12.5: Reading Articles – ANOVA

researchdesign

9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together
12.4: Commentary

researchmethods

9.1: Significant (P Value) Versus Meaningful (Effect Size) Results
12.7: Discussion Questions

researchquestion

4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables
11.7: Reading Articles That Involve a T-test

response

1.11: Discussion Questions

results

8.3: Hypotheses for Statistical Inference
10.1: The Meaning of Inferential

results interpretation

11.5: Reading T-tests

robust test

11.4: T-test Assumptions

S

sample

7.1: The Purpose of Sampling
7.5: Central Limit Theorem (CLT) and Its Implications
7.7: Discussion Questions
8.9: Type I and Type II Errors

sample bias

7: Sampling, Standard Error, Central Limit Theorem

sample mean

7.4: Sampling Bias or Error

sample size

4.2: Categorical Distributions
4.2.3: Evaluating the Plot
4.8: Evaluating a Continuous Distribution
7.2: Sampling from the Population
7.4: Sampling Bias or Error
7.6: Summary of Reasons Why You Need a Large Sample Size
7.7: Discussion Questions
8.10: The Definition of Type I and Type II Errors
9.2: What Are Effect Sizes?
9.3: Statistical Power
9.4: How to Conduct a Power Analysis
9.6: Discussion Questions
11.4: T-test Assumptions
11.8: Reading T-tests from Journal Articles

sample sizes

7: Sampling, Standard Error, Central Limit Theorem

sample statistic

2: What are Statistics?

sample statistics

2.3: Statistics as Estimates, or Making Guesses
8.9: Type I and Type II Errors

sample validity

7.3: Getting a Good Sample - Get a Random Sample

sampledemographics

4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables

samplesize

9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together

samplesrepresentation

4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables

sampling

7: Sampling, Standard Error, Central Limit Theorem
7.2: Sampling from the Population
7.3: Getting a Good Sample - Get a Random Sample
7.4: Sampling Bias or Error
8.8: Confidence Intervals

sampling bias

7.4: Sampling Bias or Error
7.6: Summary of Reasons Why You Need a Large Sample Size

sampling distribution

7.4: Sampling Bias or Error

samplingDistribution

7.5: Central Limit Theorem (CLT) and Its Implications

scaling variables

3.8: Discussion Questions

scientific method

9.6: Discussion Questions

scientificmethod

11.7: Reading Articles That Involve a T-test

score distribution

4.5: Evaluating the Quality of Frequency Distributions

settings

Front Matter

Sidak

12.3: - Interpreting the F-test and Post-Hoc Comparisons

significance

4.6: Normal Distribution for Continuous Variables
6.2: How Do You Determine Variance?
8.1: The Role of Probability
8.2: Statistical Decision - Is a Result Significant or Not Significant?
8.4: How to Determine if a Result is Significant
8.5: Interpreting a Statistical Test Result
8.6: Testing a Hypothesis
8.8: Confidence Intervals
8.9: Type I and Type II Errors
8.11: Discussion Questions
9.1: Significant (P Value) Versus Meaningful (Effect Size) Results
9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together
11.4: T-test Assumptions
11.5: Reading T-tests
12.1: When to Use the ANOVA
12.3: - Interpreting the F-test and Post-Hoc Comparisons

significance level

9.4: How to Conduct a Power Analysis

significance test

8.9: Type I and Type II Errors

significance testing

8.1: The Role of Probability

significancellevel

8.6: Testing a Hypothesis

significant results

6.1: What is Variance?
8.2: Statistical Decision - Is a Result Significant or Not Significant?
9: Effect Sizes and Power Analysis

simplicity

1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift

skewed distribution

4.7: Skewed Distribution- The Opposite of Normal Distribution
4.9: Discussion Questions
5.3: The Mean, Median, and Mode in Normal and Skewed Distributions

skewed_distribution

5.4: Discussion Questions

skewed_distributions

5: Measures of Central Tendency

skewness

4.8: Evaluating a Continuous Distribution

small effect size

9.2: What Are Effect Sizes?

social justice

0: Introduction

social science

9.2: What Are Effect Sizes?

social support

10.2: Patterns

socialjustice

0.5: Social Justice

spread

6.1: What is Variance?

SPSS

9.4: How to Conduct a Power Analysis

standard deviation

- 6: Measures of Variability
- 6.1: What is Variance?
- 6.4: Standardized Values of Reporting Scores
- 6.5: Discussion Questions
- 7.7: Discussion Questions
- 8.7: How Do We Get the P Value for the Statistical Test?
- 8.8: Confidence Intervals

standard error

- 7.1: The Purpose of Sampling
- 7.4: Sampling Bias or Error
- 7.7: Discussion Questions

standard error of the mean

- 7: Sampling, Standard Error, Central Limit Theorem

standard_deviation

- 2.4: Two Broad Categories of the Process of Using Statistics

standarddeviation

- 4.8: Evaluating a Continuous Distribution

standarderror

- 7.5: Central Limit Theorem (CLT) and Its Implications

standardization

- 1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words

statistic

- 6.1: What is Variance?

statistical

- 3: Variable Types
- 8: Probability, Hypothesis Testing, Type I and Type II Errors

statistical analyses

- 6.2: How Do You Determine Variance?

statistical analysis

- 3.1: Variables - Not Always Obvious
- 3.2: How to Organize Variation With Variables
- 3.3: Categorical Variables- Variables That Vary by Type
- 3.4: Continuous Variables- Variables That Vary by Level
- 3.7: Lessons Learned About Variables
- 3.8: Discussion Questions
- 4: Distributions
- 4.2: Categorical Distributions
- 4.2.3: Evaluating the Plot
- 4.5: Evaluating the Quality of Frequency Distributions
- 5.1: What is Central Tendency? Why Do We Use It?
- 8.1: The Role of Probability
- 8.11: Discussion Questions
- 11: T-tests
- 11.5: Reading T-tests
- 11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology (EPPP)
- Licensing Exam
- 12.6: Summary

statistical consultant

- 0.1: A Perspective on Statistics

statistical error

- 7.4: Sampling Bias or Error

statistical inference

- 8.7: How Do We Get the P Value for the Statistical Test?

statistical methods

- 2.3: Statistics as Estimates, or Making Guesses

statistical power

- 9: Effect Sizes and Power Analysis

statistical significance

- 2.1: Statistics Involves Comparison and Context
- 6.3: Partitioning of the Variance
- 6.4: Standardized Values of Reporting Scores
- 7.6: Summary of Reasons Why You Need a Large Sample Size
- 8.4: How to Determine if a Result is Significant
- 9.3: Statistical Power

statistical test

- 3.6: Making Decisions About the Variables in Your Research Study
- 8.4: How to Determine if a Result is Significant
- 8.7: How Do We Get the P Value for the Statistical Test?
- 8.8: Confidence Intervals
- 9.4: How to Conduct a Power Analysis

statistical tests

- 5.2: Three Types of Central Tendency and Why We Need Them
- 10.2: Patterns
- 11.3: Commentary
- 11.4: T-test Assumptions

statistical validity

- 0.4: Scientist-Practitioner Model; Scholar-Practitioner Model

statistical_analysis

- 2.4: Two Broad Categories of the Process of Using Statistics
- 5.4: Discussion Questions

statistical_tests

- 4.6: Normal Distribution for Continuous Variables

statisticalanalysis

- 4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables
- 8.3: Hypotheses for Statistical Inference
- 9.1: Significant (P Value) Versus Meaningful (Effect Size) Results
- 11.9: Summary
- 12.4: Commentary
- 12.7: Discussion Questions

statisticalinference

- 8.5: Interpreting a Statistical Test Result

statisticalmodels

- 10.3: General Linear Model

statisticalpower

- 9.3: Statistical Power

statisticalresults

- 11.7: Reading Articles That Involve a T-test

statisticalsignificance

- 8.10: The Definition of Type I and Type II Errors

statisticalTest

- 8.5: Interpreting a Statistical Test Result
- 9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together
- 12.2: The Concept Behind Analysis of Variance

statisticaltests

- 8.6: Testing a Hypothesis
- 10.3: General Linear Model
- 11.7: Reading Articles That Involve a T-test

statistics

- 0: Introduction
- 0.1: A Perspective on Statistics
- 0.2: Psychology Profession
- 0.3: Evidence-Based Practice
- 0.4: Scientist-Practitioner Model; Scholar-Practitioner Model
- 0.5: Social Justice
- 0.6: Description of Intended Audience
- 1: What is a Number?
- 1.1: How to Start with Statistics
- 1.2: A "Geek Out" Explanation of Variation
- 1.3: Variations Lead to Numbers
- 1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words
- 1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift
- 1.7: Statistics Do Something with Numbers
- 1.9: Seeing Numbers – What I Want You to Learn
- 1.10: Connection Between Understanding How Things Vary and What You are Trying to Do
- 2: What are Statistics?
- 2.1: Statistics Involves Comparison and Context
- 2.2: Statistics is About Finding Patterns
- 2.3: Statistics as Estimates, or Making Guesses
- 2.4: Two Broad Categories of the Process of Using Statistics
- 3.1: Variables - Not Always Obvious
- 4: Distributions
- 4.6: Normal Distribution for Continuous Variables
- 4.7: Skewed Distribution- The Opposite of Normal Distribution
- 4.8: Evaluating a Continuous Distribution
- 4.9: Discussion Questions
- 5: Measures of Central Tendency
- 5.2: Three Types of Central Tendency and Why We Need Them
- 5.3: The Mean, Median, and Mode in Normal and Skewed Distributions
- 6.2: How Do You Determine Variance?
- 7.1: The Purpose of Sampling
- 7.2: Sampling from the Population
- 7.6: Summary of Reasons Why You Need a Large Sample Size
- 7.7: Discussion Questions
- 8: Probability, Hypothesis Testing, Type I and Type II Errors
- 8.1: The Role of Probability
- 8.7: How Do We Get the P Value for the Statistical Test?
- 8.9: Type I and Type II Errors
- 9.2: What Are Effect Sizes?
- 9.6: Discussion Questions
- 10: Introduction to Inferential Statistics
- 10.4: Discussion Questions
- 11.1: When to Use the Independent T-test
- 11.3: Commentary
- 11.5: Reading T-tests
- 11.8: Reading T-tests from Journal Articles
- 12: Analysis of Variance – ANOVA
- 12.2: The Concept Behind Analysis of Variance
- 12.5: Reading Articles – ANOVA
- 12.6: Summary
- 12.7: Discussion Questions

stigma

- 0.5: Social Justice

student discipline

- 4.7: Skewed Distribution- The Opposite of Normal Distribution

students

- 0.6: Description of Intended Audience

study

- 1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words
- 3: Variable Types
- 8.3: Hypotheses for Statistical Inference

studydesign

11.7: Reading Articles That Involve a T-test

studyssetting

4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables

substance abuse

3.2: How to Organize Variation With Variables

substances

11.9: Summary

sum of squares

6.2: How Do You Determine Variance?

summarization

1.7: Statistics Do Something with Numbers

summary statistics

5.1: What is Central Tendency? Why Do We Use It?
5.2: Three Types of Central Tendency and Why We Need Them

systematic errors

6.3: Partitioning of the Variance

T

t_tests

2.4: Two Broad Categories of the Process of Using Statistics

teaching methods

0.7: Approach to Learning Statistics

technology in education

LibreTexts Information

temperature

1.7: Statistics Do Something with Numbers

theoreticalProposition

7.5: Central Limit Theorem (CLT) and Its Implications

training

0.6: Description of Intended Audience

trauma

0.3: Evidence-Based Practice
1.4: The Battle of Words Versus Numbers
11.9: Summary

treatment

2.1: Statistics Involves Comparison and Context
3.2: How to Organize Variation With Variables
8.6: Testing a Hypothesis
11.1: When to Use the Independent T-test
11.3: Commentary
12.1: When to Use the ANOVA
12.4: Commentary

treatment approach

2.2: Statistics is About Finding Patterns

Treatment Effect

8.2: Statistical Decision - Is a Result Significant or Not Significant?
8.9: Type I and Type II Errors

treatment effectiveness

8.1: The Role of Probability

treatment group

8.1: The Role of Probability

treatment groups

6.4: Standardized Values of Reporting Scores

treatment outcomes

12.6: Summary

treatmentoutcomes

9.1: Significant (P Value) Versus Meaningful (Effect Size) Results

treatmentplanning

11.7: Reading Articles That Involve a T-test

trends

7.2: Sampling from the Population

true variance

6: Measures of Variability
6.3: Partitioning of the Variance

trueVariance

12.2: The Concept Behind Analysis of Variance

ttest

8.5: Interpreting a Statistical Test Result
8.6: Testing a Hypothesis
9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together
11.1: When to Use the Independent T-test
11.7: Reading Articles That Involve a T-test
11.10: Discussion Questions
12.7: Discussion Questions

ttests

11.9: Summary

Tukey

12.3: - Interpreting the F-test and Post-Hoc Comparisons

Type I error

7.6: Summary of Reasons Why You Need a Large Sample Size
8.4: How to Determine if a Result is Significant
8.7: How Do We Get the P Value for the Statistical Test?
8.9: Type I and Type II Errors
8.11: Discussion Questions
12.3: - Interpreting the F-test and Post-Hoc Comparisons

Type II error

7.6: Summary of Reasons Why You Need a Large Sample Size
8.4: How to Determine if a Result is Significant
8.7: How Do We Get the P Value for the Statistical Test?
8.11: Discussion Questions
9.3: Statistical Power

typeIerror

8.10: The Definition of Type I and Type II Errors

typeIIerror

8.10: The Definition of Type I and Type II Errors

types

1.2: A "Geek Out" Explanation of Variation

types of variance

6.5: Discussion Questions

U

understanding

1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift
1.9: Seeing Numbers – What I Want You to Learn

url

Licensing

V

validity

1: What is a Number?
1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words
4.6: Normal Distribution for Continuous Variables

validresults

9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together

variability

1.2: A "Geek Out" Explanation of Variation

variable sorting

3.8: Discussion Questions

variable types

3.8: Discussion Questions

variables

1.7: Statistics Do Something with Numbers
2: What are Statistics?
2.2: Statistics is About Finding Patterns
2.3: Statistics as Estimates, or Making Guesses
3: Variable Types
3.3: Categorical Variables- Variables That Vary by Type
3.4: Continuous Variables- Variables That Vary by Level
3.5: Nuances in Determining if a Variable Should be a Certain Type
3.6: Making Decisions About the Variables in Your Research Study
3.7: Lessons Learned About Variables
4.1: Distributions – A Picture of Variation
8.9: Type I and Type II Errors
10.3: General Linear Model
10.4: Discussion Questions

variance

3.1: Variables - Not Always Obvious
6: Measures of Variability
6.1: What is Variance?
6.2: How Do You Determine Variance?
6.3: Partitioning of the Variance
6.4: Standardized Values of Reporting Scores
6.5: Discussion Questions
11.5: Reading T-tests
12: Analysis of Variance – ANOVA

varianceComparison

12.2: The Concept Behind Analysis of Variance

variation

1.1: How to Start with Statistics
1.3: Variations Lead to Numbers
1.4: The Battle of Words Versus Numbers
1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words
1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift
1.7: Statistics Do Something with Numbers
1.9: Seeing Numbers – What I Want You to Learn
1.10: Connection Between Understanding How Things Vary and What You are Trying to Do
2: What are Statistics?
2.1: Statistics Involves Comparison and Context
3.1: Variables - Not Always Obvious
4.1: Distributions – A Picture of Variation
4.6: Normal Distribution for Continuous Variables
5.1: What is Central Tendency? Why Do We Use It?
7.1: The Purpose of Sampling
7.2: Sampling from the Population
8.1: The Role of Probability

visual examination

4.1: Distributions – A Picture of Variation

visualization

11.1: When to Use the Independent T-test

W

Warning

Front Matter

withinGroups

12.2: The Concept Behind Analysis of Variance

Y

youth

11.8: Reading T-tests from Journal Articles

Detailed Licensing

Overview

Title: Graduate-Level Statistics in Psychology

Webpages: 119

All licenses found:

- [CC BY 4.0](#): 99.2% (118 pages)
- [Undeclared](#): 0.8% (1 page)

By Page

- Graduate-Level Statistics in Psychology - [CC BY 4.0](#)
 - Front Matter - [CC BY 4.0](#)
 - [TitlePage](#) - [CC BY 4.0](#)
 - [Acknowledgements](#) - [CC BY 4.0](#)
 - [LibreTexts Information](#) - [CC BY 4.0](#)
 - [Licensing](#) - [CC BY 4.0](#)
 - [Table of Contents](#) - [Undeclared](#)
 - 0: Introduction - [CC BY 4.0](#)
 - [0.1: A Perspective on Statistics](#) - [CC BY 4.0](#)
 - [0.2: Psychology Profession](#) - [CC BY 4.0](#)
 - [0.3: Evidence-Based Practice](#) - [CC BY 4.0](#)
 - [0.4: Scientist-Practitioner Model; Scholar-Practitioner Model](#) - [CC BY 4.0](#)
 - [0.5: Social Justice](#) - [CC BY 4.0](#)
 - [0.6: Description of Intended Audience](#) - [CC BY 4.0](#)
 - [0.7: Approach to Learning Statistics](#) - [CC BY 4.0](#)
 - 1: What is a Number? - [CC BY 4.0](#)
 - [1.1: How to Start with Statistics](#) - [CC BY 4.0](#)
 - [1.2: A "Geek Out" Explanation of Variation](#) - [CC BY 4.0](#)
 - [1.3: Variations Lead to Numbers](#) - [CC BY 4.0](#)
 - [1.4: The Battle of Words Versus Numbers](#) - [CC BY 4.0](#)
 - [1.5: "Quantifying" a Construct, or Why Numbers Are Better Than Words](#) - [CC BY 4.0](#)
 - [1.6: Now You Love Numbers as Much as You Love the White Sox and Taylor Swift](#) - [CC BY 4.0](#)
 - [1.7: Statistics Do Something with Numbers](#) - [CC BY 4.0](#)
 - [1.8: Numbers Don't Lie, but Numbers Don't Tell the Full Story](#) - [CC BY 4.0](#)
 - [1.9: Seeing Numbers – What I Want You to Learn](#) - [CC BY 4.0](#)
 - [1.10: Connection Between Understanding How Things Vary and What You are Trying to Do](#) - [CC BY 4.0](#)
 - [1.11: Discussion Questions](#) - [CC BY 4.0](#)
 - 2: What are Statistics? - [CC BY 4.0](#)
 - [2.1: Statistics Involves Comparison and Context](#) - [CC BY 4.0](#)
 - [2.2: Statistics is About Finding Patterns](#) - [CC BY 4.0](#)
 - [2.3: Statistics as Estimates, or Making Guesses](#) - [CC BY 4.0](#)
 - [2.4: Two Broad Categories of the Process of Using Statistics](#) - [CC BY 4.0](#)
 - [2.5: Discussion Questions](#) - [CC BY 4.0](#)
 - 3: Variable Types - [CC BY 4.0](#)
 - [3.1: Variables - Not Always Obvious](#) - [CC BY 4.0](#)
 - [3.2: How to Organize Variation With Variables](#) - [CC BY 4.0](#)
 - [3.3: Categorical Variables- Variables That Vary by Type](#) - [CC BY 4.0](#)
 - [3.4: Continuous Variables- Variables That Vary by Level](#) - [CC BY 4.0](#)
 - [3.5: Nuances in Determining if a Variable Should be a Certain Type](#) - [CC BY 4.0](#)
 - [3.6: Making Decisions About the Variables in Your Research Study](#) - [CC BY 4.0](#)
 - [3.7: Lessons Learned About Variables](#) - [CC BY 4.0](#)
 - [3.8: Discussion Questions](#) - [CC BY 4.0](#)
 - 4: Distributions - [CC BY 4.0](#)
 - [4.1: Distributions – A Picture of Variation](#) - [CC BY 4.0](#)
 - [4.2: Categorical Distributions](#) - [CC BY 4.0](#)
 - [4.2.3: Evaluating the Plot](#) - [CC BY 4.0](#)
 - [4.3: Reviewing Journal Articles for Frequency Distributions for Categorical Variables](#) - [CC BY 4.0](#)
 - [4.4: Continuous Distributions](#) - [CC BY 4.0](#)
 - [4.5: Evaluating the Quality of Frequency Distributions](#) - [CC BY 4.0](#)
 - [4.6: Normal Distribution for Continuous Variables](#) - [CC BY 4.0](#)

- 4.7: Skewed Distribution- The Opposite of Normal Distribution - CC BY 4.0
- 4.8: Evaluating a Continuous Distribution - CC BY 4.0
- 4.9: Discussion Questions - CC BY 4.0
- 5: Measures of Central Tendency - CC BY 4.0
 - 5.1: What is Central Tendency? Why Do We Use It? - CC BY 4.0
 - 5.2: Three Types of Central Tendency and Why We Need Them - CC BY 4.0
 - 5.3: The Mean, Median, and Mode in Normal and Skewed Distributions - CC BY 4.0
 - 5.4: Discussion Questions - CC BY 4.0
- 6: Measures of Variability - CC BY 4.0
 - 6.1: What is Variance? - CC BY 4.0
 - 6.2: How Do You Determine Variance? - CC BY 4.0
 - 6.3: Partitioning of the Variance - CC BY 4.0
 - 6.4: Standardized Values of Reporting Scores - CC BY 4.0
 - 6.5: Discussion Questions - CC BY 4.0
- 7: Sampling, Standard Error, Central Limit Theorem - CC BY 4.0
 - 7.1: The Purpose of Sampling - CC BY 4.0
 - 7.2: Sampling from the Population - CC BY 4.0
 - 7.3: Getting a Good Sample - Get a Random Sample - CC BY 4.0
 - 7.4: Sampling Bias or Error - CC BY 4.0
 - 7.5: Central Limit Theorem (CLT) and Its Implications - CC BY 4.0
 - 7.6: Summary of Reasons Why You Need a Large Sample Size - CC BY 4.0
 - 7.7: Discussion Questions - CC BY 4.0
- 8: Probability, Hypothesis Testing, Type I and Type II Errors - CC BY 4.0
 - 8.1: The Role of Probability - CC BY 4.0
 - 8.2: Statistical Decision - Is a Result Significant or Not Significant? - CC BY 4.0
 - 8.3: Hypotheses for Statistical Inference - CC BY 4.0
 - 8.4: How to Determine if a Result is Significant - CC BY 4.0
 - 8.5: Interpreting a Statistical Test Result - CC BY 4.0
 - 8.6: Testing a Hypothesis - CC BY 4.0
 - 8.7: How Do We Get the P Value for the Statistical Test? - CC BY 4.0
 - 8.8: Confidence Intervals - CC BY 4.0
 - 8.9: Type I and Type II Errors - CC BY 4.0
- 8.10: The Definition of Type I and Type II Errors - CC BY 4.0
- 8.11: Discussion Questions - CC BY 4.0
- 9: Effect Sizes and Power Analysis - CC BY 4.0
 - 9.1: Significant (P Value) Versus Meaningful (Effect Size) Results - CC BY 4.0
 - 9.2: What Are Effect Sizes? - CC BY 4.0
 - 9.3: Statistical Power - CC BY 4.0
 - 9.4: How to Conduct a Power Analysis - CC BY 4.0
 - 9.5: Pause and Process Check, Pulling P Values and Effect Sizes Together - CC BY 4.0
 - 9.6: Discussion Questions - CC BY 4.0
- 10: Introduction to Inferential Statistics - CC BY 4.0
 - 10.1: The Meaning of Inferential - CC BY 4.0
 - 10.2: Patterns - CC BY 4.0
 - 10.3: General Linear Model - CC BY 4.0
 - 10.4: Discussion Questions - CC BY 4.0
- 11: T-tests - CC BY 4.0
 - 11.1: When to Use the Independent T-test - CC BY 4.0
 - 11.2: Design Set Up - CC BY 4.0
 - 11.3: Commentary - CC BY 4.0
 - 11.4: T-test Assumptions - CC BY 4.0
 - 11.5: Reading T-tests - CC BY 4.0
 - 11.6: T-Tests You Need to Know for the Examination for Professional Practice in Psychology (EPPP) Licensing Exam - CC BY 4.0
 - 11.7: Reading Articles That Involve a T-test - CC BY 4.0
 - 11.8: Reading T-tests from Journal Articles - CC BY 4.0
 - 11.9: Summary - CC BY 4.0
 - 11.10: Discussion Questions - CC BY 4.0
- 12: Analysis of Variance – ANOVA - CC BY 4.0
 - 12.1: When to Use the ANOVA - CC BY 4.0
 - 12.2: The Concept Behind Analysis of Variance - CC BY 4.0
 - 12.3: - Interpreting the F-test and Post-Hoc Comparisons - CC BY 4.0
 - 12.4: Commentary - CC BY 4.0
 - 12.5: Reading Articles – ANOVA - CC BY 4.0
 - 12.6: Summary - CC BY 4.0
 - 12.7: Discussion Questions - CC BY 4.0
- Back Matter - CC BY 4.0
 - Index - CC BY 4.0
 - Appendices - CC BY 4.0
 - Detailed Licensing - CC BY 4.0