PSYC 330: STATISTICS FOR THE BEHAVIORAL SCIENCES WITH DR. DESOUZA

LibreTexts™

PSYC 330: Statistics for the Behavioral Sciences with Dr. DeSouza

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 adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

This text was compiled on 04/19/2026

TABLE OF CONTENTS

Licensing

Foreword

1: What are Statistics?

- 1.1: What are statistics?
- 1.2: Why do we study statistics?

2: Types of Data and How to Collect Them (by Dr. Alisa Beyer)

- 2.1: Types of Data and How to Collect Them
- 2.2: Who are your participants? Who is your population?
- 2.3: Representative Sample
- 2.4: Type of Research Designs
- 2.5: Working with Data
- 2.6: Levels of Measurement
- 2.7: What level of measurement is used for psychological variables?
- 2.8: Reliability, Validity, and Results
- 2.9: Types of Statistical Analyses
- 2.10: Mathematical Notation
- 2.E: Exercises

3: Measures of Central Tendency and Spread

- 3.1: What is Central Tendency?
- 3.2: Measures of Central Tendency
- 3.3: Spread and Variability
- 3.E: Measures of Central Tendency and Spread (Exercises)

4: Describing Data using Distributions and Graphs

- 4.1: Graphing Qualitative Variables
- 4.2: Graphing Quantitative Variables
- 4.E: Describing Data using Distributions and Graphs (Exercises)

5: Z-scores and the Standard Normal Distribution

- 5.1: Normal Distributions
- 5.2: Z-scores
- 5.3: Z-scores and the Area under the Curve
- 5.E: Z-scores and the Standard Normal Distribution (Exercises)

6: Probability

- 6.1: What is Probability
- 6.2: Probability in Graphs and Distributions
- 6.3: The Bigger Picture
- 6.E: Probability (Exercises)

7: Sampling Distributions

- 7.1: People, Samples, and Populations
- 7.2: The Sampling Distribution of Sample Means
- 7.3: Sampling Distribution, Probability and Inference
- 7.E: Sampling Distributions (Exercises)

8: Introduction to Hypothesis Testing

- 8.1: Logic and Purpose of Hypothesis Testing
- 8.2: The Probability Value
- 8.3: The Null Hypothesis
- 8.4: The Alternative Hypothesis
- 8.5: Critical values, p-values, and significance level
- 8.6: Steps of the Hypothesis Testing Process
- 8.7: Movie Popcorn
- 8.8: Effect Size
- 8.9: Office Temperature
- 8.10: Different Significance Level
- 8.11: Other Considerations in Hypothesis Testing
- 8.E: Introduction to Hypothesis Testing (Exercises)

9: Introduction to t-tests

- 9.1: The t-statistic
- 9.2: Hypothesis Testing with t
- 9.3: Confidence Intervals
- 9.E: Introduction to t-tests (Exercises)

10: Repeated Measures

- 10.1: Change and Differences
- 10.2: Hypotheses of Change and Differences
- 10.3: Increasing Satisfaction at Work
- 10.4: Bad Press
- 10.E: Repeated Measures (Exercises)

11: Independent Samples

- 11.1: Difference of Means
- 11.2: Research Questions about Independent Means
- 11.3: Hypotheses and Decision Criteria
- 11.4: Independent Samples t-statistic
- 11.5: Standard Error and Pooled Variance
- 11.6: Movies and Mood
- 11.7: Effect Sizes and Confidence Intervals
- 11.8: Homogeneity of Variance
- 11.E: Independent Samples (Exercises)

12: Analysis of Variance

- 12.1: Observing and Interpreting Variability
- 12.2: Sources of Variance

- 12.3: ANOVA Table
- 12.4: ANOVA and Type I Error
- 12.5: Hypotheses in ANOVA
- 12.6: Scores on Job Application Tests
- 12.7: Variance Explained
- 12.8: Post Hoc Tests
- 12.9: Other ANOVA Designs
- 12.10: Analysis of Variance (Exercises)

13: Two-Factor ANOVAs (by Dr. Alisa Beyer)

- 13.1: Two-Factor ANOVAs
- 13.2: Conducting a Two Factor ANOVA
- 13.3: Graphing the Results of Factorial Experiments
- 13.E: Exercises

14: Correlations

- 14.1: Variability and Covariance
- 14.2: Visualizing Relations
- 14.3: Three Characteristics
- 14.4: Pearson's r
- 14.5: Anxiety and Depression
- 14.6: Effect Size
- 14.7: Correlation versus Causation
- 14.8: Final Considerations
- 14.E: Correlations (Exercises)

15: Chi-square

- 15.1: Categories and Frequency Tables
- 15.2: Goodness-of-Fit
- 15.3: χ^2 Statistic
- 15.4: Pineapple on Pizza
- 15.5: Contingency Tables for Two Variables
- 15.6: Test for Independence
- 15.7: College Sports
- 15.E: Chi-square (Exercises)

16: Appendix A- Statistical Tables

- 16.1: F-distribution (ANOVA distribution) table
- 16.2: z-table (aka Standard Normal Distribution Table)
- 16.3: t Distribution Table

Index

Glossary

Detailed Licensing

Detailed Licensing

Licensing

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

Foreword

A Note from Dr. Kara DeSouza:

Below you'll find the original Foreword to this book from Dr. Garrett Foster, one of its original authors. Because of the unique nature of LibreText and open source textbooks, I have been able to customize Dr. Foster's text to exactly the topics relevant to our class. I have also drawn from a second text to cover some material missing from the Foster text but which is necessary and beneficial to know for your best success in this class. That second open source textbook is by Dr. Alisa Beyer of Chandler-Gilbert Community College in Chandler, AZ. You can assume all chapters are by Dr. Foster and the other authors on the original project (listed below), unless it is specifically pointed out that the chapter is by Dr. Beyer. I am also indebted to Judy Schmitt of the University of Missouri-St. Louis for the excellent illustratations she created for the Cote, et al (2021) update of the book, which I have used to make the illustrations in this edition as accessible and readable as possible. As a professor, I am deeply grateful to all of these authors for the immense time commitment put into creating high quality and free educational content. I am also very grateful the LibreText system exists to host and allow me to customize, these textbooks just for my PSYC 330 classes at Sacramento City College.

See you in class!

-Kara

Foreword by Dr. Foster:

We are constantly bombarded by information, and finding a way to filter that information in an objective way is crucial to surviving this onslaught with your sanity intact. This is what statistics, and logic we use in it, enables us to do. Through the lens of statistics, we learn to find the signal hidden in the noise when it is there and to know when an apparent trend or pattern is really just randomness. The study of statistics involves math and relies upon calculations of numbers. But it also relies heavily on how the numbers are chosen and how the statistics are interpreted.

This work was created as part of the University of Missouri's Affordable and Open Access Educational Resources Initiative (https://www.umsystem.edu/ums/aa/oer). The contents of this work have been adapted from the following Open Access Resources: Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University. Changes to the original works were made by Dr. Garett C. Foster in the Department of Psychological Sciences to tailor the text to fit the needs of the introductory statistics course for psychology majors at the University of Missouri – St. Louis. Materials from the original sources have been combined, reorganized, and added to by the current author, and any conceptual, mathematical, or typographical errors are the responsibility of the current author.

- **Garett C. Foster**, *University of Missouri-St. Louis*Follow
- **David Lane**, *Rice University*Follow
- **David Scott**, *Rice University*
- **Mikki Hebl**, *Rice University*
- **Rudy Guerra**, *Rice University*
- **Dan Osherson**, *Rice University*
- **Heidi Zimmer**, *University of Houston, Downtown Campus*

Recommended Citations

Beyer, Alisa. *Introduction to Statistics for Psychology*. Maricopa Open Digital Press, 2021, https://open.maricopa.edu/psy230mm/.

Cote, Linda R.; Gordon, Rupa; Randell, Chrislyn E.; Schmitt, Judy; and Marvin, Helena, "Introduction to Statistics in the Psychological Sciences" (2021). Open Educational Resources Collection. 25. Available at: https://irl.umsl.edu/oer/25 - Version 08/01/2023 (Minor Revision)

Foster, Garett C.; Lane, David; Scott, David; Hebl, Mikki; Guerra, Rudy; Osherson, Dan; and Zimmer, Heidi, "An Introduction to Psychological Statistics" (2018). *Open Educational Resources Collection*. 4.
https://irl.umsl.edu/oer/4

Chapters 4 and 11 cover icons by Freepik, found on Flaticon, 2024, https://www.freepik.com/.

Chapter 9 cover icon by Iconiic, found on Flaticon, 2024.

Chapter 10 cover icon by Pause8, found on Flaticon, 2024.

Chapter 12 cover icon by Afian Rochmah Afif, found on Flaticon, 2024, https://www.flaticon.com/authors/afian-rochmah-afif.

Many illustrations throughout the chapters were created by Judy Scmitt of the University of Missouri at St. Louis, and were originally published in Cote et al, 2021.

- **About this Book** is licensed CC BY-NC-SA 4.0.

CHAPTER OVERVIEW

1: What are Statistics?

1.1: What are statistics?
1.2: Why do we study statistics?

This page titled 1: What are Statistics? is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

1.1: What are statistics?

Statistics include numerical facts and figures. For instance:

- The largest earthquake measured 9.2 on the Richter scale.
- Men are at least 10 times more likely than women to commit murder.
- One in every 8 South Africans is HIV positive.
- By the year 2020, there will be 15 people aged 65 and over for every new baby born.

The study of statistics involves math and relies upon calculations of numbers. But it also relies heavily on how the numbers are chosen and how the statistics are interpreted. For example, consider the following three scenarios and the interpretations based upon the presented statistics. You will find that the numbers may be right, but the interpretation may be wrong. Try to identify a major flaw with each interpretation before we describe it.

1. A new advertisement for Ben and Jerry's ice cream introduced in late May of last year resulted in a 30% increase in ice cream sales for the following three months. Thus, the advertisement was effective. A major flaw is that ice cream consumption generally increases in the months of June, July, and August regardless of advertisements. This effect is called a history effect and leads people to interpret outcomes as the result of one variable when another variable (in this case, one having to do with the passage of time) is actually responsible.
2. The more churches in a city, the more crime there is. Thus, churches lead to crime. A major flaw is that both increased churches and increased crime rates can be explained by larger populations. In bigger cities, there are both more churches and more crime. This problem, which we will discuss in more detail in Chapter 6, refers to the third-variable problem. Namely, a third variable can cause both situations; however, people erroneously believe that there is a causal relationship between the two primary variables rather than recognize that a third variable can cause both.
3. 75% more interracial marriages are occurring this year than 25 years ago. Thus, our society accepts interracial marriages. A major flaw is that we don't have the information that we need. What is the rate at which marriages are occurring? Suppose only 1% of marriages 25 years ago were interracial and so now 1.75% of marriages are interracial (1.75 is 75% higher than 1). But this latter number is hardly evidence suggesting the acceptability of interracial marriages. In addition, the statistic provided does not rule out the possibility that the number of interracial marriages has seen dramatic fluctuations over the years and this year is not the highest. Again, there is simply not enough information to understand fully the impact of the statistics.

As a whole, these examples show that statistics are not only facts and figures; they are something more than that. In the broadest sense, “statistics” refers to a range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data.

Statistics is the language of science and data. The ability to understand and communicate using statistics enables researchers from different labs, different languages, and different fields articulate to one another exactly what they have found in their work. It is an objective, precise, and powerful tool in science and in everyday life.

What statistics are not

Many psychology students dread the idea of taking a statistics course, and more than a few have changed majors upon learning that it is a requirement. That is because many students view statistics as a math class, which is actually not true. While many of you will not believe this or agree with it, statistics isn’t math. Although math is a central component of it, statistics is a broader way of organizing, interpreting, and communicating information in an objective manner. Indeed, great care has been taken to eliminate as much math from this course as possible (students who do not believe this are welcome to ask the professor what matrix algebra is). Statistics is a way of viewing reality as it exists around us in a way that we otherwise could not.

This page titled 1.1: What are statistics? is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri’s Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **1.1: What are statistics?** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

1.2: Why do we study statistics?

Virtually every student of the behavioral sciences takes some form of statistics class. This is because statistics is how we communicate in science. It serves as the link between a research idea and usable conclusions. Without statistics, we would be unable to interpret the massive amounts of information contained in data. Even small datasets contain hundreds – if not thousands – of numbers, each representing a specific observation we made. Without a way to organize these numbers into a more interpretable form, we would be lost, having wasted the time and money of our participants, ourselves, and the communities we serve.

Beyond its use in science, however, there is a more personal reason to study statistics. Like most people, you probably feel that it is important to "take control of your life." But what does this mean? Partly, it means being able to properly evaluate the data and claims that bombard you every day. If you cannot distinguish good from faulty reasoning, then you are vulnerable to manipulation and to decisions that are not in your best interest. Statistics provides tools that you need in order to react intelligently to information you hear or read. In this sense, statistics is one of the most important things that you can study.

To be more specific, here are some claims that we have heard on several occasions. (We are not saying that each one of these claims is true!)

- 4 out of 5 dentists recommend Dentine.
- Almost 85% of lung cancers in men and 45% in women are tobacco-related.
- Condoms are effective 94% of the time.
- People tend to be more persuasive when they look others directly in the eye and speak loudly and quickly.
- Women make 75 cents to every dollar a man makes when they work the same job.
- A surprising new study shows that eating egg whites can increase one's life span.
- People predict that it is very unlikely there will ever be another baseball player with a batting average over 400.
- There is an 80% chance that in a room full of 30 people that at least two people will share the same birthday.
- 79.48% of all statistics are made up on the spot.

All of these claims are statistical in character. We suspect that some of them sound familiar; if not, we bet that you have heard other claims like them. Notice how diverse the examples are. They come from psychology, health, law, sports, business, etc. Indeed, data and data interpretation show up in discourse from virtually every facet of contemporary life.

Statistics are often presented in an effort to add credibility to an argument or advice. You can see this by paying attention to television advertisements. Many of the numbers thrown about in this way do not represent careful statistical analysis. They can be misleading and push you into decisions that you might find cause to regret. For these reasons, learning about statistics is a long step towards taking control of your life. (It is not, of course, the only step needed for this purpose.) The purpose of this course, beyond preparing you for a career in psychology, is to help you learn statistical essentials. It will make you into an intelligent consumer of statistical claims.

You can take the first step right away. To be an intelligent consumer of statistics, your first reflex must be to question the statistics that you encounter. The British Prime Minister Benjamin Disraeli is quoted by Mark Twain as having said, "There are three kinds of lies -- lies, damned lies, and statistics." This quote reminds us why it is so important to understand statistics. So let us invite you to reform your statistical habits from now on. No longer will you blindly accept numbers or findings. Instead, you will begin to think about the numbers, their sources, and most importantly, the procedures used to generate them.

The above section puts an emphasis on defending ourselves against fraudulent claims wrapped up as statistics, but let us look at a more positive note. Just as important as detecting the deceptive use of statistics is the appreciation of the proper use of statistics. You must also learn to recognize statistical evidence that supports a stated conclusion. Statistics are all around you, sometimes used well, sometimes not. We must learn how to distinguish the two cases. In doing so, statistics will likely be the course you use most in your day to day life, even if you do not ever run a formal analysis again.

This page titled 1.2: Why do we study statistics? is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **1.2: Why do we study statistics?** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

2: Types of Data and How to Collect Them (by Dr. Alisa Beyer)

2.1: Types of Data and How to Collect Them
2.2: Who are your participants? Who is your population?
2.3: Representative Sample
2.4: Type of Research Designs
2.5: Working with Data
2.6: Levels of Measurement
2.7: What level of measurement is used for psychological variables?
2.8: Reliability, Validity, and Results
2.9: Types of Statistical Analyses
2.10: Mathematical Notation
2.E: Exercises

2: Types of Data and How to Collect Them (by Dr. Alisa Beyer) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

2.1: Types of Data and How to Collect Them

In order to use statistics, we need data to analyze. Data come in an amazingly diverse range of formats, and each type gives us a unique type of information. In virtually any form, data represent the measured value of variables. A **variable** is simply a characteristic or feature of the thing we are interested in understanding. Let's imagine we want to conduct a study to measure the stress level of students who are taking PSY 230. We will administer the survey during the first week of the course. One question we will ask is, "How stressed have you been in the last 2 weeks, on a scale of 0 to 10, with 0 being not at all stressed and 10 being as stressed as possible?"

- **Variable** is a condition or characteristic that can take on different values. In our example, the variable was stress, which can take on any value between 0 and 10. Height is a variable. Social class is a variable. One's score on a creativity test is a variable. The number of people absent from work on a given day is a variable. In psychology, we are interested in people, so we might get a group of people together and measure their levels of anxiety (a variable) or their physical health (another variable). You get the point. Pretty much anything we can count or measure can be a variable.
  - Once we have data on different variables, we can use statistics to understand if and how they are related.
- A **value** is just a number, such as 4, – 81, or 367.12. A value can also be a category (word), such as male or female, or a psychological diagnosis (major depressive disorder, post-traumatic stress disorder, schizophrenia).
  - We will learn more about values and types of data a little later in this chapter.
- Each person studied has a particular **score** that is his or her value on the variable. As we've said, your score on the stress variable might have a value of 6. Another student's score might have a value of 8.

We also need to understand the nature of our data: what they represent and where they came from. We will briefly review some keys to understanding statistical studies.

Tips to understanding statistical studies

Here are a few key considerations for evaluating studies using statistics.

1. Know the basic components of a statistical investigation.
2. Know the sample. Identify if using a representative sample.
3. Identify the sample size. Evaluate if using a large enough sample.
4. Understand and evaluate the study design.
5. Identify type of data working with.
6. Understand the statistics used.
7. Evaluate conclusions made from statistical findings.

The basic components to a statistical investigation

- **Planning the study:** Start by asking a testable research question and deciding how to collect data. For example, how long was the study period of the study? How many people were recruited for the study, how were they recruited, and from where? How old were they? What other variables were recorded about the individuals, such as smoking habits, on the comprehensive lifestyle questionnaires?
- **Examining the data:** What are appropriate ways to examine the data? What graphs are relevant, and what do they reveal? What descriptive statistics can be calculated to summarize relevant aspects of the data, and what do they reveal? What patterns do you see in the data? Are there any individual observations that deviate from the overall pattern, and what do they reveal?
- **Inferring from the data:** What are valid statistical methods for drawing inferences "beyond" the data you collected? Is a 10%–15% reduction in risk of death something that can happen just by chance?
- **Drawing conclusions:** Based on what you learned from your data, what conclusions can you draw? Who do you think these conclusions apply to? Can you draw a cause-and-effect conclusion about your treatment? (note: we are about to learn more about the study design needed for this)

Notice that the *numerical analysis* ("crunching numbers" on the computer) comprises only a small part of the overall statistical investigation. In this module, you will see how we can answer some of these questions and what questions you should be asking about any statistical investigation you read about. In the end, statistics provides us a way to give a very objective "yes" or "no" answer to the question, "is this treatment or intervention effective and, if so, how effective is it?" Nearly all statistical techniques boil down to answering these questions. Statistics is all about helping make correct and reliable decisions in our chosen field of study. But even if you never plan on conducting research or pursuing a career where you have to use statistics, the material in this course will help you in your daily life. In today's world of instant gratification, information overload, and the 24-hour news cycle, statistics are thrown at us nonstop. Soon, you will be able to determine if the person or group providing these statistics is being honest or manipulating the data to suit their ideas.

Let's learn a little bit more about what is needed to know to better understand statistics.

2.1: Types of Data and How to Collect Them is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

2.2: Who are your participants? Who is your population?

Research in psychology typically begins with a general question about a specific group (or groups) of individuals or animals. For example, a researcher might want to know how many homeless people live on the streets of Phoenix. Or a researcher might want to know how often married people have sex, as reported by partners separately. In the first example, the researcher is interested in the group of homeless people. In the second example, the researcher may study heterosexual couples and compare the group of men with the group of women. In statistics, we call the entire group that a researcher wishes to study **a population.** As you can well imagine, a population can be quite large; for example, any student enrolled in college. A researcher might be more specific, limiting the population for a study to college students who have successfully completed a statistics course and who live in the United States.

Populations can obviously vary in size from extremely large to very small, depending on how the researcher defines the population. The population being studied should always be identified by the researcher. In addition, the population can include more than people and animals. A population could be corporations, parts produced in a factory, or anything else a researcher wants to study. Because populations tend to be very large it usually is impossible for a researcher to examine every individual in the population of interest. It is typically not feasible to collect data from an entire population. Therefore, researchers typically select a smaller, more manageable group from the population and limit their studies to the individuals in the selected group. A smaller more manageable group, known as a **sample**, is used to measure populations.

The participants in the research are the **sample**, and the larger group the sample represents is the **population**. In statistical terms, a set of individuals selected from a population is called a **sample** . A sample is intended to be representative of its population, and a sample should always be identified in terms of the population from which it was selected. As with populations, samples can vary in size. For example, one study might examine a sample of only 10 autistic children, and another study might use a sample of more than 10,000 people who take specific cholesterol medication. The sample is intended to represent the population in a research study.

When describing data it is necessary to distinguish whether the data come from a population or a sample.

- If data describe a **sample** it is called a **statistic.**
- If data describe a **population** it is called a **parameter.**

If I had given a statistical attitudes survey to the class, the class would be my sample. I might be interested in all students taking a statistics class for the first time, generalizing my findings to all statistics students would be applying information from my sample to a population. While it might be convenient for me to ask my class, does my class best represent all students taking statistics? I would need to carefully consider selecting the best sample for a population or critically think about the limits for generalizing my findings to a population. While our results would be most accurate if we could study the entire population rather than a sample from it, in most research situations this is not practical. Moreover, research usually is to be able to make generalizations or predictions about events beyond your reach. Additionally, sampling is an important concept to consider with the big picture of understanding statistics.

Imagine that we wanted to see if statistics anxiety was related to procrastination. We could measure everyone's levels of statistics anxiety and procrastination and observe how strongly they were related to each other. This would, however, be prohibitively expensive. A more convenient way is to select a number of individuals randomly from the population and find the relationship between their statistics anxiety and procrastination levels. We could then generalize the findings from this sample to the population. We use statistics, more specifically inferential statistics, to help us generalize from a particular sample to the whole population. Understanding the relationship between populations and their samples is the first vital concept to grasp in this course. Remember that the research started with a general question about the population but to answer the question, a researcher studies a sample and then generalizes the results from the sample to the population.

As we move into further concepts in statistics, we will see that how you get your participants (sampling) and sample size are important. The general rule is to get a large enough sample size and have the sample be a good representation of your population.

2.2: Who are your participants? Who is your population? is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

2.3: Representative Sample

Usually, the ideal method of picking out a sample to study is called random selection or sampling. The researcher starts with a **complete list of the population** and randomly selects some of them to study. . Random sampling is considered a fair way of selecting a sample from a given population since every member is given equal opportunities of being selected. This process also helps to ensure that the sample selected is more likely to be representative of the larger population. Theoretically, the only thing that can compromise its representativeness is luck. If the sample is not representative of the population, the random variation is called **sampling error.**

> **Example #1:** You have been hired by the National Election Commission to examine how the American people feel about the fairness of the voting procedures in the U.S. Who will you ask?

It is not practical to ask every single American how he or she feels about the fairness of the voting procedures. Instead, we query a relatively small number of Americans, and draw inferences about the entire country from their responses. The Americans actually queried constitute our sample of the larger population of all Americans. A **sample** is typically a small subset of the population. In the case of voting attitudes, we would sample a few thousand Americans drawn from the hundreds of millions that make up the country. In choosing a sample, it is therefore crucial that it not over-represent one kind of citizen at the expense of others. For example, something would be wrong with our sample if it happened to be made up entirely of Florida residents. If the sample held only Floridians, it could not be used to infer the attitudes of other Americans. The same problem would arise if the sample were comprised only of Republicans. Inferences from statistics are based on the assumption that sampling is representative of the population. If the sample is not representative, then the possibility of sampling bias occurs. Sampling bias means that our conclusions apply only to our sample and are not generalizable to the full population.

> **Example #2:** We are interested in examining how many math classes have been taken on average by current graduating seniors at American colleges and universities during their four years in school.

Whereas our population in the last example included all US citizens, now it involves just the graduating seniors throughout the country. This is still a large set since there are thousands of colleges and universities, each enrolling many students. (New York University, for example, enrolls 48,000 students.) It would be prohibitively costly to examine the transcript of every college senior. We therefore take a sample of college seniors and then make inferences to the entire population based on what we find. To make the sample, we might first choose some public and private colleges and universities across the United States. Then we might sample 50 students from each of these institutions. Suppose that the average number of math classes taken by the people in our sample were 3.2. Then we might speculate that 3.2 approximates the number we would find if we had the resources to examine every senior in the entire population. But we must be careful about the possibility that our sample is non-representative of the population. Perhaps we chose an overabundance of math majors, or chose too many technical institutions that have heavy math requirements. Such bad sampling makes our sample unrepresentative of the population of all seniors. To solidify your understanding of sampling bias, consider the following example. Try to identify the population and the sample, and then reflect on whether the sample is likely to yield the information desired.

> **Example #3:** A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well. What is the sample? What is the population? Can you identify any problems with choosing the sample in the way that the teacher did?

In Example #3, the population consists of all students in the class. The sample is made up of just the 10 students sitting in the front row. The sample is not likely to be representative of the population. Those who sit in the front row tend to be more interested in the class and tend to perform higher on tests. Hence, the sample may perform at a higher level than the population.

Example #4: A coach is interested in how many cartwheels the average college freshmen at his university can do. Eight volunteers from the freshman class step forward. After observing their performance, the coach concludes that college freshmen can do an average of 16 cartwheels in a row without stopping.

In Example #4, the population is the class of all freshmen at the coach's university. The sample is composed of the 8 volunteers. The sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman; people who can't do cartwheels probably did not volunteer! In the example, we are also not told of the gender of the volunteers. Were they all women, for example? That might affect the outcome, contributing to the non-representative nature of the sample (if the school is co-ed).

Simple Random Sampling

Researchers adopt a variety of sampling strategies. The most straightforward is simple random sampling. Such sampling requires every member of the population to have an equal chance of being selected into the sample. In addition, the selection of one member must be independent of the selection of every other member. That is, picking one member from the population must not increase or decrease the probability of picking any other member (relative to the others). In this sense, we can say that simple random sampling chooses a sample by pure chance. To check your understanding of simple random sampling, consider the following example.

What is the population? What is the sample? Was the sample picked by simple random sampling? Is it biased?

Example #5: A research scientist is interested in studying the experiences of twins raised together versus those raised apart. She obtains a list of twins from the National Twin Registry, and selects two subsets of individuals for her study. First, she chooses all those in the registry whose last name begins with Z. Then she turns to all those whose last name begins with B. Because there are so many names that start with B, however, our researcher decides to incorporate only every other name into her sample. Finally, she mails out a survey and compares characteristics of twins raised apart versus together.

In Example #5, the population consists of all twins recorded in the National Twin Registry. It is important that the researcher only make statistical generalizations to the twins on this list, not to all twins in the nation or world. That is, the National Twin Registry may not be representative of all twins. Even if inferences are limited to the Registry, a number of problems affect the sampling procedure we described. For instance, choosing only twins whose last names begin with Z does not give every individual an equal chance of being selected into the sample. Moreover, such a procedure risks over-representing ethnic groups with many surnames that begin with Z. There are other reasons why choosing just the Z's may bias the sample.

Perhaps such people are more patient than average because they often find themselves at the end of the line! The same problem occurs with choosing twins whose last name begins with B. An additional problem for the B's is that the "every-other-one" procedure disallowed adjacent names on the B part of the list from being both selected. Just this defect alone means the sample was not formed through simple random sampling.

Sample size matters

Recall that the definition of a random sample is a sample in which every member of the population has an equal chance of being selected. This means that the sampling procedure rather than the results of the procedure define what it means for a sample to be random. Random samples, especially if the sample size is small, are not necessarily representative of the entire population. For example, if a random sample of 20 subjects were taken from a population with an equal number of males and females, there would be a nontrivial probability (0.06) that 70% or more of the sample would be female. Such a sample would not be representative, although it would be drawn randomly. Only a large sample size makes it likely that our sample is close to representative of the population. For this reason, inferential statistics take into account the sample size when generalizing results from samples to populations. In later chapters, you'll see what kinds of mathematical techniques ensure this sensitivity to sample size.

More complex sampling

Sometimes it is not feasible to build a sample using simple random sampling. To see the problem, consider the fact that both Dallas and Houston are competed to be hosts of the 2012 Olympics. Imagine that you are hired to assess whether most Texans prefer Houston to Dallas as the host, or the reverse. Given the impracticality of obtaining the opinion of every single Texan, you must construct a sample of the Texas population. But now notice how difficult it would be to proceed by simple random sampling. For

example, how will you contact those individuals who don't vote and don't have a phone? Even among people you find in the telephone book, how can you identify those who have just relocated to California (and had no reason to inform you of their move)? What do you do about the fact that since the beginning of the study, an additional 4,212 people took up residence in the state of Texas? As you can see, it is sometimes very difficult to develop a truly random procedure. For this reason, other kinds of sampling techniques have been devised. We now discuss two of them.

Stratified Sampling

Since simple random sampling often does not ensure a representative sample, a sampling method called **stratified random sampling** is sometimes used to make the sample more representative of the population. This method can be used if the population has a number of distinct "strata" or groups. In stratified sampling, you first identify members of your sample who belong to each group. Then you randomly sample from each of those subgroups in such a way that the sizes of the subgroups in the sample are proportional to their sizes in the population.

Let's take an example: Suppose you were interested in views of capital punishment at an urban university. You have the time and resources to interview 200 students. The student body is diverse with respect to age; many older people work during the day and enroll in night courses (average age is 39), while younger students generally enroll in day classes (average age of 19). It is possible that night students have different views about capital punishment than day students. If 70% of the students were day students, it makes sense to ensure that 70% of the sample consisted of day students. Thus, your sample of 200 students would consist of 140 day students and 60 night students. The proportion of day students in the sample and in the population (the entire university) would be the same. Inferences to the entire population of students at the university would therefore be more secure.

Convenience Sampling

Unfortunately, it is often impractical or impossible to study a truly random sample. Much of the time, in fact, studies are conducted with whoever is willing or available to be a research participant – this is commonly referred to as **convenience sampling**. At best, as noted, a researcher tries to study a sample that is not systematically unrepresentative of the population in any known way. For example, suppose a study is about a process that is likely to differ for people of different age groups. In this situation, the researcher may attempt to include people of all age groups in the study. Alternatively, the researcher would be careful to draw conclusions only about the age group studied. Remember that one of the goals of research is to make conclusions about the population from the sample results. An unbiased random sample and a representative sample are important when drawing conclusions from the results of a study.

"WEIRD" Culture Samples

Psychologists have been guilty of largely recruiting samples of convenience from the thin slice of humanity—students—found at universities and colleges (Sears, 1986). This presents a problem when trying to assess the social mechanics of the public at large. Aside from being an overrepresentation of young, middle-class Caucasians, college students may also be more compliant and more susceptible to attitude change, have less stable personality traits and interpersonal relationships, and possess stronger cognitive skills than samples reflecting a wide range of age and experience (Peterson & Merunka, 2014; Visser, Krosnick, & Lavrakas, 2000). Put simply, these traditional samples (college students) may not be sufficiently representative of the broader population. Furthermore, considering that 96% of participants in psychology studies come from *western, educated, industrialized, rich, and democratic countries* (so-called **WEIRD cultures**; Henrich, Heine, & Norenzayan, 2010), and that the majority of these *are also psychology students*, the question of non-representativeness becomes even more serious. Of course, when studying a basic cognitive process (like working memory capacity) or an aspect of social behavior that appears to be fairly universal (e.g., even cockroaches exhibit social facilitation), a non-representative sample may not be a big deal. Over time research has repeatedly demonstrated the important role that individual differences (e.g., personality traits, cognitive abilities, etc.) and culture (e.g., individualism versus collectivism) play in shaping social behavior. For instance, even if we only consider a tiny sample of research on aggression, we know that narcissists are more likely to respond to criticism with aggression (Bushman &Baumeister, 1998); conservatives, who have a low tolerance for uncertainty, are more likely to prefer aggressive actions against those considered to be 'outsiders'(de Zavala et al., 2010); countries, where men hold the bulk of the power in society, have higher rates of physical aggression directed against female partners (Archer, 2006); and males from the southern part of the United States are more likely to react with aggression following an insult (Cohen et al., 1996).

Why does random sampling work?

Below is an example showing how many credit hours students are currently enrolled in at a community college. This data represents the entire population of interest, all students currently enrolled in classes at Chandler-Gilbert Community College. Let's say we randomly selected one student out of the population and asked them how many credit hours they are currently taking. How likely would it be for this one student to represent the entire population? This is the first line showing 1 student reported taking 12 hours while the average credit hours for a CGCC student was 8 (population average).

Sample size (n)	Sample average	Population average	Difference between Sample & Population
1	12	8	4
2	15	8	7
5	9.8	8	1.8
25	9.5	8	1.5
250	8.3	8	0.3
2500	7.9	8	0.1

The larger the sample size, the more closely it represents the population.

As we can see from this activity, the larger our sample is, the more accurately it will represent the population from which it was drawn. This brings up a very important rule in research design. The larger the sample size is, the more accurately the sample will represent the population from which it was drawn. Also, if you are comparing groups, consider that the more diverse, or variable, individuals in each group are, the larger the sample needs to be to detect real differences between groups. We will further dive into the importance of sample sizes with inferential statistics, but for now, consider that the larger the sample, the more likely the researcher will represent the population.

2.3: Representative Sample is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by LibreTexts.

2.4: Type of Research Designs

Type of Research Designs

Research studies come in many forms, and, just like with the different types of data we have, different types of studies tell us different things. The choice of research design is determined by the research question and the logistics involved. Though a complete understanding of different research designs is the subject for at least one full class, if not more, a basic understanding of the principles is useful here. There are three types of research designs we will discuss: non-experimental, quasi-experimental, and random experimental.

Non-Experimental Designs

Non-experimental research (sometimes called correlational research) involves observing things as they occur naturally and recording our observations as data. In **observational studies**, information is gathered from observing. This could include self-report as well as interviews.

> Consider this example: A data scientist wants to know if there is a relation between how conscientious a person is and whether that person is a good employee. She hopes to use this information to predict the job performance of future employees by measuring their personality when they are still job applicants. She randomly samples volunteer employees from several different companies, measuring their conscientiousness and having their bosses rate their performance on the job. She analyzes this data to find a relation. Conscientiousness is a person-based variable that researcher must gather data from employees as they are in order to find a relation between her variables.

This type of research design cannot establish causality, it can still be quite useful. If the relation between conscientiousness and job performance is consistent, then it doesn't necessarily matter is conscientiousness causes good performance or if they are both caused by something else – she can still measure conscientiousness to predict future performance. Additionally, these studies have the benefit of reflecting reality as it actually exists since we as researchers do not change anything.

Experimental Designs

If we want to know if a change in one variable causes a change in another variable, we must use a true experiment. *A true experiment is an experimental design with random assignment.* In an **experimental design** a researcher assigns or manipulates, the group's participants will be in. Further, each participant is **randomly assigned** to a group. If there is no random assignment, the experiment can not have cause-effect conclusions.

Types of Variables in an Experiment

When conducting research, experimenters often manipulate variables. For example, an experimenter might compare the effectiveness of four types of antidepressants. In this case, the variable is "type of antidepressant." When a variable is manipulated by an experimenter, it is called an **independent variable**. The experiment seeks to determine the effect of the independent variable on relief from depression. In this example, relief from depression is called a **dependent variable**. In general, the independent variable is manipulated by the experimenter and its effects on the dependent variable are measured.

> To understand what this means, let's look at an example: A clinical researcher wants to know if a newly developed drug is effective in treating the flu. Working with collaborators at several local hospitals, she randomly samples 40 flu patients and randomly assigns each one to one of two conditions: Group A receives the new drug and Group B received a placebo. She measures the symptoms of all participants after 1 week to see if there is a difference in symptoms between the groups.

In the example, the *independent variable* is the drug treatment; we manipulate it into 2 levels: new drug or placebo. Without the researcher administering the drug (i.e. manipulating the independent variable), there would be no difference between the groups. Each person, after being randomly sampled to be in the research, was then randomly assigned to one of the 2 groups. That is, random sampling and random assignment are *not* the same thing and cannot be used interchangeably. *For research to be a true experiment, random assignment must be used.* For research to be representative of the population, random sampling must be used. The use of both techniques helps ensure that there are no systematic differences between the groups, thus eliminating the potential for sampling bias. The *dependent variable* in the example is flu symptoms. Barring any other intervention, we would assume that people in both groups, on average, get better at roughly the same rate. Because there are no systematic differences between the 2 groups, if the researcher does find a difference in symptoms, she can confidently attribute it to the effectiveness of the new drug.

Can you identify the independent and dependent variables?

Example #1: Can blueberries slow down aging? A study indicates that antioxidants found in blueberries may slow down the process of aging. In this study, 19-month- old rats (equivalent to 60-year-old humans) were fed either their standard diet or a diet supplemented by either blueberry, strawberry, or spinach powder (randomly assigned). After eight weeks, the rats were given memory and motor skills tests. Although all supplemented rats showed improvement, those supplemented with blueberry powder showed the most notable improvement.

- What is the independent variable? (dietary supplement: none, blueberry, strawberry, and spinach)
- What are the dependent variables? (memory test and motor skills test)

Example #2: Does beta-carotene protect against cancer? Beta-carotene supplements have been thought to protect against cancer. However, a study published in the Journal of the National Cancer Institute suggests this is false. The study was conducted with 39,000 women aged 45 and up. These women were randomly assigned to receive a beta-carotene supplement or a placebo, and their health was studied over their lifetime. Cancer rates for women taking the beta-carotene supplement did not differ systematically from the cancer rates of those women taking the placebo.

- What is the independent variable? (supplements: beta-carotene or placebo)
- What is the dependent variable? (occurrence of cancer)

Example #3: How bright is right? An automobile manufacturer wants to know how bright brake lights should be in order to minimize the time required for the driver of the following car to realize that the car in front is stopping and to hit the brakes.

- What is the independent variable? (brightness of brake lights)
- What is the dependent variable? (time to hit brakes)

Levels of an Independent Variable

In order to establish that one variable must cause a change in another variable and so a researcher will likely use two groups or levels in order to observe the changes and make comparisons.

- **Experimental (treatment) group** is the group who are exposed to the independent variable (or the manipulation) by the researcher; the experimental group represents the treatment group.
- **Control group** is the group who are not exposed to the treatment variable; the control group serves as the comparison group.

If an experiment compares an experimental treatment group with a control group, then the independent variable (type of treatment) has two levels: experimental and control. Further, if an experiment were comparing five types of diets, then the independent variable (type of diet) would have 5 levels. In general, the number of levels of an independent variable is the number of experimental conditions. *Another term for levels for the independent variable is groups, treatments, or conditions.*

Scores from the experimental group are compared to scores in the control group and if there is a systematic difference between groups then there is evidence of a relationship between variables. Let's use our earlier example of stress as a way to illustrate the experimental method. Let's assume that a researcher examining stress wants to test the impact of a stress reduction program on the stress levels of students and recruits 100 students to participate. Students are randomly assigned to either the experimental group or the control group. The experimental group participates in the stress reduction program but the control group does not. The stress-reduction program is the independent variable and stress level is the dependent variable. At the end of the training program each group, the experimental group, and the control group complete a stress test, and the scores are compared. If the stress reduction program worked, then the stress levels for the experimental group should be lower than the stress levels for the control group.

Quasi-Experimental Designs

Quasi-experimental research involves getting as close as possible to the conditions of a true experiment when we cannot meet all requirements. Specifically, a **quasi- experiment** involves manipulating the independent variable but *not* randomly assigning people to groups. There are several reasons this might be used. First, it may be unethical to deny potential treatment to someone if there is good reason to believe it will be effective and that the person would unduly suffer if they did not receive it. Alternatively, it may be impossible to randomly assign people to groups.

> Consider the following example: A professor wants to test out a new teaching method to see if it improves student learning. Because he is teaching two sections of the same course, he decides to teach one section the traditional way and the other

section using the new method. At the end of the semester, he compares the grades on the final for each class to see if there is a difference.

In this example, the professor has manipulated his teaching method, which is the independent variable, hoping to find a difference in student performance, the dependent variable. *However, because students enroll in courses, he cannot randomly assign the students to a particular group, thus precluding using a true experiment to answer his research question.* Because of this, we cannot know for sure that there are no systematic differences between the classes other than teaching style and therefore cannot determine causality.

Extraneous and Confounding Variables

Sometimes in a research study things happen that make it difficult for a researcher to determine whether the independent variable caused the change in the dependent variable. These have special names.

- An **extraneous variable** is something that occurs in the environment or happens to the participants that unintentionally (accidentally) influences the outcome of the study. An extraneous variable affects everyone in a study.In an experiment on the effect of expressive writing on health, for example, extraneous variables would include participant variables (individual differences) such as their writing ability, their diet, and their shoe size. They would also include situation or task variables such as the time of day when participants write, whether they write by hand or on a computer, and the weather. Extraneous variables pose a problem because many of them are likely to have some effect on the dependent variable. For example, participants' health will be affected by many things other than whether or not they engage in expressive writing. This can make it difficult to separate the effect of the independent variable from the effects of the extraneous variables, which is why it is important to control extraneous variables by holding them constant.
- A **confounding variable** is a type of extraneous variable that changes at the same time as the independent variable, making it difficult to discern which one is causing changes in the dependent variable.

2.4: Type of Research Designs is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

2.5: Working with Data

What are data?

The first important point about data is that data *are* – meaning that the word “data” is plural (though some people disagree with me on this). You might also wonder how to pronounce “data” – I say “day-tah”, but I know many people who say “dah-tah”, and I have been able to remain friends with them in spite of this. Now, if I heard them say “the data is” then that would be a bigger issue…

Operationalizing Variables

We need to have specifically defined how we are measuring our construct or our variable. The act of defining how to measure your data is to **operationalize**. Some variables are easier to define, like height or weight. I can measure height in inches tall or weight in pounds. Some other variables can be more open to measurement, like happiness or love. How would I measure happiness? Would I simply ask are you happy (yes or no)? Would I use a questionnaire for a self-report measure? Would I rate individuals from observing them for happiness? Would I ask their partner, teacher, parent, best friend about the person’s happiness? Researchers’ decisions on how to measure data is an important factor and helps to determine what kind of data is being used.

How would you measure happiness in a research study? Image Source

Qualitative and Quantitative Variables

Data are composed of *variables*, where a variable reflects a unique measurement or quantity. An important distinction between variables is between qualitative variables and quantitative variables. **Qualitative variables** are those that express a qualitative attribute such as hair color, eye color, religion, favorite movie, gender, and so on. Qualitative means that they describe a quality rather than a numeric quantity. *Qualitative variables are sometimes referred to as categorical variables.* For qualitative variables, response options are usually limited or fixed to a set of possible values. Assigning a person, animal or event to a category is done on the basis of some qualitative property. For example, in my stats course, I generally give an introductory survey, both to obtain data to use in class and to learn more about the students. One of the questions that I ask is “What is your favorite food?”, to which some of the answers have been: blueberries, chocolate, tamales, pasta, pizza, and mango. Those data are not intrinsically numerical; we could assign numbers to each one (1=blueberries, 2=chocolate, etc), but we would just be using the numbers as labels *rather than as real numbers*.

Personality type, gender, and shirt sizes are all categorical, or qualitative, variables. The values of a qualitative variable do not necessarily imply order and do not produce numerical responses or use real numbers. For example, there is an order to shirt size but shirt size is categorical and not number based. Another example is postal Zip Code data. Those numbers are represented as integers, but they don’t actually refer to a numeric scale; each zip code basically serves as a label or category representing a different region. Because this data is not using real numbers, what we do with those numbers is constrained; for example, it wouldn’t make sense to compute the average of those numbers.

More commonly in statistics we will work with *quantitative* data, meaning data that are numerical. For example, here Table 1 shows the results from another question that I ask in my introductory class, which is “Why are you taking this class?”

Table 1: Counts of the prevalence of different responses to the question “Why are you taking this class?”

Why are you taking this class?	Number of students
It fulfills a degree plan requirement	105
It fulfills a General Education Breadth Requirement	32
It is not required but I am interested in the topic	11
Other	4

Note that the students' answers were qualitative, but we generated a quantitative summary of them by counting how many students gave each response. **Quantitative variables** are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and shoe size.

Experimental studies can involve qualitative and quantitative data. In the study on the effect of diet discussed previously, the independent variable was type of supplement: none, strawberry, blueberry, and spinach. The variable "type of supplement" is a qualitative variable; there is nothing quantitative about it. In contrast, the dependent variable "memory test" is a quantitative variable since memory performance was measured on a quantitative scale (number correct).

Discrete and Continuous Variables

Variables such as number of children in a household are called **discrete variables** since the possible scores are discrete points on the scale. For example, a household could have three children or six children, but not 4.53 children. Other variables such as "time to respond to a question" are **continuous variables** since the scale is continuous and not made up of discrete steps. The response time could be 1.64 seconds, or it could be 1.64237123922121 seconds. Of course, the practicalities of measurement preclude most measured variables from being truly continuous.

2.5: Working with Data is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

2.6: Levels of Measurement

2.6.0.0.1 Levels of Measurement

Numbers mean different things in different situations. Consider three answers below that appear to be the same, but they really are not. All three questions pertain to a running race that you just finished. The three 5s all look the same. However, the three variables (identification number, finish place, and time) are quite different. Because of these different variables, the way we interpret 5 is unique for each variable.

- What number were you wearing in the race?
- What place did you finish in?
- How many minutes did it take you to finish the race?

To illustrate the difference, consider your friend who also ran the race. Their answers to the same three questions were 10, 10, and 10. If we take the first question by itself and know that you had a score of 5, and your friend had a score of 10, what can we conclude? We can conclude that your race identification number is different from your friend's number. That is all we can conclude. On the second question, with scores of 5 and 10, what can we conclude regarding the place you and your friend finished in the race? We can state that you were faster than your friend in the race and, of course, that your finishing places are different. Comparing the 5 and 10 on the third question, what can we conclude? We could state that you ran the race twice as fast as your friend, you ran the race faster than your friend and that your time was different than your friend's time. The point of this discussion is to demonstrate the ***relationship between the questions we ask, and what the answers to those questions can tell us*** . Chances are, much of your past experience with numbers has been with pure numbers or with measurements such as time, length, and amount. "Four is twice as much as two" is true for the pure numbers themselves and for time, length, and amount –but this statement would not be true for finish places in a race. Fourth place is not twice anything in relation to 2nd place. Fourth place is not twice as slow or twice as far behind the 2nd place runner. The types of descriptive and inferential statistics we can use depend on the type of variable measured. Remember, a variable is defined as a characteristic we can measure that can assume more than one value.

For statistical analysis, exactly how the measurement is carried out depends on the type of variable involved in the analysis. Different types are measured differently. To measure the time taken to respond to a stimulus, you might use a stopwatch. Stopwatches are of no use, of course, when it comes to measuring someone's attitude towards a political candidate. A rating scale is more appropriate in this case (with labels like "very favorable," "somewhat favorable," etc.). For a dependent variable such as "favorite color," you can simply note the color-word (like "red") that the subject offers.

Although procedures for measurement differ in many ways, they can be classified using a few fundamental categories. The psychologist S. S. Stevens suggested that scores can be assigned to individuals so that they communicate more or less quantitative information about the variable of interest (Stevens, 1946). Stevens actually suggested four different levels of measurement(which he called "scales of measurement") that correspond to four different levels of quantitative information that can be communicated by a set of scores. In a given category, all of the procedures share some properties that are important for you to know about. The categories are called "scale types," or just "scales," and are described in this section.

Nominal scales

When measuring using a nominal scale, one simply names or categorizes responses. Gender, handedness, favorite color, and religion are examples of variables measured on a nominal scale. The essential point about nominal scales is that they do not imply any ordering among the responses. For example, when classifying people according to their favorite color, there is no sense in which green is placed "ahead of" blue. Responses are merely categorized. Nominal scales embody the lowest level of measurement.

Ordinal *scales*

A researcher wishing to measure consumers' satisfaction with their microwave ovens might ask them to specify their feelings as either "very dissatisfied," "somewhat dissatisfied," "somewhat satisfied," or "very satisfied." The items in this scale are ordered, ranging from least to most satisfied. This is what distinguishes ordinal from nominal scales. Unlike nominal scales, ordinal scales allow comparisons of the degree to which two subjects possess the dependent variable. For example, our satisfaction ordering makes it meaningful to assert that one person is more satisfied than

another with their microwave ovens. Such an assertion reflects the first person's use of a verbal label that comes later in the list than the label chosen by the second person.

On the other hand, ordinal scales fail to capture important information that will be present in the other scales we examine. In particular, the difference between two levels of an ordinal scale cannot be assumed to be the same as the difference between two other levels. In our satisfaction scale, for example, the difference between the responses "very dissatisfied" and "somewhat dissatisfied" is probably not equivalent to the difference between "somewhat dissatisfied" and "somewhat satisfied." Nothing in our measurement procedure allows us to determine whether the two differences reflect the same difference in psychological satisfaction.

Statisticians express this point by saying that the differences between adjacent scale values do not necessarily represent equal intervals on the underlying scale giving rise to the measurements. (In our case, the underlying scale is the true feeling of satisfaction, which we are trying to measure.)

What if the researcher had measured satisfaction by asking consumers to indicate their level of satisfaction by choosing a number from one to four? Would the difference between the responses of one and two necessarily reflect the same difference in satisfaction as the difference between the responses two and three? The answer is No. Changing the response format to numbers does not change the meaning of the scale. We still are in no position to assert that the mental step from 1 to 2 (for example) is the same as the mental step from 3 to 4.

(Equal) Interval *scales*

Interval scales are numerical scales in which intervals have the same interpretation throughout. As an example, consider the Fahrenheit scale of temperature. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 degrees. This is because each 10-degree interval has the same physical meaning (in terms of the kinetic energy of molecules).

Interval scales are not perfect, however. In particular, they do not have a true zero point even if one of the scaled values happens to carry the name "zero." The Fahrenheit scale illustrates the issue. Zero degrees Fahrenheit does not represent the complete absence of temperature (the absence of any molecular kinetic energy). In reality, the label "zero" is applied to its temperature for quite accidental reasons connected to the history of temperature measurement. Since an interval scale has no true zero point, it does not make sense to compute ratios of temperatures. For example, there is no sense in which the ratio of 40 to 20 degrees Fahrenheit is the same as the ratio of 100 to 50 degrees; no interesting physical property is preserved across the two ratios. After all, if the "zero" label were applied at the temperature that Fahrenheit happens to label as 10 degrees, the two ratios would instead be 30 to 10 and 90 to 40, no longer the same! For this reason, it does not make sense to say that 80 degrees is "twice as hot" as 40 degrees. Such a claim would depend on an arbitrary decision about where to "start" the temperature scale, namely, what temperature to call zero (whereas the claim is intended to make a more fundamental assertion about the underlying physical reality).

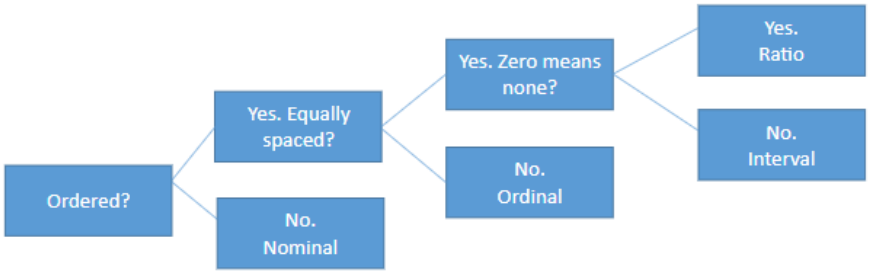
Ratio scales (Absolute zero)

The ratio scale of measurement is the most informative scale. It is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured. You can think of a ratio scale as the three earlier scales rolled up in one. Like a nominal scale, it provides a name or category for each object (the numbers serve as labels). Like an ordinal scale, the objects are ordered (in terms of the ordering of the numbers). Like an interval scale, the same difference at two places on the scale has the same meaning. And in addition, the same ratio at two places on the scale also carries the same meaning.

The Fahrenheit scale for temperature has an arbitrary zero point and is therefore not a ratio scale. However, zero on the Kelvin scale is absolute zero. This makes the Kelvin scale a ratio scale. For example, if one temperature is twice as high as another as measured on the Kelvin scale, then it has twice the kinetic energy of the other temperature.

Another example of a ratio scale is the amount of money you have in your pocket right now (25 cents, 55 cents, etc.). Money is measured on a ratio scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have zero money, this implies the absence of money. Since money has a true zero point, it makes sense to

say that someone with 50 cents has twice as much money as someone with 25 cents (or that Bill Gates has a million times more money than you do).



Here is a decision tree that you might find helpful for classifying data into N-O-I-R

Kara's note: I use the acronym NOIR (pronounced like the French word meaning dark or black: nuh-waar) to help me remember the order of the levels.

Digging deeper: What about the number value? It is important to know what number values mean. Is the number meaningful or it is a category? This section briefly reviews how numbers can be categorized according to meaning.

Binary numbers. The simplest are binary numbers – that is, zero or one. We will often use binary numbers to represent whether something is true or false, or present or absent. For example, I might ask 10 people if they have ever experienced a migraine headache, recording their answers as "Yes" or "No". It's often useful to instead use *logical* values, which take the value of either `TRUE` or `FALSE`. This can be especially useful for programming languages to analyze data, since these languages already understand the concepts of TRUE and FALSE. In fact, most programming languages treat truth values and binary numbers equivalently. The number 1 is equal to the logical value `TRUE`, and the number zero is equal to the logical value `FALSE`.

Integers. Integers are whole numbers with no fractional or decimal part. We most commonly encounter integers when we count things, but they also often occur in psychological measurement. For example, in my introductory survey I administer a set of questions about attitudes towards statistics (such as "Statistics seems very mysterious to me."), on which the students respond with a number between 1 ("Disagree strongly") and 7 ("Agree strongly"). Integers are discontinuous.

Real numbers. Most commonly in statistics we work with real numbers, which have a fractional/decimal part. For example, we might measure someone's weight, which can be measured to an arbitrary level of precision, from kilograms down to micrograms. Real numbers can be discontinuous or continuous.

2.6: Levels of Measurement is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

2.7: What level of measurement is used for psychological variables?

2.7.0.0.1 What level of measurement is used for psychological variables?

Rating scales are used frequently in psychological research. For example, experimental subjects may be asked to rate their level of pain, how much they like a consumer product, their attitudes about capital punishment, their confidence in an answer to a test question. Typically these ratings are made on a 5-point or a 7-point scale. These scales are ordinal scales since there is no assurance that a given difference represents the same thing across the range of the scale. For example, there is no way to be sure that a treatment that reduces pain from a rated pain level of 3 to a rated pain level of 2 represents the same level of relief as a treatment that reduces pain from a rated pain level of 7 to a rated pain level of 6.

In memory experiments, the dependent variable is often the number of items correctly recalled. What scale of measurement is this? You could reasonably argue that it is a ratio scale. First, there is a true zero point; some subjects may get no items correct at all. Moreover, a difference of one represents a difference of one item recalled across the entire scale. It is certainly valid to say that someone who recalled 12 items recalled twice as many items as someone who recalled only 6 items.

But number-of-items recalled is a more complicated case than it appears at first. Consider the following example in which subjects are asked to remember as many items as possible from a list of 10. Assume that (a) there are 5 easy items and 5 difficult items, (b) half of the subjects are able to recall all the easy items and different numbers of difficult items, while (c) the other half of the subjects are unable to recall any of the difficult items but they do remember different numbers of easy items. Some sample data are shown below.

Subject	Easy Items					Difficult Items					Score
A	0	0	1	1	0	0	0	0	0	0	2
B	1	0	1	1	0	0	0	0	0	0	3
C	1	1	1	1	1	1	1	0	0	0	7
D	1	1	1	1	1	0	1	1	0	1	8

Let's compare (i) the difference between Subject A's score of 2 and Subject B's score of 3 and (ii) the difference between Subject C's score of 7 and Subject D's score of 8. The former difference is a difference of one easy item; the latter difference is a difference of one difficult item. Do these two differences necessarily signify the same difference in memory? We are inclined to respond "No" to this question since only a little more memory may be needed to retain the additional easy item whereas a lot more memory may be needed to retain the additional hard item. *The general point is that it is often inappropriate to consider psychological measurement scales as either interval or ratio. You will often see in statistical software that the distinction is between nominal, ordinal, and interval/ratio.*

2.7.0.0.1 Consequences of level of measurement

Why are we so interested in the type of scale that measures a dependent variable? The crux of the matter is the relationship between the variable's level of measurement and the statistics that can be meaningfully computed with that variable. For example, consider

a hypothetical study in which 5 children are asked to choose their favorite color from blue, red, yellow, green, and purple. The researcher codes the results as follows:

Color	Code
Blue	1
Red	2
Yellow	3
Green	4
Purple	5

This means that if a child said her favorite color was “Red,” then the choice was coded as “2,” if the child said her favorite color was “Purple,” then the response was coded as 5, and so forth. Consider the following hypothetical data:

Subject	Color	Code
1	Blue	1
2	Blue	1
3	Green	4
4	Green	4
5	Purple	5

Each code is a number, so nothing prevents us from computing the average code assigned to the children. The average happens to be 3, but you can see that it would be senseless to conclude that the average favorite color is yellow (the color with a code of 3). Such nonsense arises because favorite color is a nominal scale, and taking the average of its numerical labels is like counting the number of letters in the name of a snake to see how long the beast is.

Does it make sense to compute the mean of numbers measured on an ordinal scale? This is a difficult question, one that statisticians have debated for decades. The prevailing (but by no means unanimous) opinion of statisticians is that for almost all practical situations, the mean of an ordinally-measured variable is a meaningful statistic. However, there are extreme situations in which computing the mean of an ordinally-measured variable can be very misleading.

2.7.1 What makes a good measurement?

In many fields such as psychology, the thing that we are measuring is not a physical feature, but instead is an unobservable theoretical concept, which we usually refer to as a *construct*. For example, let’s say that I want to test how well you understand the distinction between the different types of numbers described above. I could give you a pop quiz that would ask you several questions about these concepts and count how many you got right. This test might or might not be a good measurement of the construct of your actual knowledge – for example, if I were to write the test in a confusing way or use language that you don’t understand, then the test might suggest you don’t understand the concepts when really you do. On the other hand, if I give a multiple-choice test with very obvious wrong answers, then you might be able to perform well on the test even if you don’t actually understand the material.

It is usually impossible to measure a construct without some amount of error. In the example above, you might know the answer, but you might misread the question and get it wrong. In other cases, there is error intrinsic to the thing being measured, such as when we measure how long it takes a person to respond on a simple reaction time test, which will vary from trial to trial for many reasons. We generally want our measurement error to be as low as possible, which we can acheive either by improving the quality

of the measurement (for example, using a better time to measure reaction time), or by averaging over a larger number of individual measurements.

Sometimes there is a standard against which other measurements can be tested, which we might refer to as a “gold standard” – for example, measurement of sleep can be done using many different devices (such as devices that measure movement in bed), but they are generally considered inferior to the gold standard of polysomnography (which uses measurement of brain waves to quantify the amount of time a person spends in each stage of sleep). Often the gold standard is more difficult or expensive to perform, and the cheaper method is used even though it might have greater error.

When we think about what makes a good measurement, we usually distinguish two different aspects of a good measurement: it should be *reliable*, and it should be *valid*.

2.7: What level of measurement is used for psychological variables? is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

2.8: Reliability, Validity, and Results

2.8.0.1 Reliability

Reliability refers to the consistency of our measurements. One common form of reliability, known as “test-retest reliability”, measures how well the measurements agree if the same measurement is performed twice. For example, I might give you a questionnaire about your attitude towards statistics today, repeat this same questionnaire tomorrow, and compare your answers on the two days; we would hope that they would be very similar to one another, unless something happened in between the two tests that should have changed your view of statistics (like reading this book!).

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

Reliability is important if we want to compare one measurement to another, because the relationship between two different variables can’t be any stronger than the relationship between either of the variables and itself (i.e., its reliability). This means that an unreliable measure can never have a strong statistical relationship with any other measure. For this reason, researchers developing a new measurement (such as a new survey) will often go to great lengths to establish and improve its reliability.

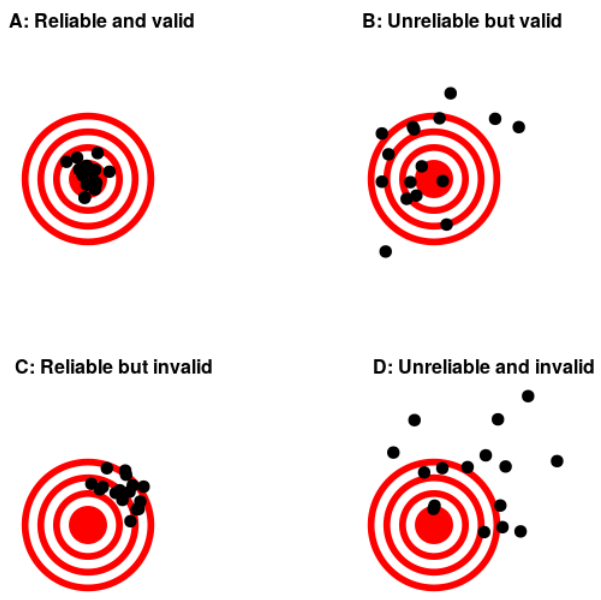


Figure 1: A figure demonstrating the distinction between reliability and validity, using shots at a bullseye. Reliability refers to the consistency of location of shots, and validity refers to the accuracy of the shots with respect to the center of the bullseye.

2.8.0.1 Validity

Reliability is important, but on its own it’s not enough: After all, I could create a perfectly reliable measurement on a personality test by re-coding every answer using the same number, regardless of how the person actually answers. We want our measurements to also be *valid* — that is, we want to make sure that we are actually measuring the construct that we think we are measuring (Figure 1). There are many different types of validity that are commonly discussed; we will focus on three of them.

Face validity. Does the measurement make sense on its face? If I were to tell you that I was going to measure a person’s blood pressure by looking at the color of their tongue, you would probably think that this was not a valid measure on its face. On the

other hand, using a blood pressure cuff would have face validity. This is usually a first reality check before we dive into more complicated aspects of validity.

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

Predictive validity. If our measurements are truly valid, then they should also be predictive of other outcomes. For example, let's say that we think that the psychological trait of sensation seeking (the desire for new experiences) is related to risk taking in the real world. To test for predictive validity of a measurement of sensation seeking, we would test how well scores on the test predict scores on a different survey that measures real-world risk taking.

2.8.1 Critical Evaluation of Statistical Results

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include:

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions.
- Undue influence: collecting data or asking questions in a way that influences the response
- Non-response or refusal of a participant to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

2.8: Reliability, Validity, and Results is shared under a CC BY-NC license and was authored, remixed, and/or curated by LibreTexts.

2.9: Types of Statistical Analyses

2.9.1 Types of Statistical Analyses

Now that we understand the nature of our data, let's turn to the types of statistics we can use to interpret them. As mentioned at the end of chapter 1, there are 2 types of statistics: descriptive and inferential.

2.9.1.0.1 Descriptive Statistics

Descriptive statistics are numbers that are used to summarize and describe data. The word "data" refers to the information that has been collected from an experiment, a survey, an historical record, etc. (By the way, "data" is plural. One piece of information is called a "datum.") If we are analyzing birth certificates, for example, a descriptive statistic might be the percentage of certificates issued in New York State, or the average age of the mother. Any other number we choose to compute also counts as a descriptive statistic for the data from which the statistic is computed. Several descriptive statistics are often used at one time to give a full picture of the data.

Descriptive statistics are just descriptive. They do not involve generalizing beyond the data at hand. Generalizing from our data to another set of cases is the business of inferential statistics, which you'll be studying in another section. Here we focus on (mere) descriptive statistics.

Some descriptive statistics are shown in Table 2. The table shows the average salaries for various occupations in the United States in 1999.

Salary 1999	Salary 2019	Occupation
$112,760	$175,310	pediatricians
$106,130	$155,600	dentists
$100,090	$126,240	podiatrists
$76,140	$97,152	physicists
$53,410	$80,750	architects
$49,720	$78,200	school, clinical, and counseling psychologists
$47,910	$56,640	flight attendants
$39,560	$59,670	elementary school teachers
$38,710	$65,170	police officers
$18,980	$28, 040	floral designers

Table 2. Average salaries for various occupations in 1999 and 2019 (median salaries reported by Bureau of Labor Statistics).

Descriptive statistics like these offer insight into American society. It is interesting to note, for example, that we pay the people who educate our children and who protect our citizens a great deal less than we pay people who take care of our feet or our teeth.

For more descriptive statistics, consider Table 3. It shows the number of employed single young men to single young women for large metro areas in the US (reported in 2014). From this table we see that men outnumber women most in the San Jose, CA area, and women outnumber

men most in the Memphis, TN area. You can see that descriptive statistics can be useful if we are looking for an opposite-sex partner between the ages of 25-34 years old! (These data come from Pew Research)

Highest Ratios of Employed Single Men to Single Women (25-34 y/o)	Men per 100 Women	Lowest Ratios of Employed Single Men to Single Women (25-34 y/o)	Men per 100 Women
1. San-Jose-Sunnyvale-Santa Clara, CA	114	1. Memphis, TN-MS-AR	59
2. Denver-Aurora-Lakewood, CO	101	2. Jacksonville, FL	70
3. San Diego-Carlsbad, CA	99	3. Detroit-Warren-Dearborn, MI	71
4. Minneapolis-St. Paul-Bloomington, MN-WI	98	4. Charlotte-Concord-Gastonia, NC-SC	73
5. Seattle-Tacoma-Bellevue, WA	96	5. Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	74
6. San Francisco-Oakland-Hayward, CA	93	6. Kansas City, MO-KS	75
7. Washington-Arlington-Alexandra, DC-VA-MD-WV	92	7. Nashville-Davidson-Murfreesboro-Franklin, TN	77
8. Los Angeles-Long Beach-Anaheim, CA	91	8. Miami-Fort Lauderdale-West-Palm Beach, FL	78
9. Pittsburgh, PA	90	9. New Orleans-Metairie, LA	78
10. Orlando-Kissimmee-Sanford, FL	90	10. Cincinnati, OH-KY-IN	78

Table 3. Number of employed, 25-34 year old ratio of men to women in large metro areas of the U.S. (Pew Research, 2014)

These descriptive statistics may make us ponder why there are ratio differences in these large metropolitan areas. You probably know that descriptive statistics are central to the world of sports. Every sporting event produces numerous statistics such as the shooting percentage of players on a basketball team. For the Olympic marathon (a foot race of 26.2 miles), we possess data that cover more than a century of

competition. (The first modern Olympics took place in 1896.) The following table shows the winning times for both men and women (the latter have only been allowed to compete since 1984).

Women			
Year	**Winner**	**Country**	**Time**
1984	Joan Benoit	USA	2:24:52
1988	Rosa Mota	POR	2:25:40
1992	Valentina Yegorova	UT	2:32:41
1996	Fatuma Roba	ETH	2:26:05
2000	Naoko Takahashi	JPN	2:23:14
2004	Mizuki Noguchi	JPN	2:26:20
2008	Constantina Dita-Tomescu	Romania	2:26:44
2012	Tiki Gelana	ETH	2:23:07
2016	Jemima Sumgong	Kenya	2:24:04
2020	Peres Jepchirchir	Kenya	2:27:20
Men			
Year	**Winner**	**Country**	**Time**
1896	Spiridon Louis	GRE	2:58:50
1900	Michel Theato	FRA	2:59:45
1904	Thomas Hicks	USA	3:28:53
1906	Billy Sherring	CAN	2:51:23
1908	Johnny Hayes	USA	2:55:18
1912	Kenneth McArthur	S. Afr.	2:36:54
1920	Hannes Kolehmainen	FIN	2:32:35
1924	Albin Stenroos	FIN	2:41:22
1928	Boughra El Ouafi	FRA	2:32:57
1932	Juan Carlos Zabala	ARG	2:31:36
1936	Sohn Kee-Chung	JPN	2:29:19
1948	Delfo Cabrera	ARG	2:34:51
1952	Emil Ztopek	CZE	2:23:03

1956	Alain Mimoun	FRA	2:25:00
1960	Abebe Bikila	ETH	2:15:16
1964	Abebe Bikila	ETH	2:12:11
1968	Mamo Wolde	ETH	2:20:26
1972	Frank Shorter	USA	2:12:19
1976	Waldemar Cierpinski	E.Ger	2:09:55
1980	Waldemar Cierpinski	E.Ger	2:11:03
1984	Carlos Lopes	POR	2:09:21
1988	Gelindo Bordin	ITA	2:10:32
1992	Hwang Young-Cho	S. Kor	2:13:23
1996	Josia Thugwane	S. Afr.	2:12:36
2000	Gezahenge Abera	ETH	2:10.10
2004	Stefano Baldini	ITA	2:10:55
2008	Samuel Wanjiru	Kenya	2:06:32
2012	Stephen Kiprotich	Uganda	2:08:01
2016	Eliud Kipchoge	Kenya	2:08:44
2020	Eliud Kipchoge	Kenya	2:08:38

Table 4. Winning Olympic marathon times.

There are many descriptive statistics that we can compute from the data in the table. To gain insight into the improvement in speed over the years, let us divide the men's times into two pieces, namely, the first 13 races (up to 1952) and the second 13 (starting from 1956). The mean winning time for the first 13 races is 2 hours, 44 minutes, and 22 seconds (written 2:44:22). The mean winning time for the second 13 races is 2:13:18. This is quite a difference (over half an hour). Does this prove that the fastest men are running faster? Or is the difference just due to chance, no more than what often emerges from chance differences in performance from year to year? We can't answer this question with descriptive statistics alone. All we can affirm is that the two means are "suggestive."

Examining Table 4 leads to many other questions. We note that Takahashi (the lead female runner in 2000) would have beaten the male runner in 1956 and all male runners in the first 12 marathons. This fact leads us to ask whether the gender gap will close or remain constant. When we look at the times within each gender, we also wonder how far they will decrease (if at all) in the next century of the Olympics. Might we one day witness a sub-2 hour marathon? The study of statistics can help you make reasonable guesses about the answers to these questions.

It is also important to differentiate what we use to describe populations vs what we use to describe samples. A population is described by a parameter; the parameter is the true value of the descriptive in the population, but one that we can never know for sure. For example, the Bureau of Labor Statistics reports that the average hourly wage of chefs or head cooks is $25.66[1]. However, even if this number was computed using information from every single chef in the United States (making it a parameter), it would quickly become slightly off as one chef retires and a new chef enters the job market. Additionally, as noted above, there is virtually no way to collect data from every single person in a population. In order to understand a variable, we estimate the population parameter using a sample statistic. Here, the term "statistic" refers to the specific number we compute from the data (e.g. the average), not the field of statistics. A sample statistic is an estimate

of the true population parameter, and if our sample is representative of the population, then the statistic is considered to be a good estimator of the parameter.

Even the best sample will be somewhat off from the full population, earlier referred to as sampling bias, and as a result, there will always be a tiny discrepancy between the parameter and the statistic we use to estimate it. This difference is known as sampling error, and, as we will see throughout the course, understanding sampling error is the key to understanding statistics. Every observation we make about a variable, be it a full research study or observing an individual's behavior, is incapable of being completely representative of all possibilities for that variable.

Knowing where to draw the line between an unusual observation and a true difference is what statistics is all about.

2.9.1.0.1 Inferential Statistics

Descriptive statistics are wonderful at telling us what our data look like. However, what we often want to understand is how our data behave. What variables are related to other variables? Under what conditions will the value of a variable change? Are two groups different from each other, and if so, are people within each group different or similar? These are the questions answered by inferential statistics, and inferential statistics are how we generalize from our sample back up to our population. Units 2 and 3 are all about inferential statistics, the formal analyses and tests we run to make conclusions about our data.

For example, we will learn how to use a t statistic to determine whether people change over time when enrolled in an intervention. We will also use an F statistic to determine if we can predict future values on a variable based on current known values of a variable. There are many types of inferential statistics, each allowing us insight into a different behavior of the data we collect. This course will only touch on a small subset (or a *sample*) of them, but the principles we learn along the way will make it easier to learn new tests, as most inferential statistics follow the same structure and format.

2.9: Types of Statistical Analyses is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

2.10: Mathematical Notation

2.10.0.1 Mathematical Notation

As noted above, statistics is not math. It does, however, use math as a tool. Many statistical formulas involve summing numbers. Fortunately there is a convenient notation for expressing summation. This section covers the basics of this summation notation.

Let's say we have a variable X that represents the weights (in grams) of 4 grapes:

Grape	X
1	4.6
2	5.1
3	4.9
4	4.4

$$\sum_{i=1}^{4} X_i$$

We label Grape 1's weight X_1, Grape 2's weight X_2, etc. The following formula means to sum up the weights of the four grapes:

The Greek letter Σ indicates summation. The "i = 1" at the bottom indicates that the summation is to start with X_1 and the 4 at the top indicates that the summation will end with X_4. The "X_i" indicates that X is the variable to be summed as i goes from 1 to 4. Therefore,

$$\sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4 = 4.6 + 5.1 + 4.9 + 4.4 = 19$$

The symbol

$$\sum_{i=1}^{3} X_i$$

indicates that only the first 3 scores are to be summed. The index variable i goes from 1 to 3.

When all the scores of a variable (such as X) are to be summed, it is often convenient to use the following abbreviated notation:

$$\sum X$$

Thus, when no values of i are shown, it means to sum all the values of X.

Many formulas involve squaring numbers before they are summed. This is indicated as

$$\sum X^2 = 4.6^2 + 5.1^2 + 4.9^2 + 4.4^2$$

$$= 21.16 + 26.01 + 24.01 + 19.36 = 90.54$$

Notice that:

$$\left(\sum X\right)^2 \neq \sum X^2$$

because the expression on the left means to sum up all the values of X and then square the sum ($19^2 = 361$), whereas the expression on the right means to square the numbers and then sum the squares (90.54, as shown).

Some formulas involve the sum of cross products. Below are the data for variables X and Y. The cross products (XY) are shown in the third column. The sum of the cross products is 3 + 4 + 21 = 28.

X	Y	XY
1	3	3
2	2	4
3	7	21

In summation notation, this is written as:

$$\sum XY = 28$$

Three key concepts for statistical formulas:

1. Perform summation in the correct order following the order of operations (PEMDAS).
2. Typically we will use a set of scores for the mathematical operations/formulas used in statistics.
3. Each operation, except for summation, creates a new column of numbers (we will see this in action in chapter 4). Summation adds up the sum for the column and is typically seen as the last row.

2.10: Mathematical Notation is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

2.E: Exercises

2.E.0.1 Exercises – Ch. 2

1. In your own words, describe why we study statistics.
2. For each of the following, determine if the variable is continuous or discrete:
    1. Time taken to read a book chapter
    2. Favorite food
    3. Cognitive ability
    4. Temperature
    5. Letter grade received in a class
3. For each of the following, determine the level of measurement:
    1. T-shirt size
    2. Time taken to run 100 meter race
    3. First, second, and third place in 100 meter race
    4. Birthplace
    5. Temperature in Celsius
4. What is the difference between a population and a sample? Which is described by a parameter and which is described by a statistic?
5. What is sampling bias? What is sampling error?
6. What is the difference between a simple random sample and a stratified random sample?
7. What are the two key characteristics of a true experimental design?
8. When would we use a quasi-experimental design?
9. Use the following dataset for the computations below:

X	Y
1	1
2	3
5	5
7	1

    Computations to use for the above data set:

    1. ΣX
    2. ΣY^2
    3. ΣXY
    4. $(\Sigma Y)^2$
10. What are the most common measures of central tendency and spread?

Answers to Odd-Numbered Exercises – Ch. 2 - Highlight to reveal them

1. [redacted]
2. [redacted]
    1. [redacted]
    2. [redacted]
    3. [redacted]
    4. [redacted]
    5. [redacted]

3. [redacted]
4. [redacted]
5. [redacted]
    1. [redacted]
    2. [redacted]
    3. [redacted]
    4. [redacted]

2.E: Exercises is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

3: Measures of Central Tendency and Spread

3.1: What is Central Tendency?
3.2: Measures of Central Tendency
3.3: Spread and Variability
3.E: Measures of Central Tendency and Spread (Exercises)

This page titled 3: Measures of Central Tendency and Spread is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

3.1: What is Central Tendency?

What is “central tendency,” and why do we want to know the central tendency of a group of scores? Let us first try to answer these questions intuitively. Then we will proceed to a more formal discussion.

Imagine this situation: You are in a class with just four other students, and the five of you took a 5-point pop quiz. Today your instructor is walking around the room, handing back the quizzes. She stops at your desk and hands you your paper. Written in bold black ink on the front is “3/5.” How do you react? Are you happy with your score of 3 or disappointed? How do you decide? You might calculate your percentage correct, realize it is 60%, and be appalled. But it is more likely that when deciding how to react to your performance, you will want additional information. What additional information would you like?

If you are like most students, you will immediately ask your neighbors, “Whad'ja get?” and then ask the instructor, “How did the class do?” In other words, the additional information you want is how your quiz score compares to other students' scores. You therefore understand the importance of comparing your score to the class distribution of scores. Should your score of 3 turn out to be among the higher scores, then you'll be pleased after all. On the other hand, if 3 is among the lower scores in the class, you won't be quite so happy.

This idea of comparing individual scores to a distribution of scores is fundamental to statistics. So let's explore it further, using the same example (the pop quiz you took with your four classmates). Three possible outcomes are shown in Table **3.1.1**. They are labeled “Dataset A,” “Dataset B,” and “Dataset C.” Which of the three datasets would make you happiest? In other words, in comparing your score with your fellow students' scores, in which dataset would your score of 3 be the most impressive?

Table **3.1.1**: Three possible datasets for the 5-point make-up quiz.

Student	Dataset A	Dataset B	Dataset C
You	3	3	3
John's	3	4	2
Maria's	3	4	2
Shareecia's	3	4	2
Luther's	3	5	1

In Dataset A, everyone's score is 3. This puts your score at the exact center of the distribution. You can draw satisfaction from the fact that you did as well as everyone else. But of course it cuts both ways: everyone else did just as well as you.

Now consider the possibility that the scores are described as in Dataset B. This is a depressing outcome even though your score is no different than the one in Dataset A. The problem is that the other four students had higher grades, putting yours below the center of the distribution.

Finally, let's look at Dataset C. This is more like it! All of your classmates score lower than you so your score is above the center of the distribution.

Now let's change the example in order to develop more insight into the center of a distribution. Figure **3.1.1** shows the results of an experiment on memory for chess positions. Subjects were shown a chess position and then asked to reconstruct it on an empty

chess board. The number of pieces correctly placed was recorded. This was repeated for two more chess positions. The scores represent the total number of chess pieces correctly placed for the three chess positions. The maximum possible score was 89.

Non-Chess Players	Chess Players
40	80
43	85
39	76
34	75
32	71
30	63
22	63
22	62
26	62
	58
	56
	51
	46
	40

Figure **3.1.1**: Table of values from the chess memory experiment. The left side shows the memory scores of the non-players. The right side shows the scores of the tournament players.

Two groups are compared. On the left are people who don't play chess. On the right are people who play a great deal (tournament players). It is clear that the location of the center of the distribution for the non-players is much lower than the center of the distribution for the tournament players.

We're sure you get the idea now about the center of a distribution. It is time to move beyond intuition. We need a formal definition of the center of a distribution. In fact, we'll offer you three definitions! This is not just generosity on our part. There turn out to be (at least) three different ways of thinking about the center of a distribution, all of them useful in various contexts. In the remainder of this section we attempt to communicate the idea behind each concept. In the succeeding sections we will give statistical measures for these concepts of central tendency.

Definitions of Center

Now we explain the three different ways of defining the center of a distribution. All three are called measures of central tendency.

Balance Scale

One definition of central tendency is the point at which the distribution is in balance. Figure **3.1.2** shows the distribution of the five numbers 2, 3, 4, 9, 16 placed upon a balance scale. If each number weighs one pound, and is placed at its position along the number line, then it would be possible to balance them by placing a fulcrum at 6.8.

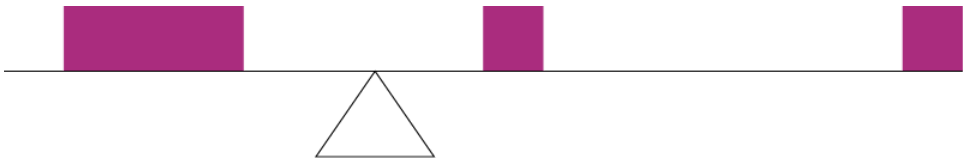


Figure **3.1.2**: A balance scale.

Image Credit: Judy Schmitt, from Cote et al, 2021

For another example, consider the distribution shown in Figure 3.1.3. It is balanced by placing the fulcrum in the geometric middle.

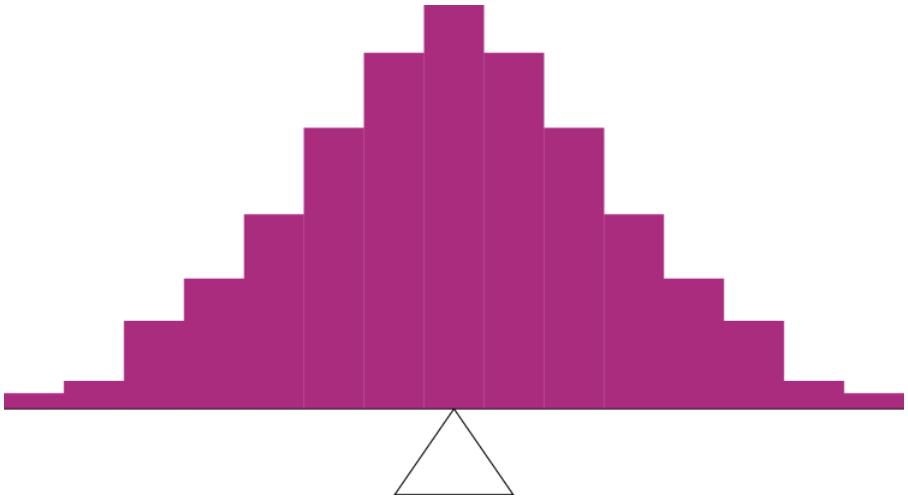


Figure 3.1.3: A distribution balanced on the tip of a triangle.

Image Credit: Judy Schmitt, from Cote et al, 2021

Figure 3.1.4 illustrates that the same distribution can't be balanced by placing the fulcrum to the left of center.

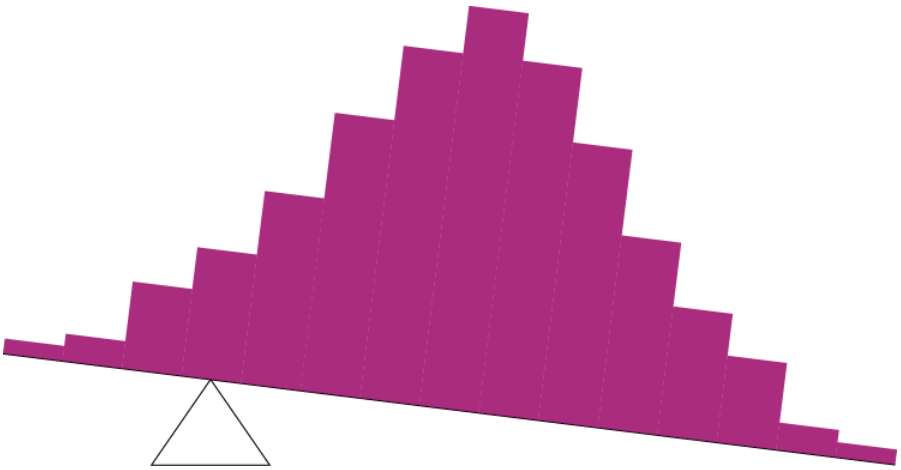


Figure 3.1.4: The distribution is not balanced.

Image Credit: Judy Schmitt, from Cote et al, 2021

Figure 3.1.5 shows an asymmetric distribution. To balance it, we cannot put the fulcrum halfway between the lowest and highest values (as we did in Figure 3.1.3). Placing the fulcrum at the “half way” point would cause it to tip towards the left.

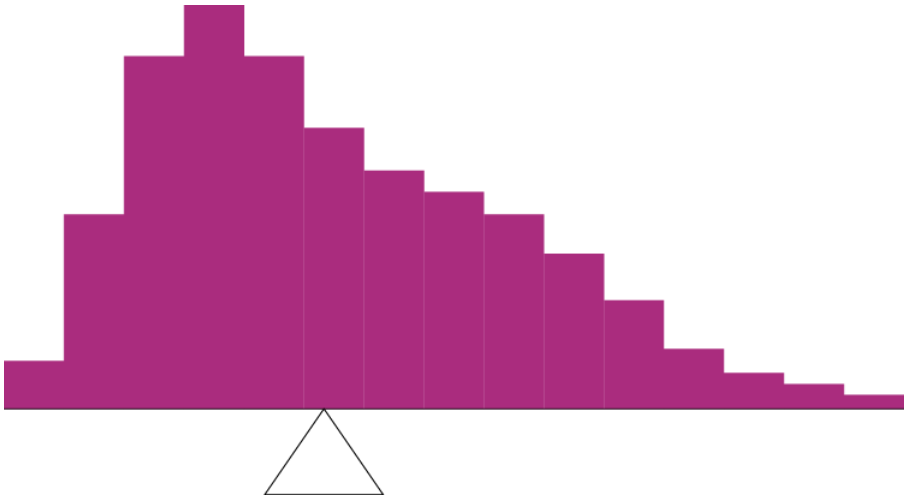


Figure 3.1.5 An asymmetric distribution balanced on the tip of a triangle.

Image Credit: Judy Schmitt, from Cote et al, 2021

This page titled 3.1: What is Central Tendency? is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **3.1: What is Central Tendency?** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

3.2: Measures of Central Tendency

In the previous section we saw that there are several ways to define central tendency. This section defines the three most common measures of central tendency: the mean, the median, and the mode. The relationships among these measures of central tendency and the definitions given in the previous section will probably not be obvious to you.

This section gives only the basic definitions of the mean, median and mode. A further discussion of the relative merits and proper applications of these statistics is presented in a later section.

Arithmetic Mean

The arithmetic mean is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers. The symbol "μ" (pronounced "mew") is used for the mean of a population. The symbol "$\overline{X}$" (pronounced "X-bar") is used for the mean of a sample. The formula for μ is shown below:

$$\mu = \frac{\sum X}{N} \tag{3.2.1}$$

where ΣX is the sum of all the numbers in the population and N is the number of numbers in the population.

The formula for $\overline{X}$ is essentially identical:

$$\overline{X} = \frac{\sum X}{N} \tag{3.2.2}$$

where ΣX is the sum of all the numbers in the sample and N is the number of numbers in the sample. The only distinction between these two equations is whether we are referring to the population (in which case we use the parameter μ) or a sample of that population (in which case we use the statistic $\overline{X}$).

As an example, the mean of the numbers 1, 2, 3, 6, 8 is 20/5 = 4 regardless of whether the numbers constitute the entire population or just a sample from the population.

Figure 3.2.1 shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season. The mean number of touchdown passes thrown is 20.45 as shown below.

$$\mu = \frac{\sum X}{N} = \frac{634}{31} = 20.45$$

37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20, 19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6

Figure 3.2.1: Number of touchdown passes.

Although the arithmetic mean is not the only "mean" (there is also a geometric mean, a harmonic mean, and many others that are all beyond the scope of this course), it is by far the most commonly used. Therefore, if the term "mean" is used without specifying whether it is the arithmetic mean, the geometric mean, or some other mean, it is assumed to refer to the arithmetic mean.

Median

The median is also a frequently used measure of central tendency. The median is the midpoint of a distribution: the same number of scores is above the median as below it. For the data in Figure 3.2.1, there are 31 scores. The 16th highest score (which equals 20) is the median because there are 15 scores below the 16th score and 15 scores above the 16th score. The median can also be thought of as the 50th percentile.

When there is an odd number of numbers, the median is simply the middle number. For example, the median of 2, 4, and 7 is 4. When there is an even number of numbers, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 7, 12 is:

$$\frac{4+7}{2} = 5.5$$

When there are numbers with the same values, each appearance of that value gets counted. For example, in the set of numbers 1, 3, 4, 4, 5, 8, and 9, the median is 4 because there are three numbers (1, 3, and 4) below it and three numbers (5, 8, and 9) above it. If we only counted 4 once, the median would incorrectly be calculated at 4.5 (4+5 divided by 2). When in doubt, writing out all of the numbers in order and marking them off one at a time from the top and bottom will always lead you to the correct answer.

Mode

The mode is the most frequently occurring value in the dataset. For the data in Figure 3.2.1, the mode is 18 since more teams (4) had 18 touchdown passes than any other number of touchdown passes. With continuous data, such as response time measured to many decimals, the frequency of each value is one since no two scores will be exactly the same (see discussion of continuous variables). Therefore the mode of continuous data is normally computed from a grouped frequency distribution. Table 3.2.1 shows a grouped frequency distribution for the target response time data. Since the interval with the highest frequency is 600-700, the mode is the middle of that interval (650). Though the mode is not frequently used for continuous data, it is nevertheless an important measure of central tendency as it is the only measure we can use on qualitative or categorical data.

Table 3.2.1: Grouped frequency distribution

Range	Frequency
500 - 600	3
600 - 700	6
700 - 800	5
800 - 900	5
900 - 1000	0
1000 - 1100	1

More on the Mean and Median

Figure 3.2.2 shows that the distribution balances at the mean of 6.8 and not at the median of 4. The relative advantages and disadvantages of the mean and median are discussed in the section “Comparing Measures” later in this chapter.

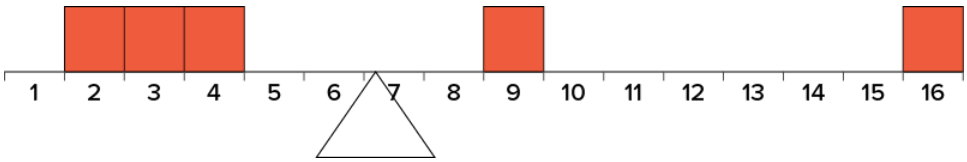


Figure 3.2.2: The distribution balances at the mean of 6.8 and not at the median of 4.0.

Image Credit: Judy Schmitt, from Cote et al, 2021

When a distribution is symmetric, then the mean and the median are the same. Consider the following distribution: 1, 3, 4, 5, 6, 7, 9. The mean and median are both 5. The mean, median, and mode are identical in the bell-shaped normal distribution.

Comparing Measures of Central Tendency

How do the various measures of central tendency compare with each other? For symmetric distributions, the mean and median, as is the mode except in bimodal distributions. Differences among the measures occur with skewed distributions. Figure 3.2.3 shows the distribution of 642 scores on an introductory psychology test. Notice this distribution has a slight positive skew.

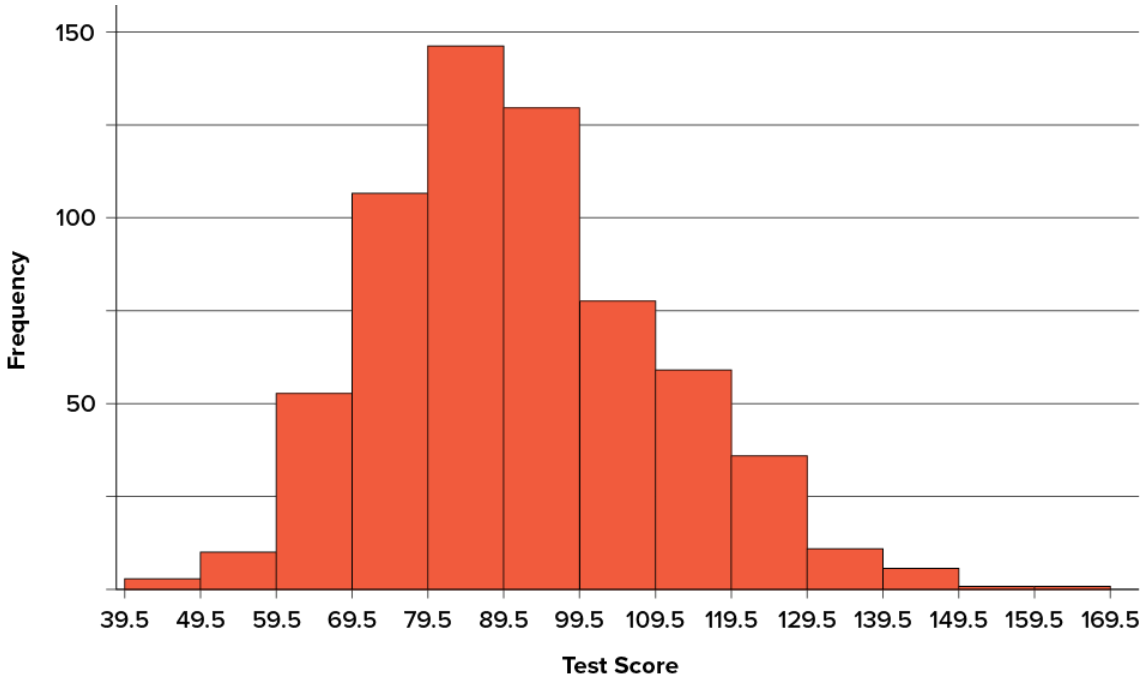


Figure 3.2.3: A distribution with a positive skew.

Image Credit: Judy Schmitt, from Cote et al, 2021

Measures of central tendency are shown in Table 3.2.3. Notice they do not differ greatly, with the exception that the mode is considerably lower than the other measures. When distributions have a positive skew, the mean is typically higher than the median, although it may not be in bimodal distributions. For these data, the mean of 91.58 is higher than the median of 90. This pattern holds true for any skew: the mode will remain at the highest point in the distribution, the median will be pulled slightly out into the skewed tail (the longer end of the distribution), and the mean will be pulled the farthest out. Thus, the mean is more sensitive to skew than the median or mode, and in cases of extreme skew, the mean may no longer be appropriate to use.

Table 3.2.3: Measures of central tendency for the test scores.

Measure	Value
Mode	84.00
Median	90.00
Mean	91.58

The distribution of baseball salaries (in 1994) shown in Figure 3.2.4 has a much more pronounced skew than the distribution in Figure 3.2.3.

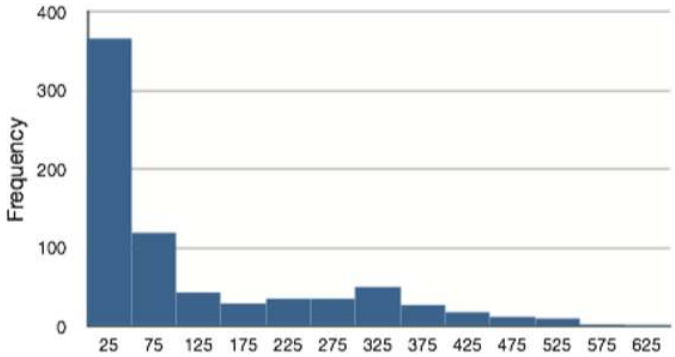


Figure **3.2.3**: A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars).

Image Credit: Judy Schmitt, from Cote et al, 2021

Table **3.2.4** shows the measures of central tendency for these data. The large skew results in very different values for these measures. No single measure of central tendency is sufficient for data such as these. If you were asked the very general question: "So, what do baseball players make?" and answered with the mean of $1,183,000, you would not have told the whole story since only about one third of baseball players make that much. If you answered with the mode of $250,000 or the median of $500,000, you would not be giving any indication that some players make many millions of dollars. Fortunately, there is no need to summarize a distribution with a single number. When the various measures differ, our opinion is that you should report the mean and median. Sometimes it is worth reporting the mode as well. In the media, the median is usually reported to summarize the center of skewed distributions. You will hear about median salaries and median prices of houses sold, etc. This is better than reporting only the mean, but it would be informative to hear more statistics.

Table **3.2.4**: Measures of central tendency for baseball salaries (in thousands of dollars).

Measure	Value
Mode	250
Median	500
Mean	1,183

This page titled 3.2: Measures of Central Tendency is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **3.2: Measures of Central Tendency** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

3.3: Spread and Variability

Variability refers to how “spread out” a group of scores is. To see what we mean by spread out, consider graphs in Figure 3.3.1. These graphs represent the scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, you can see that the distributions are quite different. Specifically, the scores on Quiz 1 are more densely packed and those on Quiz 2 are more spread out. The differences among students were much greater on Quiz 2 than on Quiz 1.

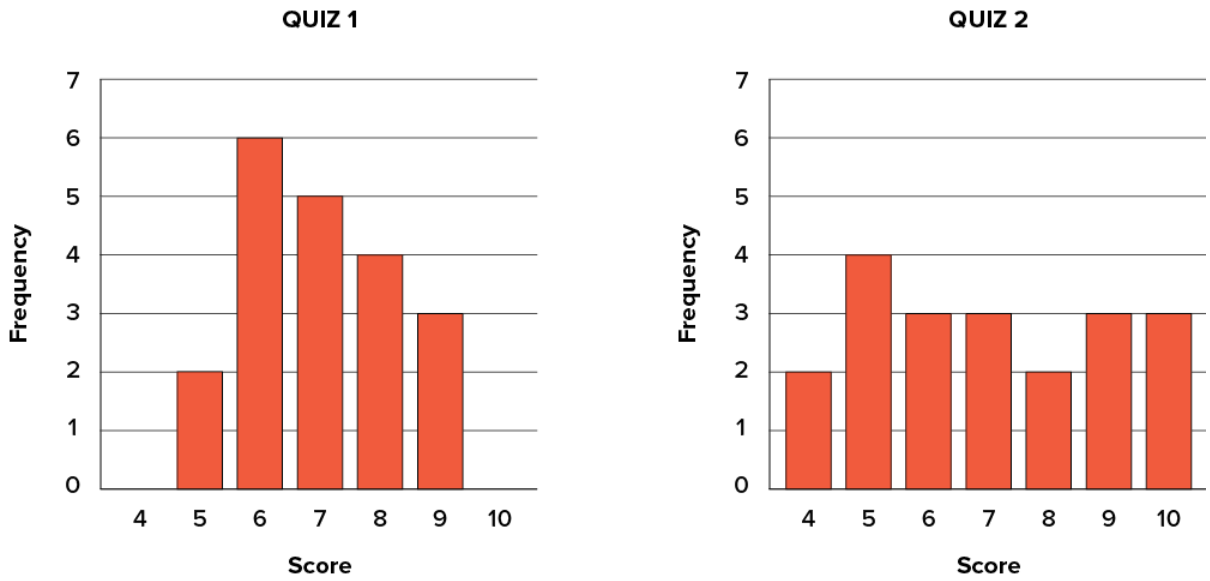


Figure 3.3.1: Bar chart of quiz one.

Image Credit: Judy Schmitt, from Cote et al, 2021

Range

The range is the simplest measure of variability to calculate, and one you have probably encountered many times in your life. The range is simply the highest score minus the lowest score. Let’s take a few examples. What is the range of the following group of numbers: 10, 2, 5, 6, 7, 3, 4? Well, the highest number is 10, and the lowest number is 2, so 10 - 2 = 8. The range is 8. Let’s take another example. Here’s a dataset with 10 numbers: 99, 45, 23, 67, 45, 91, 82, 78, 62, 51. What is the range? The highest number is 99 and the lowest number is 23, so 99 - 23 equals 76; the range is 76. Now consider the two quizzes shown in Figure 3.3.1 and Figure 3.3.2. On Quiz 1, the lowest score is 5 and the highest score is 9. Therefore, the range is 4. The range on Quiz 2 was larger: the lowest score was 4 and the highest score was 10. Therefore the range is 6.

The problem with using range is that it is extremely sensitive to outliers, and one number far away from the rest of the data will greatly alter the value of the range. For example, in the set of numbers 1, 3, 4, 4, 5, 8, and 9, the range is 8 (9 – 1).

However, if we add a single person whose score is nowhere close to the rest of the scores, say, 20, the range more than doubles from 8 to 19.

There are technically two different ways to describe range, known as **exclusive range** and **inclusive range.** The exclusive range is found with the simple formula used above, **h-l** (meaning highest number minus the lowest number). Due to the nature of this formula, you are literally excluding the end number of a range in the total count. For example, if you take 10-1=9, you are only accounting for the values between 1 and 10, but not 10 itself. Comparatively, the inclusive range adds back in that missing end number with the formula **h-l+1**. In this example you would subtract 10-1=9+1, to get 10 again, or essentially a complete count of all the values present in the range (included in the range). The inclusive range is very helpful in data science when wanting to count the number of rows in a spreadsheet, values in a dataset, or total number of participants in a study, to name a few ideas. If we used the exclusive range in these situations, we would come up one short of the true number of values present.

(The following section "Revisiting Percentiles", is borrowed and edited from Dr. Alisa Beyer's text, Chapter 6).

Revisiting Percentiles

The percentile rank of a score is the percentage of scores in the distribution that are lower than that score. Percentiles are useful for comparing values. For any score in the distribution, we can find its percentile rank by counting the number of scores in a distribution that are lower than that score and converting that number to a percentage of the total number of scores. Percentile ranks are often used to report the results of standardized tests of ability or achievement. If your percentile rank on a test of verbal ability were 40, for example, this would mean that you scored higher than 40% of the people who took the test.

(End of section borrowed from the Beyer text).

Interquartile Range

The interquartile range (IQR) is the range of the middle 50% of the scores in a distribution and is sometimes used to communicate where the bulk of the data in the distribution are located. It is computed as follows:

$$\text{IQR} = 75\text{th percentile} - 25\text{th percentile} \tag{3.3.1}$$

For Quiz 1, the 75th percentile is 8 and the 25th percentile is 6. The interquartile range is therefore 2. For Quiz 2, which has greater spread, the 75th percentile is 9, the 25th percentile is 5, and the interquartile range is 4.

Sum of Squares

Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution. Using the mean as the measure of the middle of the distribution, we can see how far, on average, each data point is from the center. The data from Quiz 1 are shown in Table 3.3.1. The mean score is 7.0 ($\Sigma X/N = 140/20 = 7$). Therefore, the column "$X - \overline{X}$" contains deviations (how far each score deviates from the mean), here calculated as the score minus 7. The column "$(X - \overline{X})^2$" has the "Squared Deviations" and is simply the previous column squared.

There are a few things to note about how Table 3.3.1 is formatted, as this is the format you will use to calculate variance (and, soon, standard deviation). The raw data scores (X) are always placed in the left-most column. This column is then summed at the bottom to facilitate calculating the mean (simply divided this number by the number of scores in the table). Once you have the mean, you can easily work your way down the middle column calculating the deviation scores. This column is also summed and has a very important property: it will always sum to 0 (or close to zero if you have rounding error due to many decimal places). This step is used as a check on your math to make sure you haven't made a mistake. If this column sums to 0, you can move on to filling in the third column of squared deviations. This column is summed as well and has its own name: the **Sum of Squares (abbreviated as** SS and given the formula $\sum (X - \overline{X})^2$). As we will see, the Sum of Squares appears again and again in different formulas – it is a very important value, and this table makes it simple to calculate without error.

Table 3.3.1: Calculation of Variance for Quiz 1 scores.

X	$X-\overline{X}$	$(X-\overline{X})^2$
9	2	4
9	2	4
9	2	4
8	1	1
8	1	1
8	1	1
8	1	1
7	0	0
7	0	0
7	0	0
7	0	0
7	0	0
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1
5	-2	4
5	-2	4
$\Sigma = 140$	$\Sigma = 0$	$\Sigma = 30$

Variance

Now that we have the Sum of Squares calculated, we can use it to compute our formal measure of average distance from the mean, the variance. The variance is defined as the average squared difference of the scores from the mean. We square the deviation scores because, as we saw in the Sum of Squares table, the sum of raw deviations is always 0, and there's nothing we can do mathematically without changing that.

The population parameter for variance is σ^2 ("sigma-squared") and is calculated as:

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \tag{3.3.2}$$

Notice that the numerator that formula is identical to the formula for Sum of Squares presented above with $\overline{X}$ replaced by μ. Thus, we can use the Sum of Squares table to easily calculate the numerator then simply divide that value by N to get variance. If we assume that the values in Table 3.3.1 represent the full population, then we can take our value of Sum of Squares and divide it by N to get our population variance:

$$\sigma^2 = \frac{30}{20} = 1.5$$

So, on average, scores in this population are 1.5 squared units away from the mean. This measure of spread is much more robust (a term used by statisticians to mean resilient or resistant to) outliers than the range, so it is a much more useful value to compute. Additionally, as we will see in future chapters, variance plays a central role in inferential statistics.

The sample statistic used to estimate the variance is s^2 ("s-squared"):

$$s^2 = \frac{\sum (X - \overline{X})^2}{N - 1} \tag{3.3.3}$$

This formula is very similar to the formula for the population variance with one change: we now divide by N–1 instead of N. The value N–1 has a special name: the **degrees of freedom (abbreviated as df)**. You don't need to understand in depth what degrees of freedom are (essentially they account for the fact that we have to use a sample statistic to estimate the mean ($\overline{X}$) before we estimate the variance) in order to calculate variance, but knowing that the denominator is called df provides a nice shorthand for the variance formula: SS/df.

Going back to the values in Table 3.3.1 and treating those scores as a sample, we can estimate the sample variance as:

$$s^2 = \frac{30}{20 - 1} = 1.58 \tag{3.3.4}$$

Notice that this value is slightly larger than the one we calculated when we assumed these scores were the full population. This is because our value in the denominator is slightly smaller, making the final value larger. In general, as your sample size N gets bigger, the effect of subtracting 1 becomes less and less. Comparing a sample size of 10 to a sample size of 1000; 10 – 1 = 9, or 90% of the original value, whereas 1000 – 1 = 999, or 99.9% of the original value. Thus, larger sample sizes will bring the estimate of the sample variance closer to that of the population variance. This is a key idea and principle in statistics that we will see over and over again: larger sample sizes better reflect the population.

Standard Deviation

The standard deviation is simply the square root of the variance. This is a useful and interpretable statistic because taking the square root of the variance (recalling that variance is the average squared difference) puts the standard deviation back into the original units of the measure we used. Thus, when reporting descriptive statistics in a study, scientists virtually always report mean and standard deviation. Standard deviation is therefore the most commonly used measure of spread for our purposes.

The population parameter for standard deviation is σ ("sigma"), which, intuitively, is the square root of the variance parameter σ^2 (on occasion, the symbols work out nicely that way). The formula is simply the formula for variance under a square root sign:

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}} \tag{3.3.5}$$

Back to our earlier example from Table 3.3.1:

$$\sigma = \sqrt{\frac{30}{20}} = \sqrt{1.5} = 1.22$$

The sample statistic follows the same conventions and is given as s:

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}} = \sqrt{\frac{SS}{df}} \tag{3.3.6}$$

The sample standard deviation from Table 3.3.1 is:

$$s = \sqrt{\frac{30}{20 - 1}} = \sqrt{1.58} = 1.26$$

The standard deviation is an especially useful measure of variability when the distribution is normal or approximately normal because the proportion of the distribution within a given number of standard deviations from the mean can be calculated. For example, 68% of the distribution is within one standard deviation (above and below) of the mean and approximately 95% of the distribution is within two standard deviations of the mean. Therefore, if you had a normal distribution with a mean of 50 and a

standard deviation of 10, then 68% of the distribution would be between 50 - 10 = 40 and 50 +10 =60. Similarly, about 95% of the distribution would be between 50 - 2 x 10 = 30 and 50 + 2 x 10 = 70.

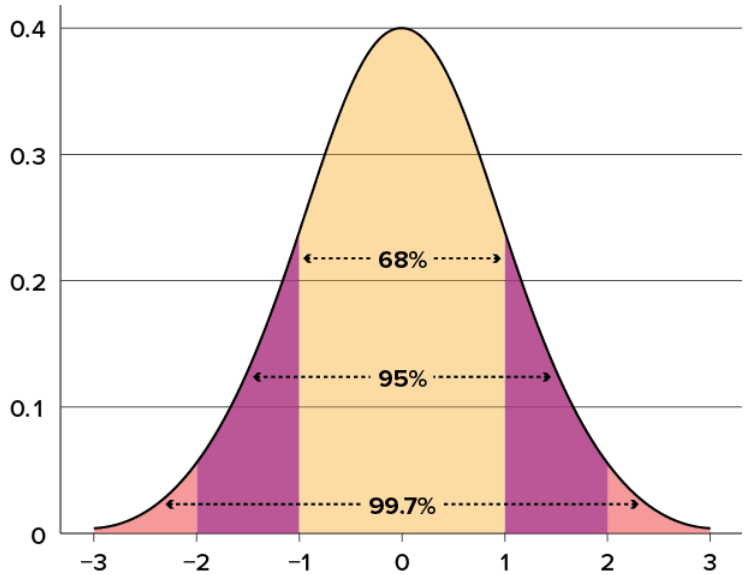


Figure 3.3.3: Percentages of the normal distribution

Image Credit: Judy Schmitt, from Cote et al, 2021

Figure 3.3.4 shows two normal distributions. The red distribution has a mean of 40 and a standard deviation of 5; the blue distribution has a mean of 60 and a standard deviation of 10. For the red distribution, 68% of the distribution is between 45 and 55; for the blue distribution, 68% is between 50 and 70. Notice that as the standard deviation gets smaller, the distribution becomes much narrower, regardless of where the center of the distribution (mean) is. Figure 3.3.5 presents several more examples of this effect.

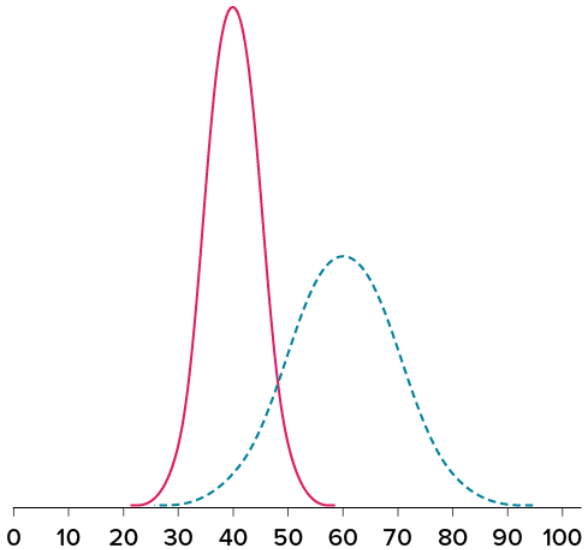


Figure 3.3.4: Normal distributions with standard deviations of 5 and 10.

Image Credit: Judy Schmitt, from Cote et al, 2021

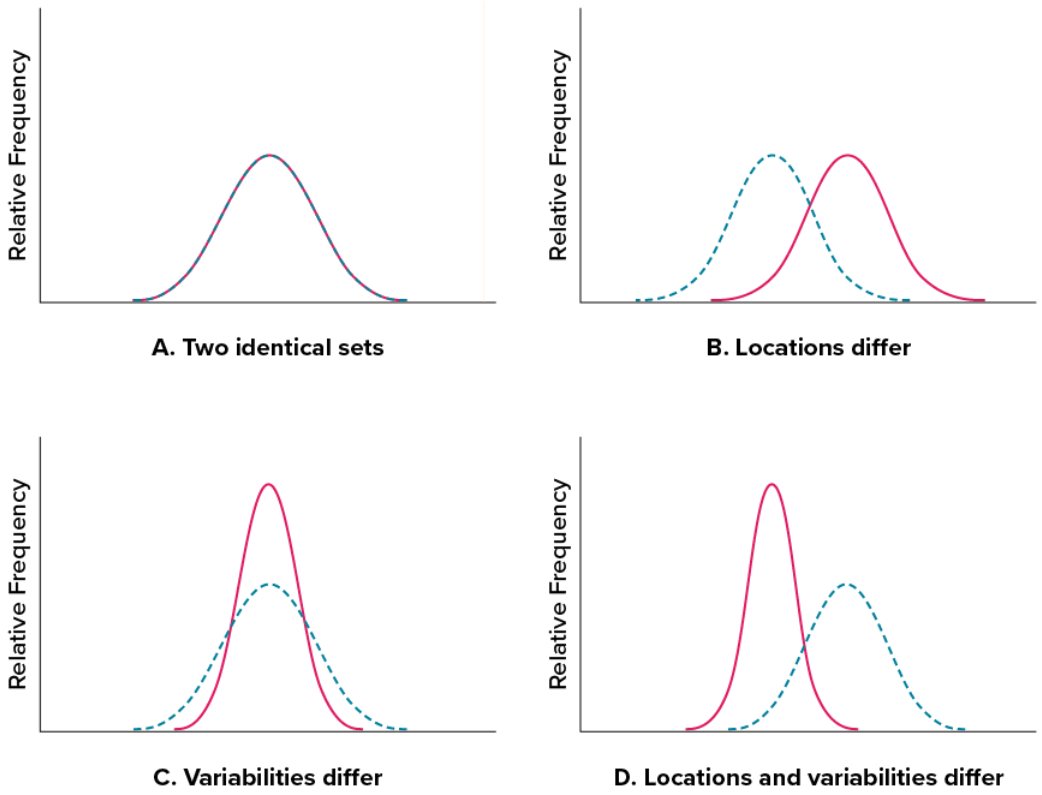


Figure **3.3.5**: Differences between two datasets.

Image Credit: Judy Schmitt, from Cote et al, 2021

This page titled 3.3: Spread and Variability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **3.3: Spread and Variability** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

3.E: Measures of Central Tendency and Spread (Exercises)

1. If the mean time to respond to a stimulus is much higher than the median time to respond, what can you say about the shape of the distribution of response times?

Answer:

If the mean is higher, that means it is farther out into the right-hand tail of the distribution. Therefore, we know this distribution is positively skewed.

2. Compare the mean, median, and mode in terms of their sensitivity to extreme scores.
3. Your younger brother comes home one day after taking a science test. He says that some- one at school told him that “60% of the students in the class scored above the median test grade.” What is wrong with this statement? What if he had said “60% of the students scored above the mean?”

Answer:

The median is defined as the value with 50% of scores above it and 50% of scores below it; therefore, 60% of score cannot fall above the median. If 60% of scores fall above the mean, that would indicate that the mean has been pulled down below the value of the median, which means that the distribution is negatively skewed

4. Make up three data sets with 5 numbers each that have:
   a. the same mean but different standard deviations.
   b. the same mean but different medians.
   c. the same median but different means.
5. Compute the population mean and population standard deviation for the following scores (remember to use the Sum of Squares table): 5, 7, 8, 3, 4, 4, 2, 7, 1, 6

Answer:

$\mu = 4.80, \sigma^2 = 2.36$

6. For the following problem, use the following scores: 5, 8, 8, 8, 7, 8, 9, 12, 8, 9, 8, 10, 7, 9, 7, 6, 9, 10, 11, 8
   a. Create a histogram of these data. What is the shape of this histogram?
   b. How do you think the three measures of central tendency will compare to each other in this dataset?
   c. Compute the sample mean, the median, and the mode
   d. Draw and label lines on your histogram for each of the above values. Do your results match your predictions?
7. Compute the range, sample variance, and sample standard deviation for the following scores: 25, 36, 41, 28, 29, 32, 39, 37, 34, 34, 37, 35, 30, 36, 31, 31

Answer:

range = 16, $s^2 = 18.40$, $s = 4.29$

8. Using the same values from problem 7, calculate the range, sample variance, and sample standard deviation, but this time include 65 in the list of values. How did each of the three values change?
9. Two normal distributions have exactly the same mean, but one has a standard deviation of 20 and the other has a standard deviation of 10. How would the shapes of the two distributions compare?

Answer:

If both distributions are normal, then they are both symmetrical, and having the same mean causes them to overlap with one another. The distribution with the standard deviation of 10 will be narrower than the other distribution

10. Compute the sample mean and sample standard deviation for the following scores: -8, -4, -7, -6, -8, -5, -7, -9, -2, 0

This page titled 3.E: Measures of Central Tendency and Spread (Exercises) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **3.E: Measures of Central Tendency and Spread (Exercises)** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

4: Describing Data using Distributions and Graphs

4.1: Graphing Qualitative Variables
4.2: Graphing Quantitative Variables
4.E: Describing Data using Distributions and Graphs (Exercises)

This page titled 4: Describing Data using Distributions and Graphs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

4.1: Graphing Qualitative Variables

When Apple Computer introduced the iMac computer in August 1998, the company wanted to learn whether the iMac was expanding Apple's market share. Was the iMac just attracting previous Macintosh owners? Or was it purchased by newcomers to the computer market and by previous Windows users who were switching over? To find out, 500 iMac customers were interviewed. Each customer was categorized as a previous Macintosh owner, a previous Windows owner, or a new computer purchaser.

This section examines graphical methods for displaying the results of the interviews. We'll learn some general lessons about how to graph data that fall into a small number of categories. A later section will consider how to graph numerical data in which each observation is represented by a number in some range. The key point about the qualitative data that occupy us in the present section is that they do not come with a pre-established ordering (the way numbers are ordered). For example, there is no natural sense in which the category of previous Windows users comes before or after the category of previous Macintosh users. This situation may be contrasted with quantitative data, such as a person's weight. People of one weight are naturally ordered with respect to people of a different weight.

Frequency Tables

All of the graphical methods shown in this section are derived from frequency tables. Table 1 shows a frequency table for the results of the iMac study; it shows the frequencies of the various response categories. It also shows the relative frequencies, which are the proportion of responses in each category. For example, the relative frequency for "none" of 0.17 = 85/500.

Table **4.1.1**: Frequency Table for the iMac Data.

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1

Pie Charts

The pie chart in Figure **4.1.1** shows the results of the iMac study. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. This is simply the relative frequency multiplied by 100. Although most iMac purchasers were Macintosh owners, Apple was encouraged by the 12% of purchasers who were former Windows users, and by the 17% of purchasers who were buying a computer for the first time.

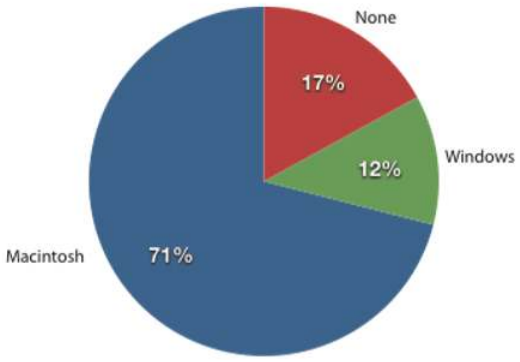


Figure **4.1.1**: Pie chart of iMac purchases illustrating frequencies of previous computer ownership.

Pie charts are effective for displaying the relative frequencies of a small number of categories. They are not recommended, however, when you have a large number of categories. Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or experiments. In an influential book on the use of graphs, Edward Tufte asserted "The only worse design than a pie chart is several of them." Here is another important point about pie charts. If they are based on a small

number of observations, it can be misleading to label the pie slices with percentages. For example, if just 5 people had been interviewed by Apple Computers, and 3 were former Windows users, it would be misleading to display a pie chart with the Windows slice showing 60%. With so few people interviewed, such a large percentage of Windows users might easily have occurred since chance can cause large errors with small samples. In this case, it is better to alert the user of the pie chart to the actual numbers involved. The slices should therefore be labeled with the actual frequencies observed (e.g., 3) instead of with percentages.

Bar charts Bar charts can also be used to represent frequencies of different categories. A bar chart of the iMac purchases is shown in Figure 4.1.2. Frequencies are shown on the Yaxis and the type of computer previously owned is shown on the X-axis. Typically, the Y-axis shows the number of observations in each category rather than the percentage of observations in each category as is typical in pie charts.

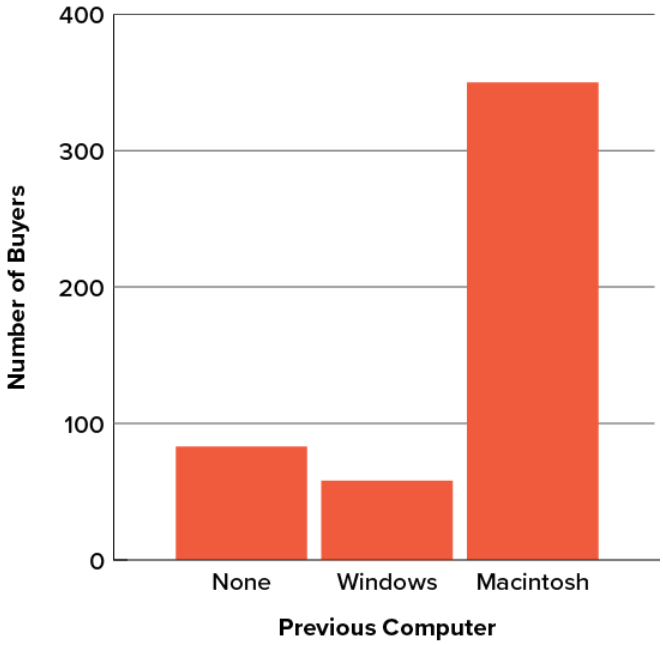


Figure 4.1.2: Bar chart of iMac purchases as a function of previous computer ownership.

Image Credit: Judy Schmitt, from Cote et al, 2021.

Comparing Distributions Often we need to compare the results of different surveys, or of different conditions within the same overall survey. In this case, we are comparing the “distributions” of responses between the surveys or conditions. Bar charts are often excellent for illustrating differences between two distributions. Figure 4.1.3 shows the number of people playing card games at the Yahoo web site on a Sunday and on a Wednesday in the spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.

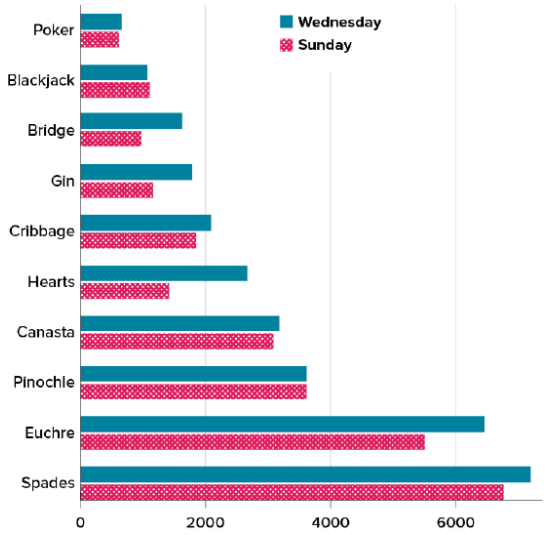


Figure 4.1.3: A bar chart of the number of people playing different card games on Sunday and Wednesday.

Image Credit: Judy Schmitt, from Cote et al, 2021.

The bars in Figure 4.1.3 are oriented horizontally rather than vertically. The horizontal format is useful when you have many categories because there is more room for the category labels. We'll have more to say about bar charts when we consider numerical quantities later in this chapter.

Some graphical mistakes to avoid Don't get fancy! People sometimes add features to graphs that don't help to convey their information. For example, 3-dimensional bar charts such as the one shown in Figure 4.1.4 are usually not as effective as their two-dimensional counterparts.

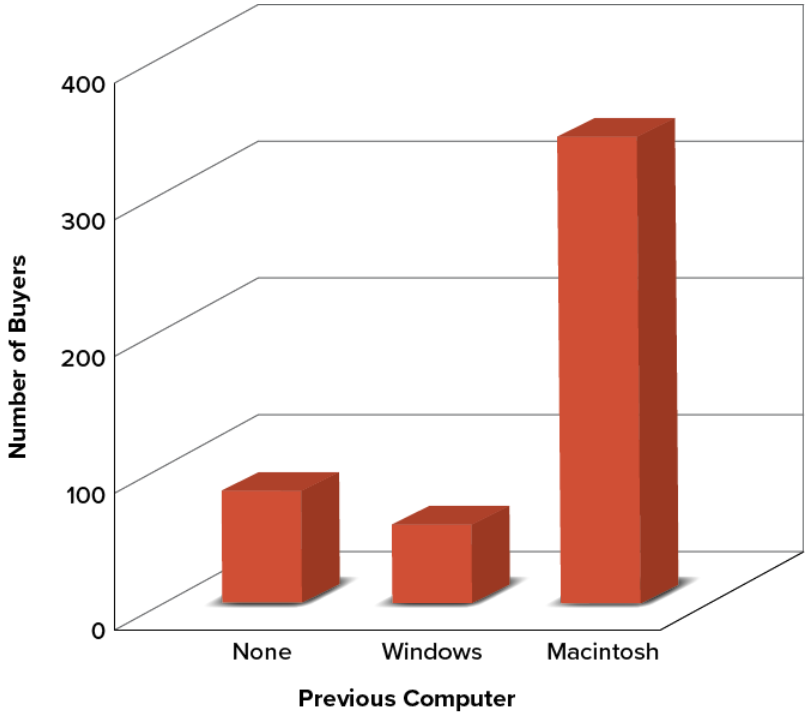


Figure 4.1.4: A three-dimensional version of Figure 4.1.2

Image Credit: Judy Schmitt, from Cote et al, 2021.

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure 4.1.5 presents the iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure 4.1.5 is misleading because the viewer's attention will be captured by areas. The areas can exaggerate the size differences between the groups. In terms of percentages, the ratio of previous Macintosh owners to previous Windows owners is about 6 to 1. But the ratio of the two areas in Figure 4.1.5 is about 35 to 1. A biased person wishing to hide the fact that many Windows owners purchased iMacs would be tempted to use Figure 4.1.5 instead of Figure 4.1.2! Edward Tufte coined the term “lie factor” to refer to the ratio of the size of the effect shown in a graph to the size of the effect shown in the data. He suggests that lie factors greater than 1.05 or less than 0.95 produce unacceptable distortion.

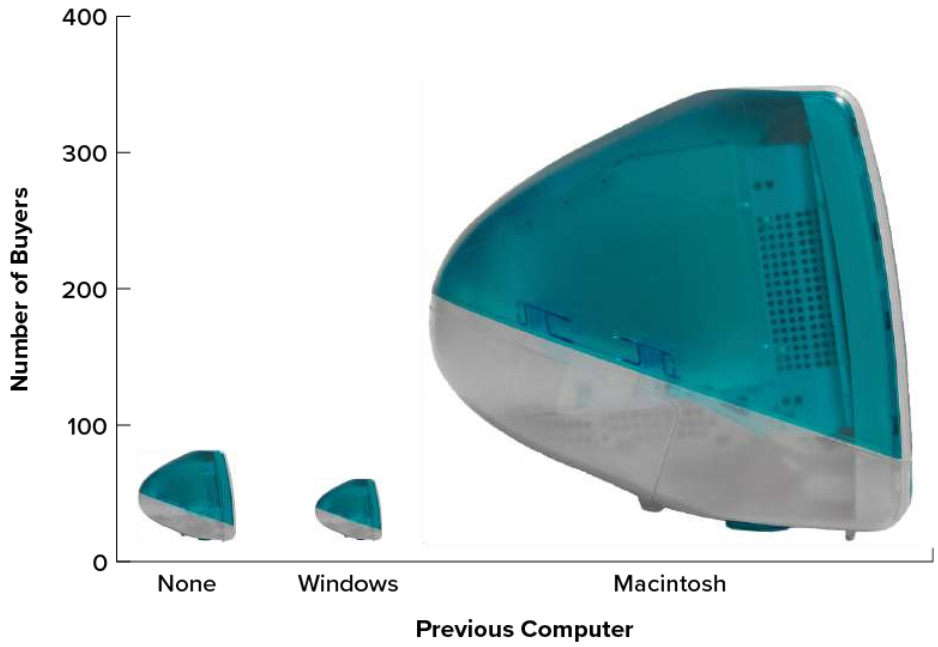


Figure 4.1.5: A redrawing of Figure 4.1.2 with a lie factor greater than 8.

Image Credit: Judy Schmitt, from Cote et al, 2021.

Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the Y-axis, representing the least number of cases that could have occurred in a category. Normally, but not always, this number should be zero. Figure 6 shows the iMac data with a baseline of 50. Once again, the differences in areas suggests a different story than the true differences in percentages. The number of Windows-switchers seems minuscule compared to its true value of 12%.

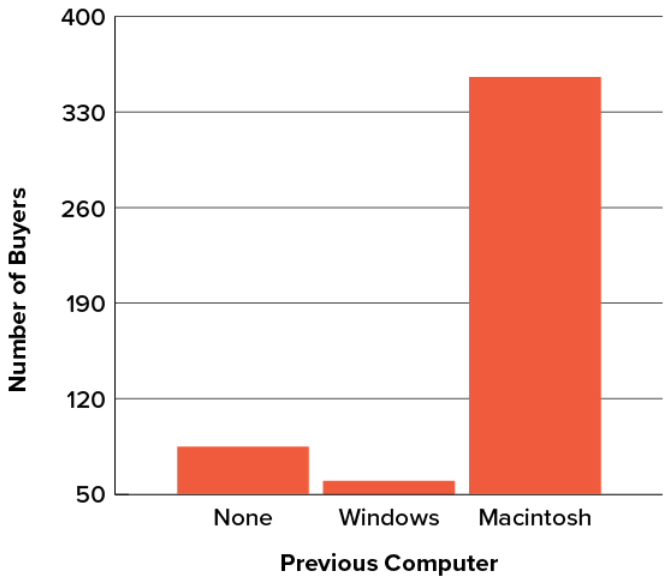


Figure 4.1.6: A redrawing of Figure 4.1.2 with a baseline of 50.

Image Credit: Judy Schmitt, from Cote et al, 2021.

Finally, we note that it is a serious mistake to use a line graph when the X-axis contains merely qualitative variables. A line graph is essentially a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). Figure 4.1.7 inappropriately shows a line graph of the card game data from Yahoo. The drawback to Figure 7 is that it gives the false impression that the games are naturally ordered in a numerical way when, in fact, they are ordered alphabetically.

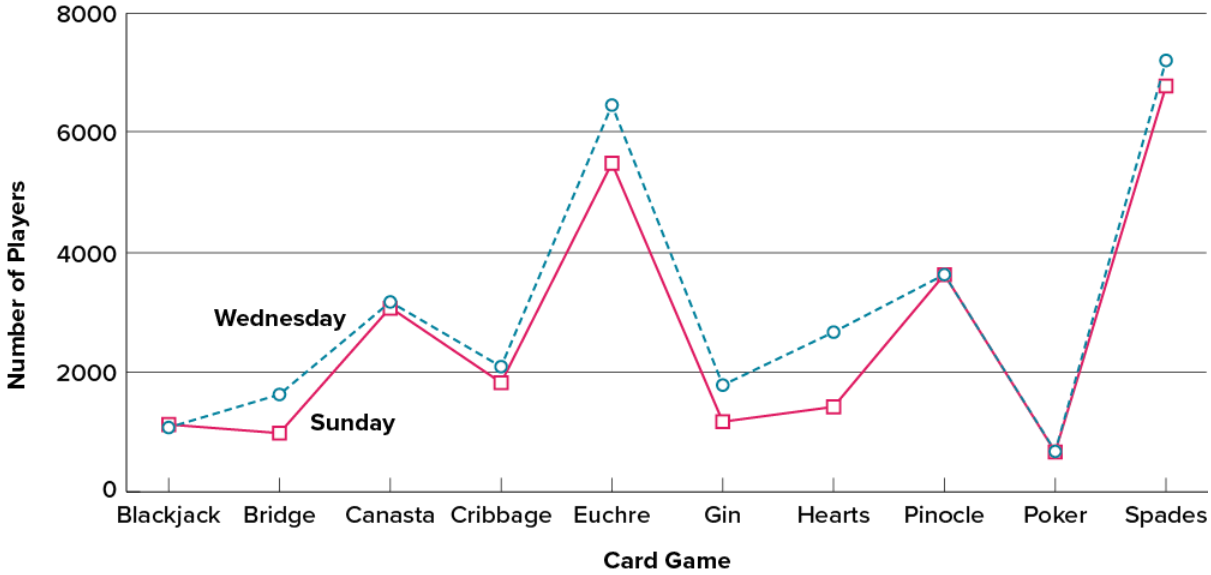


Figure 4.1.7: A line graph used inappropriately to depict the number of people playing different card games on Sunday and Wednesday.

Image Credit: Judy Schmitt, from Cote et al, 2021.

Summary

Pie charts and bar charts can both be effective methods of portraying qualitative data. Bar charts are better when there are more than just a few categories and for comparing two or more distributions. Be careful to avoid creating misleading graphs.

This page titled 4.1: Graphing Qualitative Variables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **2.1: Graphing Qualitative Variables** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

4.2: Graphing Quantitative Variables

As discussed in the section on variables in Chapter 1, quantitative variables are variables measured on a numeric scale. Height, weight, response time, subjective rating of pain, temperature, and score on an exam are all examples of quantitative variables. Quantitative variables are distinguished from categorical (sometimes called qualitative) variables such as favorite color, religion, city of birth, favorite sport in which there is no ordering or measuring involved.

There are many types of graphs that can be used to portray distributions of quantitative variables. The upcoming sections cover the following types of graphs:

- histograms
- frequency polygons
- bar charts
- line graphs
- dot plots
- scatter plots (discussed in a different chapter)

Some graph types are best-suited for small to moderate amounts of data, whereas others such as histograms are best suited for large amounts of data. Graph types such as box plots are good at depicting differences between distributions. Scatter plots are used to show the relationship between two variables.

Histograms

A histogram is a graphical method for displaying the shape of a distribution. It is particularly useful when there are a large number of observations. We begin with an example consisting of the scores of 642 students on a psychology test. The test consists of 197 items each graded as “correct” or “incorrect.” The students' scores ranged from 46 to 167.

The first step is to create a frequency table. Unfortunately, a simple frequency table would be too big, containing over 100 rows. To simplify the table, we group scores together as shown in Table 4.2.1.

Table 4.2.1: Grouped Frequency Distribution of Psychology Test Scores

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

To create this table, the range of scores was broken into intervals, called class intervals. The first interval is from 39.5 to 49.5, the second from 49.5 to 59.5, etc. Next, the number of scores falling into each interval was counted to obtain the class frequencies. There are three scores in the first interval, 10 in the second, etc.

Class intervals of width 10 provide enough detail about the distribution to be revealing without making the graph too “choppy.” More information on choosing the widths of class intervals is presented later in this section. Placing the limits of the class intervals midway between two numbers (e.g., 49.5) ensures that every score will fall in an interval rather than on the boundary between intervals.

In a histogram, the class frequencies are represented by bars. The height of each bar corresponds to its class frequency. A histogram of these data is shown in Figure 4.2.8.

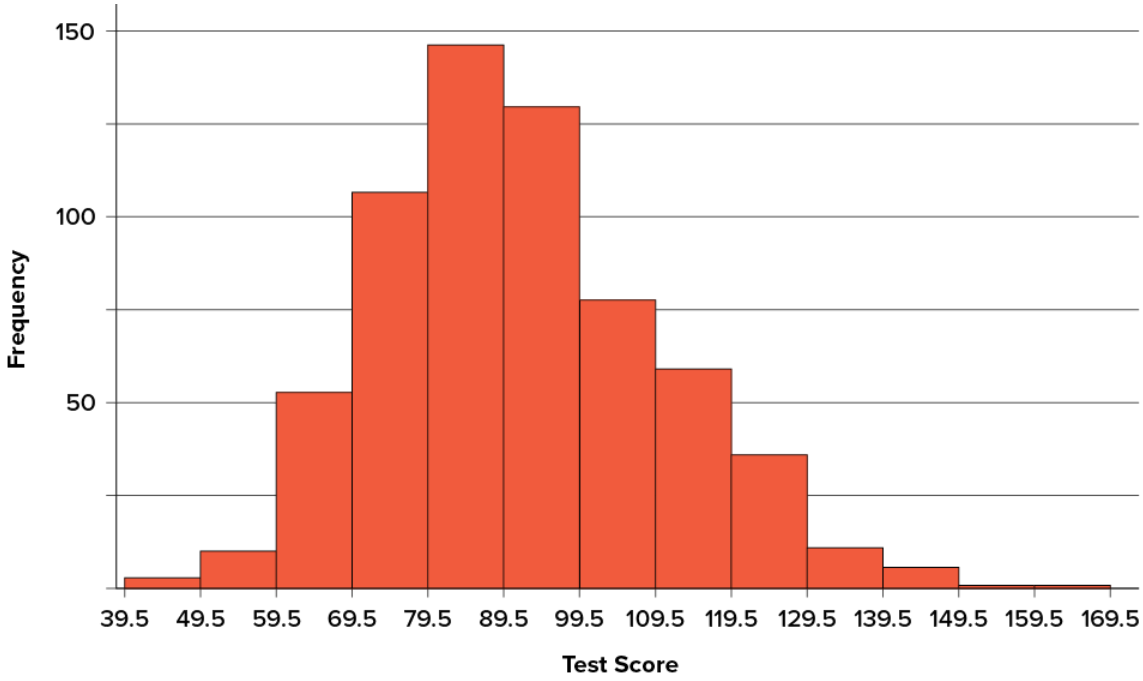


Figure 4.2.8: Histogram of scores on a psychology test.

Image Credit: Judy Schmitt, from Cote et al, 2021.

The histogram makes it plain that most of the scores are in the middle of the distribution, with fewer scores in the extremes. You can also see that the distribution is not symmetric: the scores extend to the right farther than they do to the left. The distribution is therefore said to be skewed. (We'll have more to say about shapes of distributions in Chapter 3.)

In our example, the observations are whole numbers. Histograms can also be used when the scores are measured on a more continuous scale such as the length of time (in milliseconds) required to perform a task. In this case, there is no need to worry about fence sitters since they are improbable. (It would be quite a coincidence for a task to require exactly 7 seconds, measured to the nearest thousandth of a second.) We are therefore free to choose whole numbers as boundaries for our class intervals, for example, 4000, 5000, etc. The class frequency is then the number of observations that are greater than or equal to the lower bound, and strictly less than the upper bound. For example, one interval might hold times from 4000 to 4999 milliseconds. Using whole numbers as boundaries avoids a cluttered appearance, and is the practice of many computer programs that create histograms. Note also that some computer programs label the middle of each interval rather than the end points.

Histograms can be based on relative frequencies instead of actual frequencies. Histograms based on relative frequencies show the proportion of scores in each interval rather than the number of scores. In this case, the Y-axis runs from 0 to 1 (or somewhere in between if there are no extreme proportions). You can change a histogram based on frequencies to one based on relative frequencies by (a) dividing each class frequency by the total number of observations, and then (b) plotting the quotients on the Y-axis (labeled as proportion).

There is more to be said about the widths of the class intervals, sometimes called bin widths. Your choice of bin width determines the number of class intervals. This decision, along with the choice of starting point for the first interval, affects the shape of the histogram. The best advice is to experiment with different choices of width, and to choose a histogram according to how well it communicates the shape of the distribution.

Frequency Polygons

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful for comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.

To create a frequency polygon, start just as for histograms, by choosing a class interval. Then draw an X-axis representing the values of the scores in your data. Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class. Draw the Y-axis to indicate the frequency of each class. Place a point in the middle of each class interval at the height corresponding to its frequency. Finally, connect the points. You should include one class interval below the lowest value in your data and one above the highest value. The graph will then touch the X-axis on both sides.

A frequency polygon for 642 psychology test scores shown in Figure 4.2.8 was constructed from the frequency table shown in Table 4.2.2.

Table 4.2.2: Frequency Distribution of Psychology Test Scores

Lower Limit	Upper Limit	Count	Cumulative Count
29.5	39.5	0	0
39.5	49.5	3	3
49.5	59.5	10	13
59.5	69.5	53	66
69.5	79.5	107	173
79.5	89.5	147	320
89.5	99.5	130	450
99.5	109.5	78	528
109.5	119.5	59	587
119.5	129.5	36	623
129.5	139.5	11	634
139.5	149.5	6	640
149.5	159.5	1	641
159.5	169.5	1	642
169.5	170.5	0	642

The first label on the X-axis is 35. This represents an interval extending from 29.5 to 39.5. Since the lowest test score is 46, this interval has a frequency of 0. The point labeled 45 represents the interval from 39.5 to 49.5. There are three scores in this interval. There are 147 scores in the interval that surrounds 85.

You can easily discern the shape of the distribution from Figure 4.2.9. Most of the scores are between 65 and 115. It is clear that the distribution is not symmetric inasmuch as good scores (to the right) trail off more gradually than poor scores (to the left). In the terminology of Chapter 3 (where we will study shapes of distributions more systematically), the distribution is skewed.

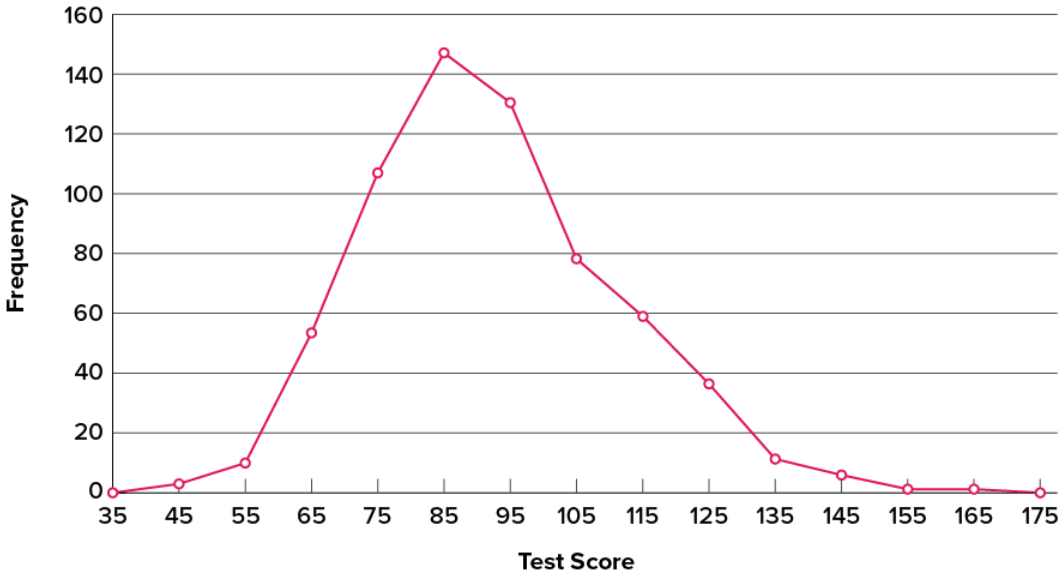


Figure 4.2.9: Frequency polygon for the psychology test scores.

Image Credit: Judy Schmitt, from Cote et al, 2021.

A cumulative frequency polygon for the same test scores is shown in Figure 4.2.10. The graph is the same as before except that the Y value for each point is the number of students in the corresponding class interval plus all numbers in lower intervals. For example, there are no scores in the interval labeled “35,” three in the interval “45,” and 10 in the interval “55.” Therefore, the Y value corresponding to “55” is 13. Since 642 students took the test, the cumulative frequency for the last interval is 642.

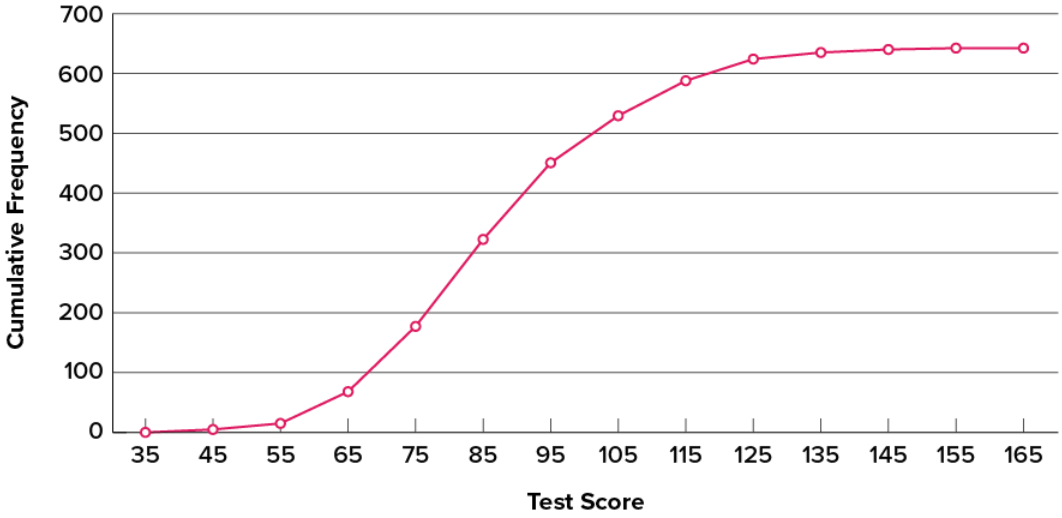


Figure 4.2.10: Cumulative frequency polygon for the psychology test scores.

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets. Figure 2.1.3 provides an example. The data come from a task in which the goal is to move a computer cursor to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial. The two distributions (one for each target) are plotted together in Figure 4.2.11. The figure shows that, although there is some overlap in times, it generally took longer to move the cursor to the small target than to the large one.

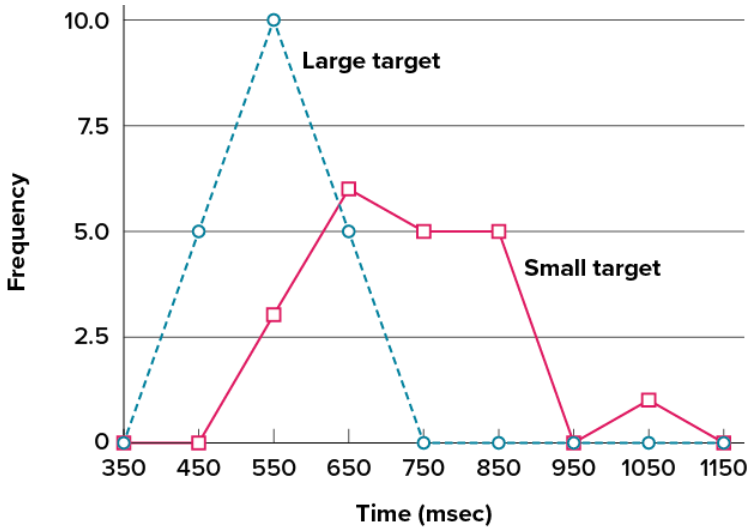


Figure 4.2.11: Overlaid frequency polygons.

It is also possible to plot two cumulative frequency distributions in the same graph. This is illustrated in Figure 4.2.12 using the same data from the cursor task. The difference in distributions for the two targets is again evident.

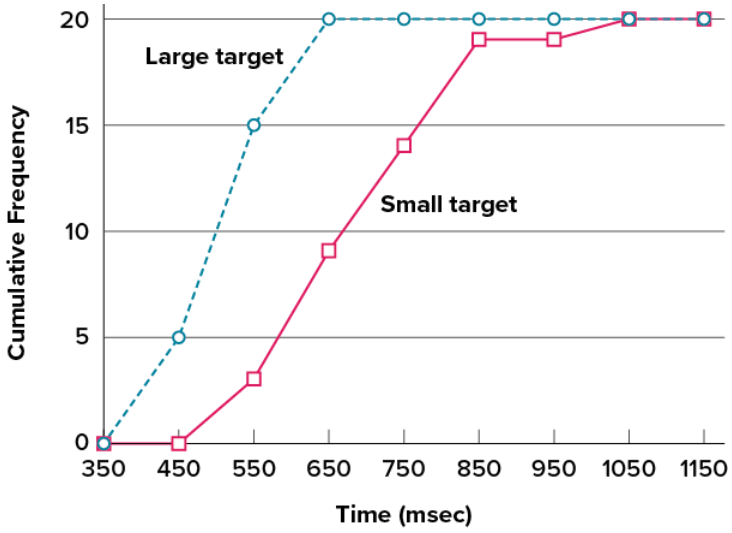


Figure 4.2.12: Overlaid cumulative frequency polygons.

Bar Charts

In the section on qualitative variables, we saw how bar charts could be used to illustrate the frequencies of different categories. For example, the bar chart shown in Figure 4.2.19 shows how many purchasers of iMac computers were previous Macintosh users, previous Windows users, and new computer purchasers.

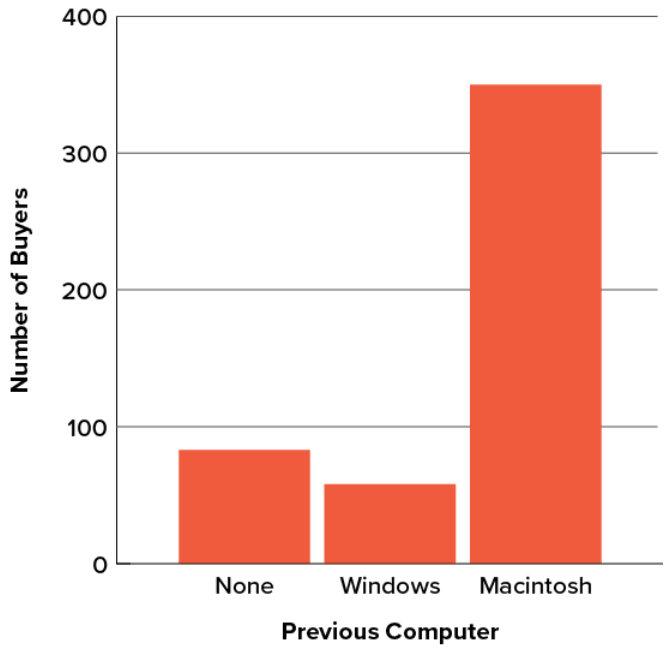


Figure 4.2.19: iMac buyers as a function of previous computer ownership.

Image Credit: Judy Schmitt, from Cote et al, 2021.

In this section we show how bar charts can be used to present other kinds of quantitative information, not just frequency counts. The bar chart in Figure 4.2.20 shows the percent increases in the Dow Jones, Standard and Poor 500 (S & P), and Nasdaq stock indexes from May 24th 2000 to May 24th 2001. Notice that both the S & P and the Nasdaq had "negative increases" which means that they decreased in value. In this bar chart, the Y-axis is not frequency but rather the signed quantity percentage increase.

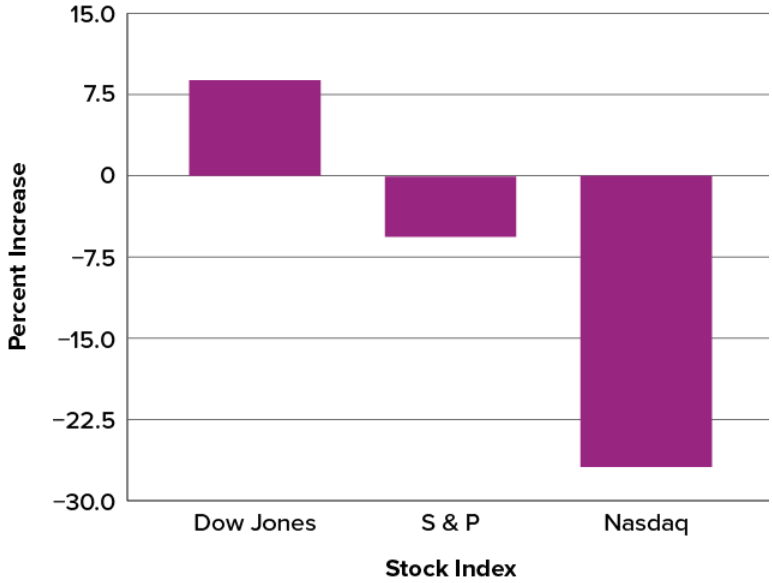


Figure 4.2.20: Percent increase in three stock indexes from May 24th 2000 to May 24th 2001.

Image Credit: Judy Schmitt, from Cote et al, 2021.

Bar charts are particularly effective for showing change over time. Figure 4.2.21, for example, shows the percent increase in the Consumer Price Index (CPI) over four three-month periods. The fluctuation in inflation is apparent in the graph.

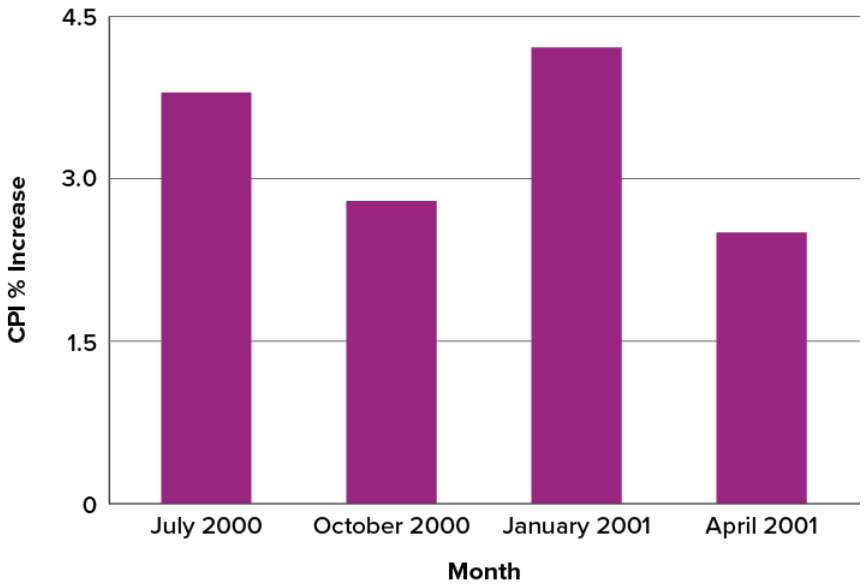


Figure 4.2.21: Percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

Image Credit: Judy Schmitt, from Cote et al, 2021.

Bar charts are often used to compare the means of different experimental conditions. Figure 2.1.4 shows the mean time it took one of us (DL) to move the cursor to either a small target or a large target. On average, more time was required for small targets than for large ones.

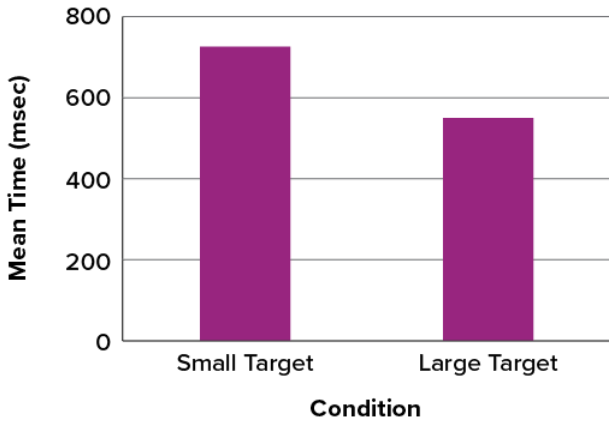


Figure 4.2.22: Bar chart showing the means for the two conditions.

Image Credit: Judy Schmitt, from Cote et al, 2021.

The section on qualitative variables presented earlier in this chapter discussed the use of bar charts for comparing distributions. Some common graphical mistakes were also noted. The earlier discussion applies equally well to the use of bar charts to display quantitative variables.

Line Graphs

A line graph is a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). For example, Figure 4.2.24 was presented in the section on bar charts and shows changes in the Consumer Price Index (CPI) over time.

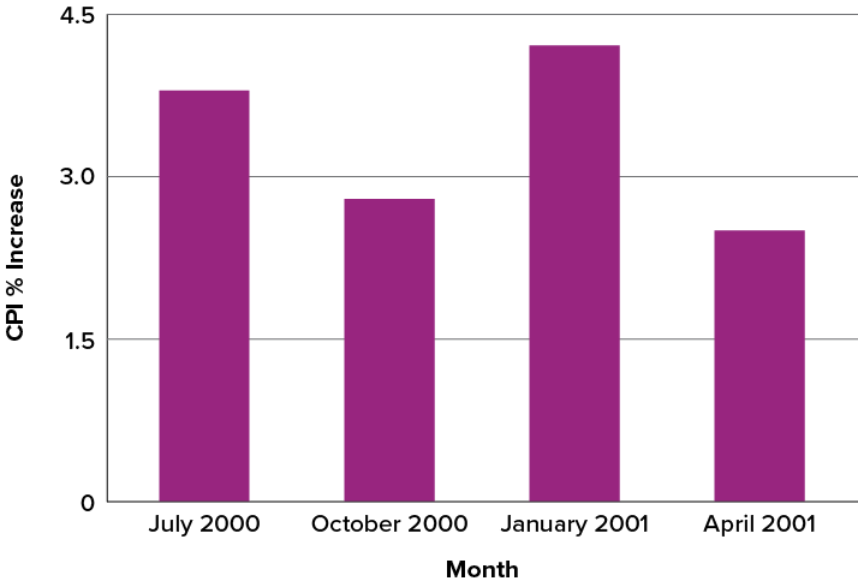


Figure 4.2.24: A bar chart of the percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

Image Credit: Judy Schmitt, from Cote et al, 2021.

A line graph of these same data is shown in Figure 4.2.25. Although the figures are similar, the line graph emphasizes the change from period to period.

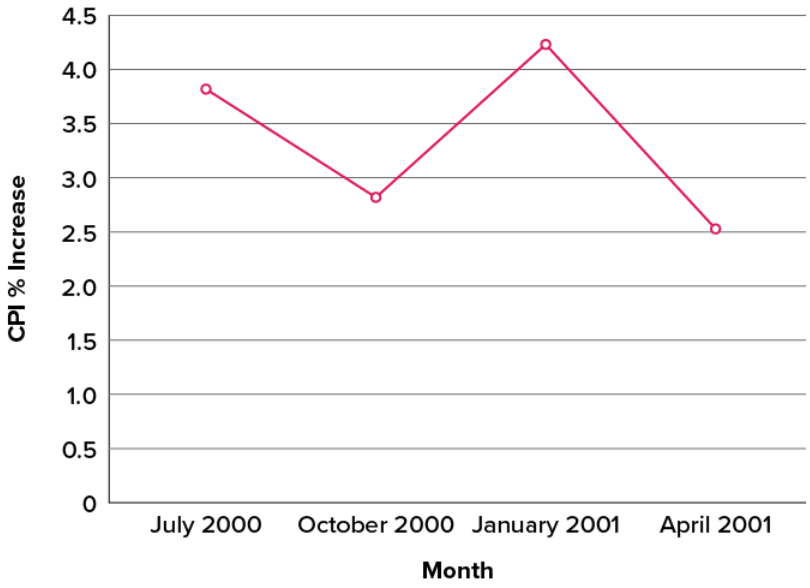


Figure 4.2.25: A line graph of the percent change in the CPI over time. Each point represents percent increase for the three months ending at the date indicated.

Image Credit: Judy Schmitt, from Cote et al, 2021.

Line graphs are appropriate only when both the X- and Y-axes display ordered (rather than qualitative) variables. Although bar charts can also be used in this situation, line graphs are generally better at comparing changes over time. Figure 4.2.26, for example, shows percent increases and decreases in five components of the CPI. The figure makes it easy to see that medical costs had a steadier progression than the other components. Although you could create an analogous bar chart, its interpretation would not be as easy.

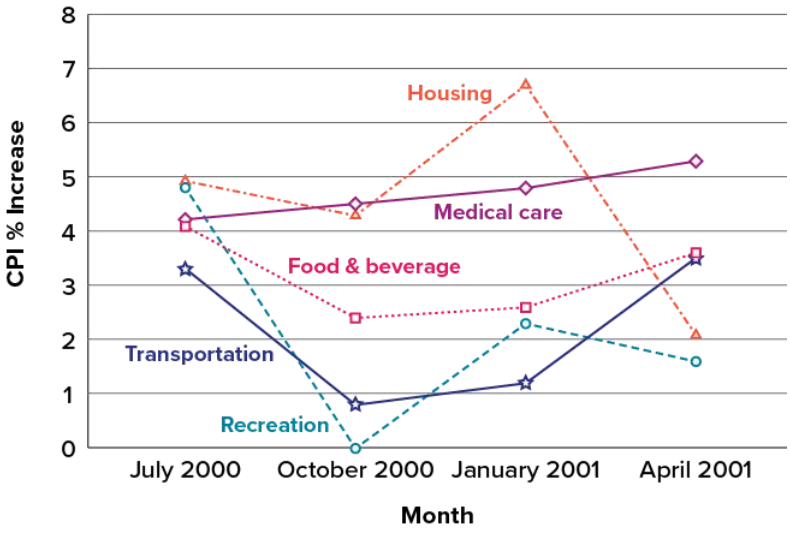


Figure 4.2.26: A line graph of the percent change in five components of the CPI over time.

Image Credit: Judy Schmitt, from Cote et al, 2021.

Let us stress that it is **misleading** to use a line graph when the X-axis contains merely qualitative variables.

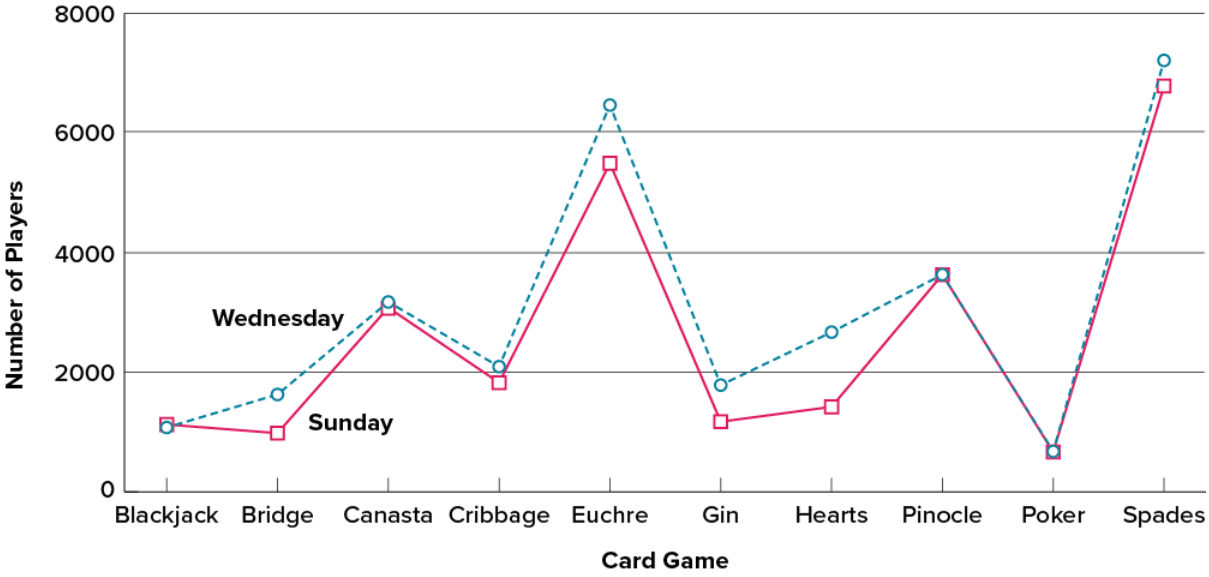


Image Credit: Judy Schmitt, from Cote et al, 2021.

The Shape of Distribution

Finally, it is useful to present discussion on how we describe the shapes of distributions, which we will revisit in the next chapter to learn how different shapes affect our numerical descriptors of data and distributions.

The primary characteristic we are concerned about when assessing the shape of a distribution is whether the distribution is symmetrical or skewed. A symmetrical distribution, as the name suggests, can be cut down the center to form 2 mirror images. Although in practice we will never get a perfectly symmetrical distribution, we would like our data to be as close to symmetrical as possible for reasons we delve into in Chapter 3. Many types of distributions are symmetrical, but by far the most common and pertinent distribution at this point is the normal distribution, shown in Figure 4.2.28. Notice that although the symmetry is not perfect (for instance, the bar just to the right of the center is taller than the one just to the left), the two sides are roughly the same shape. The normal distribution has a single peak, known as the center, and two tails that extend out equally, forming what is known as a bell shape or bell curve.

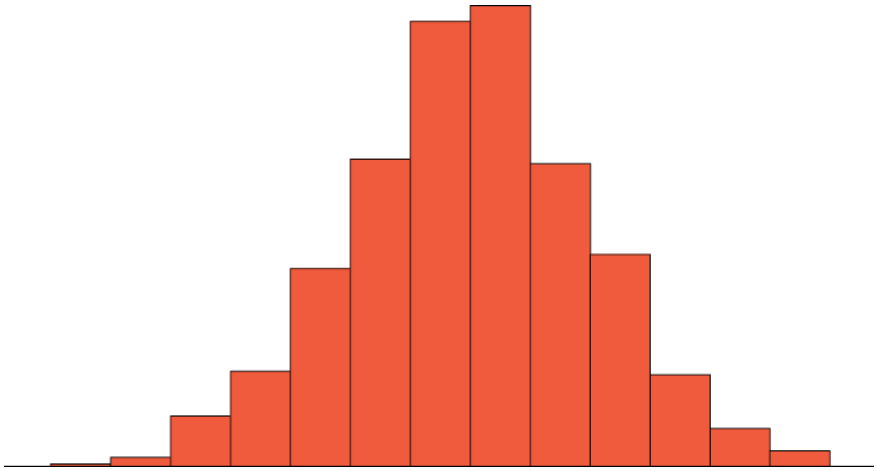


Figure 4.2.28: A symmetrical distribution

Image Credit: Judy Schmitt, from Cote et al, 2021.

Symmetrical distributions can also have multiple peaks. Figure 4.2.29 shows a bimodal distribution, named for the two peaks that lie roughly symmetrically on either side of the center point. As we will see in the next chapter, this is not a particularly desirable characteristic of our data, and, worse, this is a relatively difficult characteristic to detect numerically. Thus, it is important to visualize your data before moving ahead with any formal analyses.

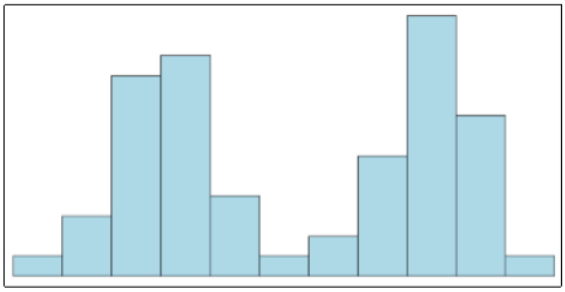
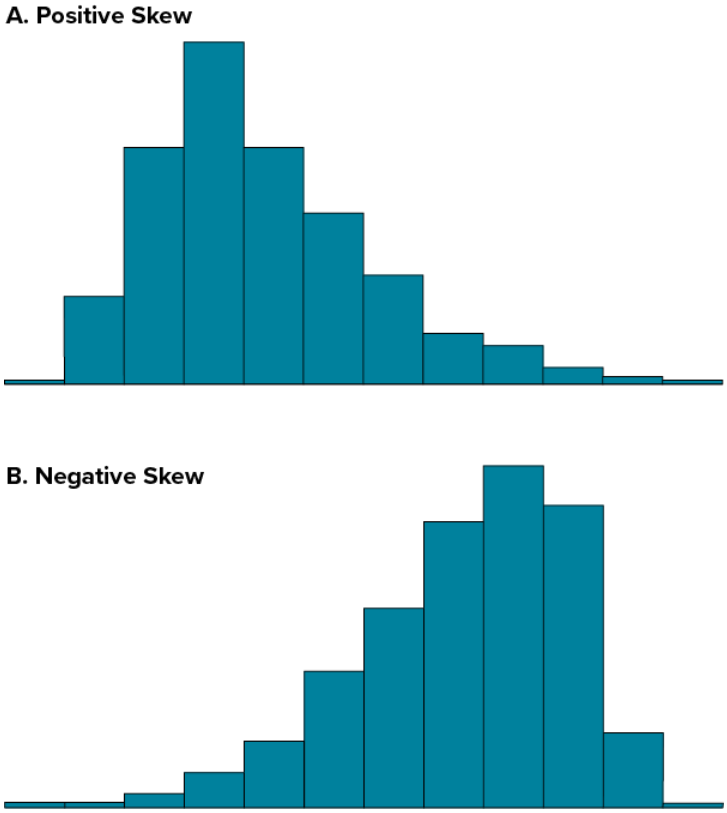


Figure 4.2.29: A bimodal distribution.

Distributions that are not symmetrical also come in many forms, more than can be described here. The most common asymmetry to be encountered is referred to as skew, in which one of the two tails of the distribution is disproportionately longer than the other. This property can affect the value of the averages we use in our analyses and make them an inaccurate representation of our data, which causes many problems.

Skew can either be positive or negative (also known as right or left, respectively), based on which tail is longer. It is very easy to get the two confused at first; many students want to describe the skew by where the bulk of the data (larger portion of the histogram, known as the body) is placed, but the correct determination is based on which tail is longer. You can think of the tail as an arrow: whichever direction the arrow is pointing is the direction of the skew. Figures 4.2.30 and 4.2.31 show positive (right) and negative (left) skew, respectively.



Figures 4.2.30 and 4.2.31

Image Credit: Judy Schmitt, from Cote et al, 2021.

This page titled 4.2: Graphing Quantitative Variables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **2.2: Graphing Quantitative Variables** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

4.E: Describing Data using Distributions and Graphs (Exercises)

1. Name some ways to graph quantitative variables and some ways to graph qualitative variables.

Answer:

Qualitative variables are displayed using pie charts and bar charts. Quantitative variables are displayed as box plots, histograms, etc.

2. Given the following data, construct a pie chart and a bar chart. Which do you think is the more appropriate or useful way to display the data?

Favorite Movie Genre	Frequency
Comedy	14
Horror	9
Romance	8
Action	12

3. Pretend you are constructing a histogram for describing the distribution of salaries for individuals who are 40 years or older, but are not yet retired.
   a. What is on the Y-axis? Explain.
   b. What is on the X-axis? Explain.
   c. What would be the probable shape of the salary distribution? Explain why.

Answer:

[You do not need to draw the histogram, only describe it below]

a. The Y-axis would have the frequency or proportion because this is always the case in histograms
b. The X-axis has income, because this is out quantitative variable of interest
c. Because most income data are positively skewed, this histogram would likely be skewed positively too

4. A graph appears below showing the number of adults and children who prefer each type of soda. There were 130 adults and kids surveyed. Discuss some ways in which the graph below could be improve

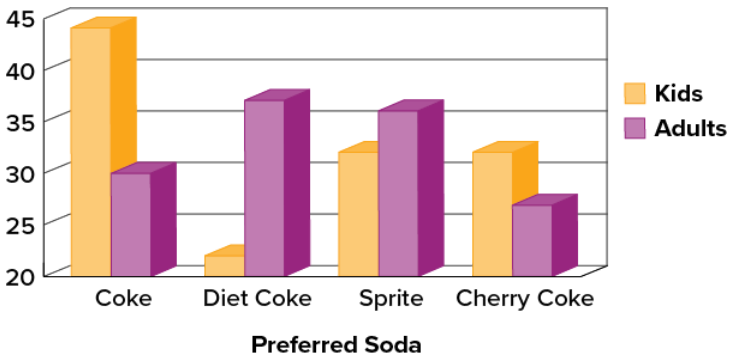


Image Credit: Judy Schmitt, from Cote et al, 2021.

4. Create a histogram of the following data representing how many shows children said they watch each day:

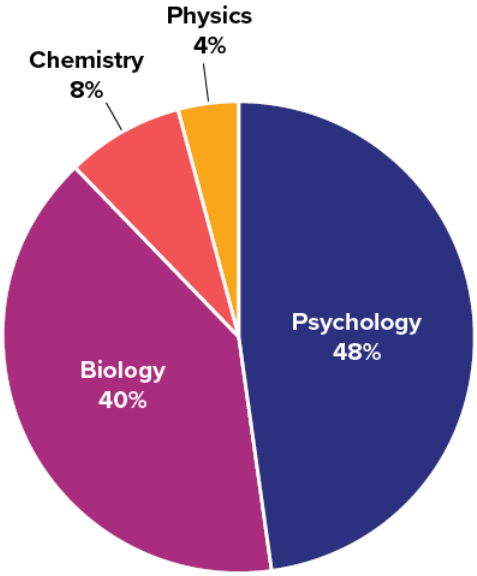
Number of TV Shows	Frequency
0	2
1	18
2	36
3	7
4	3

6. Explain the differences between bar charts and histograms. When would each be used?

Answer:

In bar charts, the bars do not touch; in histograms, the bars do touch. Bar charts are appropriate for qualitative variables, whereas histograms are better for quantitative variables.

8. Draw a histogram of a distribution that is
   a. Negatively skewed
   b. Symmetrical
   c. Positively skewed
9. Based on the pie chart below, which was made from a sample of 300 students, construct a frequency table of college majors.



Major	Frequency
Psychology	144
Biology	120
Chemistry	24
Physics	12

10. Create a histogram of the following data. Label the tails and body and determine if it is skewed (and direction, if so) or symmetrical.

Hours worked per week	Proportion
0 -10	4
10 -20	8
20 - 30	11
30 - 40	51
40 - 50	12
50 - 60	9
60+	5

This page titled 4.E: Describing Data using Distributions and Graphs (Exercises) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **2.E: Describing Data using Distributions and Graphs (Exercises)** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

5: Z-scores and the Standard Normal Distribution

5.1: Normal Distributions
5.2: Z-scores
5.3: Z-scores and the Area under the Curve
5.E: Z-scores and the Standard Normal Distribution (Exercises)

This page titled 5: Z-scores and the Standard Normal Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

5.1: Normal Distributions

The normal distribution is the most important and most widely used distribution in statistics. It is sometimes called the “bell curve,” although the tonal qualities of such a bell would be less than pleasing. It is also called the “Gaussian curve” of Gaussian distribution after the mathematician Karl Friedrich Gauss.

Strictly speaking, it is not correct to talk about “the normal distribution” since there are many normal distributions. Normal distributions can differ in their means and in their standard deviations. Figure 1 shows three normal distributions. The green (left-most) distribution has a mean of -3 and a standard deviation of 0.5, the distribution in red (the middle distribution) has a mean of 0 and a standard deviation of 1, and the distribution in black (right-most) has a mean of 2 and a standard deviation of 3. These as well as all other normal distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails. What is consistent about all normal distribution is the shape and the proportion of scores within a given distance along the x-axis. We will focus on the Standard Normal Distribution (also known as the Unit Normal Distribution), which has a mean of 0 and a standard deviation of 1 (i.e. the red distribution in Figure **5.1.1**).

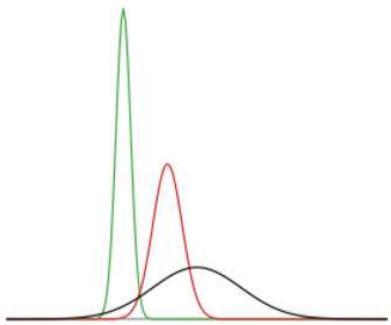


Figure **5.1.1**: Normal distributions differing in mean and standard deviation.

Seven features of normal distributions are listed below.

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).
6. 68% of the area of a normal distribution is within one standard deviation of the mean.
7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.

These properties enable us to use the normal distribution to understand how scores relate to one another within and across a distribution. But first, we need to learn how to calculate the standardized score than make up a standard normal distribution.

This page titled 5.1: Normal Distributions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri’s Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **4.1: Normal Distributions** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

5.2: Z-scores

A z-score is a standardized version of a raw score (x) that gives information about the relative location of that score within its distribution. The formula for converting a raw score into a z-score is:

$$z = \frac{x - \mu}{\sigma} \tag{5.2.1}$$

for values from a population and for values from a sample:

$$z = \frac{x - \overline{X}}{s} \tag{5.2.2}$$

As you can see, z-scores combine information about where the distribution is located (the mean/center) with how wide the distribution is (the standard deviation/spread) to interpret a raw score (x). Specifically, z-scores will tell us how far the score is away from the mean in units of standard deviations and in what direction.

The value of a z-score has two parts: the sign (positive or negative) and the magnitude (the actual number). The sign of the z-score tells you in which half of the distribution the z-score falls: a positive sign (or no sign) indicates that the score is above the mean and on the right hand-side or upper end of the distribution, and a negative sign tells you the score is below the mean and on the left-hand side or lower end of the distribution. The magnitude of the number tells you, in units of standard deviations, how far away the score is from the center or mean. The magnitude can take on any value between negative and positive infinity, but for reasons we will see soon, they generally fall between -3 and 3.

Let's look at some examples. A z-score value of -1.0 tells us that this z-score is 1 standard deviation (because of the magnitude 1.0) below (because of the negative sign) the mean. Similarly, a z-score value of 1.0 tells us that this z-score is 1 standard deviation above the mean. Thus, these two scores are the same distance away from the mean but in opposite directions. A z-score of -2.5 is two-and-a-half standard deviations below the mean and is therefore farther from the center than both of the previous scores, and a z -score of 0.25 is closer than all of the ones before. In Unit 2, we will learn to formalize the distinction between what we consider "close" to the center or "far" from the center. For now, we will use a rough cut-off of 1.5 standard deviations in either direction as the difference between close scores (those within 1.5 standard deviations or between z = -1.5 and z = 1.5) and extreme scores (those farther than 1.5 standard deviations – below z = -1.5 or above z = 1.5).

We can also convert raw scores into z-scores to get a better idea of where in the distribution those scores fall. Let's say we get a score of 68 on an exam. We may be disappointed to have scored so low, but perhaps it was just a very hard exam. Having information about the distribution of all scores in the class would be helpful to put some perspective on ours. We find out that the class got an average score of 54 with a standard deviation of 8. To find out our relative location within this distribution, we simply convert our test score into a z-score.

$$z = \frac{X - \mu}{\sigma} = \frac{68 - 54}{8} = 1.75$$

We find that we are 1.75 standard deviations above the average, above our rough cut off for close and far. Suddenly our 68 is looking pretty good!

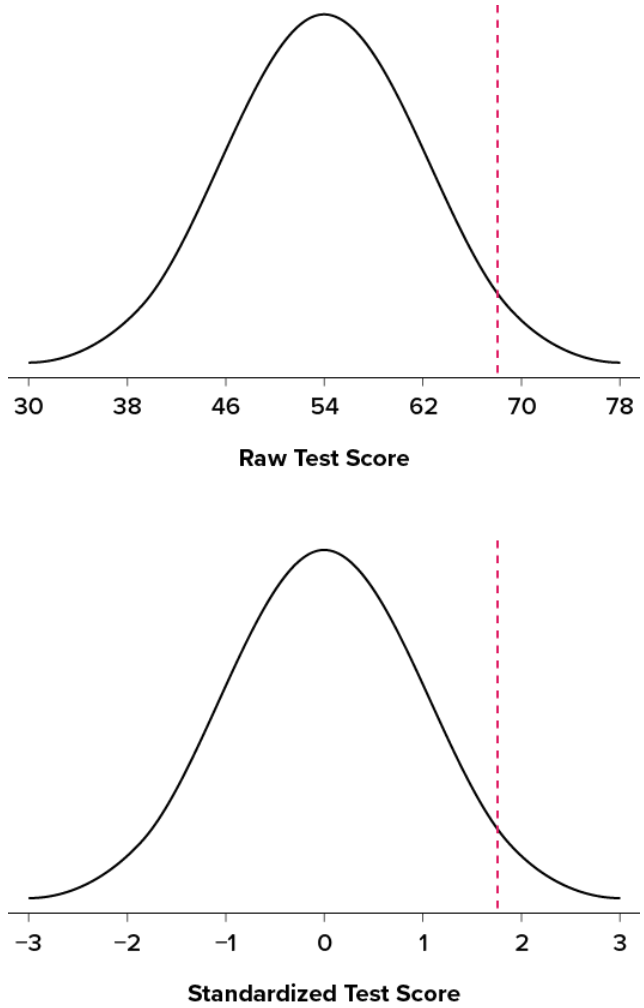


Figure 5.2.1: Raw and standardized versions of a single score

Image Credit: Judy Schmitt, from Cote et al, 2021

Figure 5.2.1 shows both the raw score and the z-score on their respective distributions. Notice that the red line indicating where each score lies is in the same relative spot for both. This is because transforming a raw score into a z-score does not change its relative location, it only makes it easier to know precisely where it is.

Z-scores are also useful for comparing scores from different distributions. Let's say we take the SAT and score 501 on both the math and critical reading sections. Does that mean we did equally well on both? Scores on the math portion are distributed normally with a mean of 511 and standard deviation of 120, so our z-score on the math section is

$$z_{math} = \frac{501 - 511}{120} = -0.08$$

which is just slightly below average (note that use of "math" as a subscript; subscripts are used when presenting multiple versions of the same statistic in order to know which one is which and have no bearing on the actual calculation). The critical reading section has a mean of 495 and standard deviation of 116, so

$$z_{CR} = \frac{501 - 495}{116} = 0.05$$

So even though we were almost exactly average on both tests, we did a little bit better on the critical reading portion relative to other people.

Finally, z-scores are incredibly useful if we need to combine information from different measures that are on different scales. Let's say we give a set of employees a series of tests on things like job knowledge, personality, and leadership. We may want to combine

these into a single score we can use to rate employees for development or promotion, but look what happens when we take the average of raw scores from different scales, as shown in Table 5.2.1:

Table 5.2.1: Raw test scores on different scales (ranges in parentheses).

Raw Scores	Job Knowledge (0 – 100)	Personality (1 –5)	Leadership (1 – 5)	Average
Employee 1	98	4.2	1.1	34.43
Employee 2	96	3.1	4.5	34.53
Employee 3	97	2.9	3.6	34.50

Because the job knowledge scores were so big and the scores were so similar, they overpowered the other scores and removed almost all variability in the average. However, if we standardize these scores into z-scores, our averages retain more variability and it is easier to assess differences between employees, as shown in Table 5.2.2.

Table 5.2.2: Standardized scores.

z-Scores	Job Knowledge (0 – 100)	Personality (1 –5)	Leadership (1 – 5)	Average
Employee 1	1.00	1.14	-1.12	0.34
Employee 2	-1.00	-0.43	0.81	-0.20
Employee 3	0.00	-0.71	0.30	-0.14

Setting the scale of a distribution

Another convenient characteristic of z-scores is that they can be converted into any "scale" that we would like. Here, the term scale means how far apart the scores are (their spread) and where they are located (their central tendency). This can be very useful if we don't want to work with negative numbers or if we have a specific range we would like to present. The formulas for transforming z to x are:

$$x = z\sigma + \mu \tag{5.2.3}$$

for a population and

$$x = zs + \overline{X} \tag{5.2.4}$$

for a sample. Notice that these are just simple rearrangements of the original formulas for calculating z from raw scores.

Let's say we create a new measure of intelligence, and initial calibration finds that our scores have a mean of 40 and standard deviation of 7. Three people who have scores of 52, 43, and 34 want to know how well they did on the measure. We can convert their raw scores into z-scores:

$$z = \frac{52 - 40}{7} = 1.71$$
$$z = \frac{43 - 40}{7} = 0.43$$
$$z = \frac{34 - 40}{7} = -0.86$$

A problem is that these new z-scores aren't exactly intuitive for many people. We can give people information about their relative location in the distribution (for instance, the first person scored well above average), or we can translate these z scores into the more familiar metric of IQ scores, which have a mean of 100 and standard deviation of 16:

$$\text{IQ} = 1.71 * 16 + 100 = 127.36$$
$$\text{IQ} = 0.43 * 16 + 100 = 106.88$$
$$\text{IQ} = -0.80 * 16 + 100 = 87.20$$

We would also likely round these values to 127, 107, and 87, respectively, for convenience.

This page titled 5.2: Z-scores is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **4.2: Z-scores** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

5.3: Z-scores and the Area under the Curve

Z-scores and the standard normal distribution go hand-in-hand. A z-score will tell you exactly where in the standard normal distribution a value is located, and any normal distribution can be converted into a standard normal distribution by converting all of the scores in the distribution into z-scores, a process known as standardization.

We saw in the previous chapter that standard deviations can be used to divide the normal distribution: 68% of the distribution falls within 1 standard deviation of the mean, 95% within (roughly) 2 standard deviations, and 99.7% within 3 standard deviations. Because z-scores are in units of standard deviations, this means that 68% of scores fall between z = -1.0 and z = 1.0 and so on. We call this 68% (or any percentage we have based on our z-scores) the proportion of the area under the curve. Any area under the curve is bounded by (defined by, delineated by, etc.) by a single z-score or pair of z-scores.

An important property to point out here is that, by virtue of the fact that the total area under the curve of a distribution is always equal to 1.0 (see section on Normal Distributions at the beginning of this chapter), these areas under the curve can be added together or subtracted from 1 to find the proportion in other areas. For example, we know that the area between z = -1.0 and z = 1.0 (i.e. within one standard deviation of the mean) contains 68% of the area under the curve, which can be represented in decimal form at 0.6800 (to change a percentage to a decimal, simply move the decimal point 2 places to the left). Because the total area under the curve is equal to 1.0, that means that the proportion of the area outside z= -1.0 and z = 1.0 is equal to 1.0 – 0.6800 = 0.3200 or 32% (see Figure **5.3.1** below). This area is called the area in the tails of the distribution. Because this area is split between two tails and because the normal distribution is symmetrical, each tail has exactly one-half, or 16%, of the area under the curve.

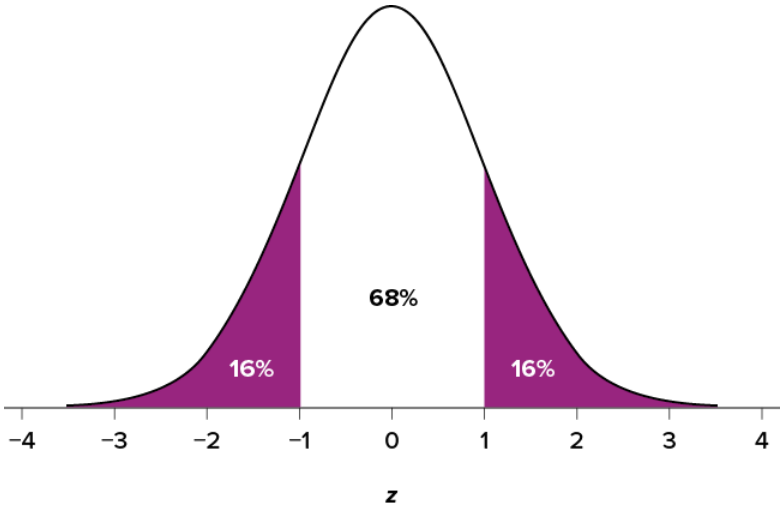


Figure **5.3.1**: Shaded areas represent the area under the curve in the tails

Image Credit: Judy Schmitt, from Cote et al, 2021

We will have much more to say about this concept in the coming chapters. As it turns out, this is a quite powerful idea that enables us to make statements about how likely an outcome is and what that means for research questions we would like to answer and hypotheses we would like to test. But first, we need to make a brief foray into some ideas about probability.

This page titled 5.3: Z-scores and the Area under the Curve is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **4.3: Z-scores and the Area under the Curve** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

5.E: Z-scores and the Standard Normal Distribution (Exercises)

1. What are the two pieces of information contained in a z-score?

Answer:

The location above or below the mean (from the sign of the number) and the distance in standard deviations away from the mean (from the magnitude of the number).

2. A z-score takes a raw score and standardizes it into units of _____.
3. Assume the following 5 scores represent a sample: 2, 3, 5, 5, 6. Transform these scores into z-scores.

Answer:

$\overline{X}$= 4.2, s = 1.64; z = -1.34, -0.73, 0.49, 0.49, 1.10

4. True or false:
   a. All normal distributions are symmetrical
   b. All normal distributions have a mean of 1.0
   c. All normal distributions have a standard deviation of 1.0
   d. The total area under the curve of all normal distributions is equal to 1
5. Interpret the location, direction, and distance (near or far) of the following z-scores:
   a. -2.00
   b. 1.25
   c. 3.50
   d. -0.34

Answer:

   a. 2 standard deviations below the mean, far
   b. 1.25 standard deviations above the mean, near
   c. 3.5 standard deviations above the mean, far
   d. 0.34 standard deviations below the mean, near

6. Transform the following z-scores into a distribution with a mean of 10 and standard deviation of 2: -1.75, 2.20, 1.65, -0.95
7. Calculate z-scores for the following raw scores taken from a population with a mean of 100 and standard deviation of 16: 112, 109, 56, 88, 135, 99

Answer:

z = 0.75, 0.56, -2.75, -0.75, 2.19, -0.06

8. What does a z-score of 0.00 represent?
9. For a distribution with a standard deviation of 20, find z-scores that correspond to:
   a. One-half of a standard deviation below the mean
   b. 5 points above the mean
   c. Three standard deviations above the mean
   d. 22 points below the mean

Answer:

   a. -0.50
   b. 0.25
   c. 3.00
   d. 1.10

10. Calculate the raw score for the following z-scores from a distribution with a mean of 15 and standard deviation of 3:

10. Calculate the raw score for the following z-scores from a distribution with a mean of 15 and standard deviation of 3.

a. 4.0

b. 2.2

c. -1.3

d. 0.46

This page titled 5.E: Z-scores and the Standard Normal Distribution (Exercises) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **4.E: Z-scores and the Standard Normal Distribution (Exercises)** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

6: Probability

6.1: What is Probability
6.2: Probability in Graphs and Distributions
6.3: The Bigger Picture
6.E: Probability (Exercises)

This page titled 6: Probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

6.1: What is Probability

When we speak of the probability of something happening, we are talking how likely it is that “thing” will happen based on the conditions present. For instance, what is the probability that it will rain? That is, how likely do we think it is that it will rain today under the circumstances or conditions today? To define or understand the conditions that might affect how likely it is to rain, we might look out the window and say, “it’s sunny outside, so it’s not very likely that it will rain today.” Stated using probability language: given that it is sunny outside, the probability of rain is low. “Given” is the word we use to state what the conditions are. As the conditions change, so does the probability. Thus, if it were cloudy and windy outside, we might say, “given the current weather conditions, there is a high probability that it is going to rain.”

In these examples, we spoke about whether or not it is going to rain. Raining is an example of an event, which is the catch-all term we use to talk about any specific thing happening; it is a generic term that we specified to mean “rain” in exactly the same way that “conditions” is a generic term that we specified to mean “sunny” or “cloudy and windy.”

It should also be noted that the terms “low” and “high” are relative and vague, and they will likely be interpreted different by different people (in other words: given how vague the terminology was, the probability of different interpretations is high). Most of the time we try to use more precise language or, even better, numbers to represent the probability of our event. Regardless, the basic structure and logic of our statements are consistent with how we speak about probability using numbers and formulas.

Let’s look at a slightly deeper example. Say we have a regular, six-sided die (note that “die” is singular and “dice” is plural, a distinction that Dr. Foster has yet to get correct on his first try) and want to know how likely it is that we will roll a 1. That is, what is the probability of rolling a 1, given that the die is not weighted (which would introduce what we call a bias, though that is beyond the scope of this chapter). We could roll the die and see if it is a 1 or not, but that won’t tell us about the probability, it will only tell us a single result. We could also roll the die hundreds or thousands of times, recording each outcome and seeing what the final list looks like, but this is time consuming, and rolling a die that many times may lead down a dark path to gambling or, worse, playing Dungeons & Dragons. What we need is a simple equation that represents what we are looking for and what is possible.

To calculate the probability of an event, which here is defined as rolling a 1 on an unbiased die, we need to know two things: how many outcomes satisfy the criteria of our event (stated different, how many outcomes would count as what we are looking for) and the total number of outcomes possible. In our example, only a single outcome, rolling a 1, will satisfy our criteria, and there are a total of six possible outcomes (rolling a 1, rolling a 2, rolling a 3, rolling a 4, rolling a 5, and rolling a 6). Thus, the probability of rolling a 1 on an unbiased die is 1 in 6 or 1/6. Put into an equation using generic terms, we get:

$$\text{Probability of an event } = \frac{\text{number of outcomes that satisfy our criteria}}{\text{total number of possible outcomes}} \tag{6.1.1}$$

We can also using P() as shorthand for probability and A as shorthand for an event:

$$P(A) = \frac{\text{number of outcomes that count a } A}{\text{total number of possible outcomes}} \tag{6.1.2}$$

Using this equation, let’s now calculate the probability of rolling an even number on this die:

$$P(\text{Even Number}) = \frac{2, 4, \text{or } 6}{1, 2, 3, 4, 5, \text{or } 6} = \frac{3}{6} = \frac{1}{2}$$

So we have a 50% chance of rolling an even number of this die. The principles laid out here operate under a certain set of conditions and can be elaborated into ideas that are complex yet powerful and elegant. However, such extensions are not necessary for a basic understanding of statistics, so we will end our discussion on the math of probability here. Now, let’s turn back to more familiar topics.

This page titled 6.1: What is Probability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri’s Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **5.1: What is Probability** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

6.2: Probability in Graphs and Distributions

We will see shortly that the normal distribution is the key to how probability works for our purposes. To understand exactly how, let's first look at a simple, intuitive example using pie charts.

Probability in Pie

Charts Recall that a pie chart represents how frequently a category was observed and that all slices of the pie chart add up to 100%, or 1. This means that if we randomly select an observation from the data used to create the pie chart, the probability of it taking on a specific value is exactly equal to the size of that category's slice in the pie chart.

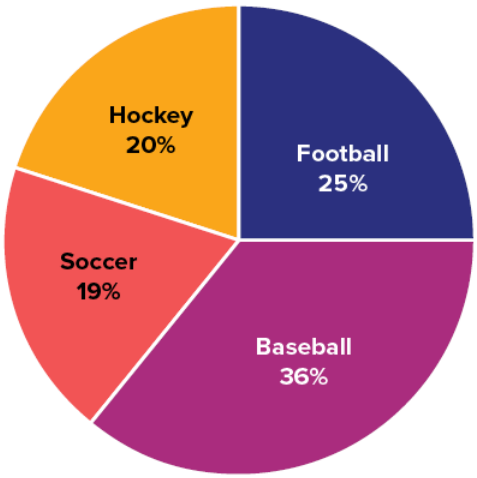


Figure **6.2.1**: Favorite sports

Image Credit: Judy Schmitt, from Cote et al, 2021.

Take, for example, the pie chart in Figure **6.2.1** representing the favorite sports of 100 people. If you put this pie chart on a dart board and aimed blindly (assuming you are guaranteed to hit the board), the likelihood of hitting the slice for any given sport would be equal to the size of that slice. So, the probability of hitting the baseball slice is the highest at 36%. The probability is equal to the proportion of the chart taken up by that section.

We can also add slices together. For instance, maybe we want to know the probability to finding someone whose favorite sport is usually played on grass. The outcomes that satisfy this criteria are baseball, football, and soccer. To get the probability, we simply add their slices together to see what proportion of the area of the pie chart is in that region: 36% + 25% + 19% = 81%. We can also add sections together even if they do not touch. If we want to know the likelihood that someone's favorite sport is not called football somewhere in the world (i.e. baseball and hockey), we can add those slices even though they aren't adjacent or continuous in the chart itself: 36% + 20% = 56%. We are able to do all of this because 1) the size of the slice corresponds to the area of the chart taken up by that slice, 2) the percentage for a specific category can be represented as a decimal (this step was skipped for ease of explanation above), and 3) the total area of the chart is equal to 100% or 1.0, which makes the size of the slices interpretable.

Probability in Normal Distributions

If the language at the end of the last section sounded familiar, that's because its exactly the language used in the last chapter to describe the normal distribution. Recall that the normal distribution has an area under its curve that is equal to 1 and that it can be split into sections by drawing a line through it that corresponds to a given z-score. Because of this, we can interpret areas under the normal curve as probabilities that correspond to z-scores.

First, let's look back at the area between z = -1.00 and z = 1.00 presented in Figure **6.2.2**. We were told earlier that this region contains 68% of the area under the curve. Thus, if we randomly chose a z-score from all possible z-scores, there is a 68% chance that it will be between z = -1.00 and z = 1.00 because those are the z-scores that satisfy our criteria.

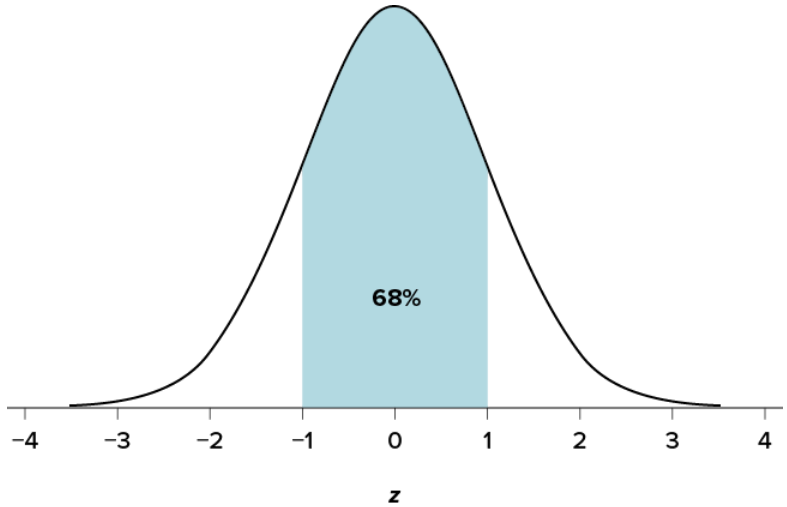


Figure 6.2.2: There is a 68% chance of selection a z-score from the blue-shaded region

Image Credit: Judy Schmitt, from Cote et al, 2021.

Just like a pie chart is broken up into slices by drawing lines through it, we can also draw a line through the normal distribution to split it into sections. Take a look at the normal distribution in Figure 6.2.3 which has a line drawn through it as z = 1.25. This line creates two sections of the distribution: the smaller section called the tail and the larger section called the body. Differentiating between the body and the tail does not depend on which side of the distribution the line is drawn. All that matters is the relative size of the pieces: bigger is always body.

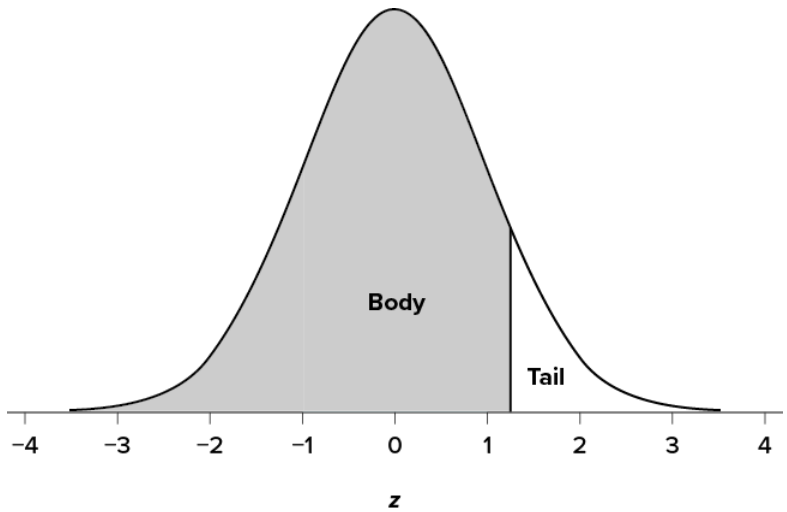


Figure 6.2.3: Body and tail of the normal distribution

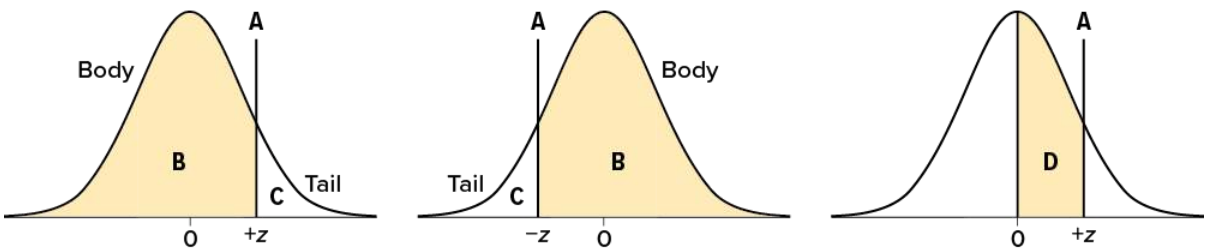
Image Credit: Judy Schmitt, from Cote et al, 2021.

As you can see, we can break up the normal distribution into 3 pieces (lower tail, body, and upper tail) as in Figure 6.2.2 or into 2 pieces (body and tail) as in Figure 6.2.3. We can then find the proportion of the area in the body and tail based on where the line was drawn (i.e. at what z-score). Mathematically this is done using calculus. Fortunately, the exact values are given you to you in the Standard Normal Distribution Table, also known at the z-table. Using the values in this table, we can find the area under the normal curve in any body, tail, or combination of tails no matter which z-scores are used to define them.

The z-table presents the values for the area under the curve to the left of the positive z-scores from 0.00-3.00 (technically 3.09), as indicated by the shaded region of the distribution at the top of the table. To find the appropriate value, we first find the row corresponding to our z-score then follow it over until we get to the column that corresponds to the number in the hundredths place of our z-score. For example, suppose we want to find the area in the body for a z-score of 1.62. We would first find the row for 1.60

then follow it across to the column labeled 0.02 (1.60 + 0.02 = 1.62) and find 0.9474 (see Table 5.1). Thus, the odds of randomly selecting someone with a z-score less than (to the left of) z = 1.62 is 94.74% because that is the proportion of the area taken up by values that satisfy our criteria.

TABLE 5.1. Standard normal distribution table (z table).



(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z	(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
1.60	.9452	.0548	.4452	1.80	.9641	.0359	.4641
1.61	.9463	.0537	.4463	1.81	.9649	.0351	.4649
1.62	.9474	.0526	.4474	1.82	.9656	.0344	.4656
1.63	.9484	.0516	.4484	1.83	.9664	.0336	.4664
1.64	.9495	.0505	.4495	1.84	.9671	.0329	.4671
1.65	.9505	.0495	.4505	1.85	.9678	.0322	.4678
1.66	.9515	.0485	.4515	1.86	.9686	.0314	.4686
1.67	.9525	.0475	.4525	1.87	.9693	.0307	.4693
1.68	.9535	.0465	.4535	1.88	.9699	.0301	.4699
1.69	.9545	.0455	.4545	1.89	.9706	.0294	.4706
1.70	.9554	.0446	.4554	1.90	.9713	.0287	.4713
1.71	.9564	.0436	.4564	1.91	.9719	.0281	.4719
1.72	.9573	.0427	.4573	1.92	.9726	.0274	.4726
1.73	.9582	.0418	.4582	1.93	.9732	.0268	.4732
1.74	.9591	.0409	.4591	1.94	.9738	.0262	.4738
1.75	.9599	.0401	.4599	1.95	.9744	.0256	.4744
1.76	.9608	.0392	.4608	1.96	.9750	.0250	.4750
1.77	.9616	.0384	.4616	1.97	.9756	.0244	.4756
1.78	.9625	.0375	.4625	1.98	.9761	.0239	.4761
1.79	.9633	.0367	.4633	1.99	.9767	.0233	.4767

("z Table Curves" by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)

Table 5.1: Using the z-table to find the area in the body to the left of z = 1.62

Image Credit: Judy Schmitt, from Cote et al, 2021.

The z-table only presents the area in the body for positive z-scores because the normal distribution is symmetrical. Thus, the area in the body of z = 1.62 is equal to the area in the body for z = -1.62, though now the body will be the shaded area to the right of z (because the body is always larger). When in doubt, drawing out your distribution and shading the area you need to find will always help. The table also only presents the area in the body because the total area under the normal curve is always equal to 1.00, so if we need to find the area in the tail for z = 1.62, we simply find the area in the body and subtract it from 1.00 (1.00 – 0.9474 = 0.0526).

Let's look at another example. This time, let's find the area corresponding to z-scores more extreme than z = -1.96 and z = 1.96. That is, let's find the area in the tails of the distribution for values less than z = -1.96 (farther negative and therefore more extreme) and greater than z = 1.96 (farther positive and therefore more extreme). This region is illustrated in Figure **6.2.5**.

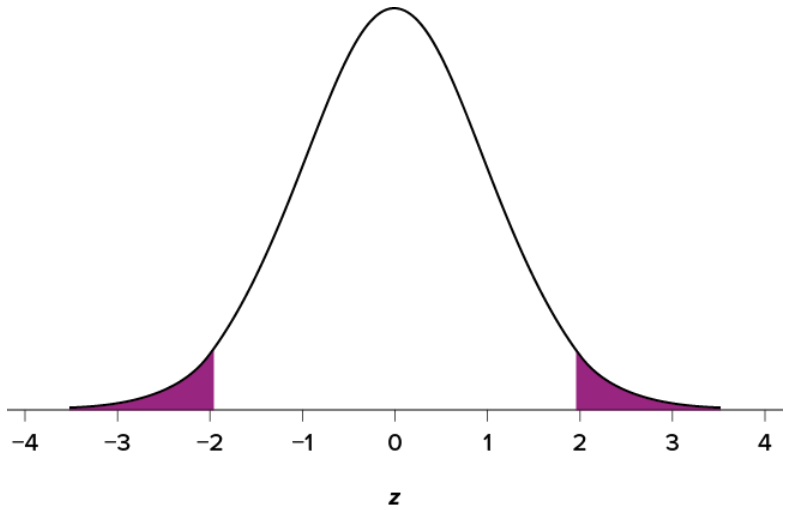


Figure **6.2.5**: Area in the tails beyond z = -1.96 and z = 1.96

Image Credit: Judy Schmitt, from Cote et al, 2021.

Let's start with the tail for z = 1.96. If we go to the z-table we will find that the body to the left of z = 1.96 is equal to 0.9750. To find the area in the tail, we subtract that from 1.00 to get 0.0250. Because the normal distribution is symmetrical, the area in the tail for z = -1.96 is the exact same value, 0.0250. Finally, to get the total area in the shaded region, we simply add the areas together to get 0.0500. Thus, there is a 5% chance of randomly getting a value more extreme than z = -1.96 or z = 1.96 (this particular value and region will become incredibly important in Unit 2).

Finally, we can find the area between two z-scores by shading and subtracting. Figure **6.2.6** shows the area between z = 0.50 and z = 1.50. Because this is a subsection of a body (rather than just a body or a tail), we must first find the larger of the two bodies, in this case the body for z = 1.50, and subtract the smaller of the two bodies, or the body for z = 0.50. Aligning the distributions vertically, as in Figure 6, makes this clearer. From the z-table, the area in the body for z = 1.50 is 0.9332 and the area in the body for z = 0.50 is 0.6915. Subtracting these gives us 0.9332 – 0.6915 = 0.2417.

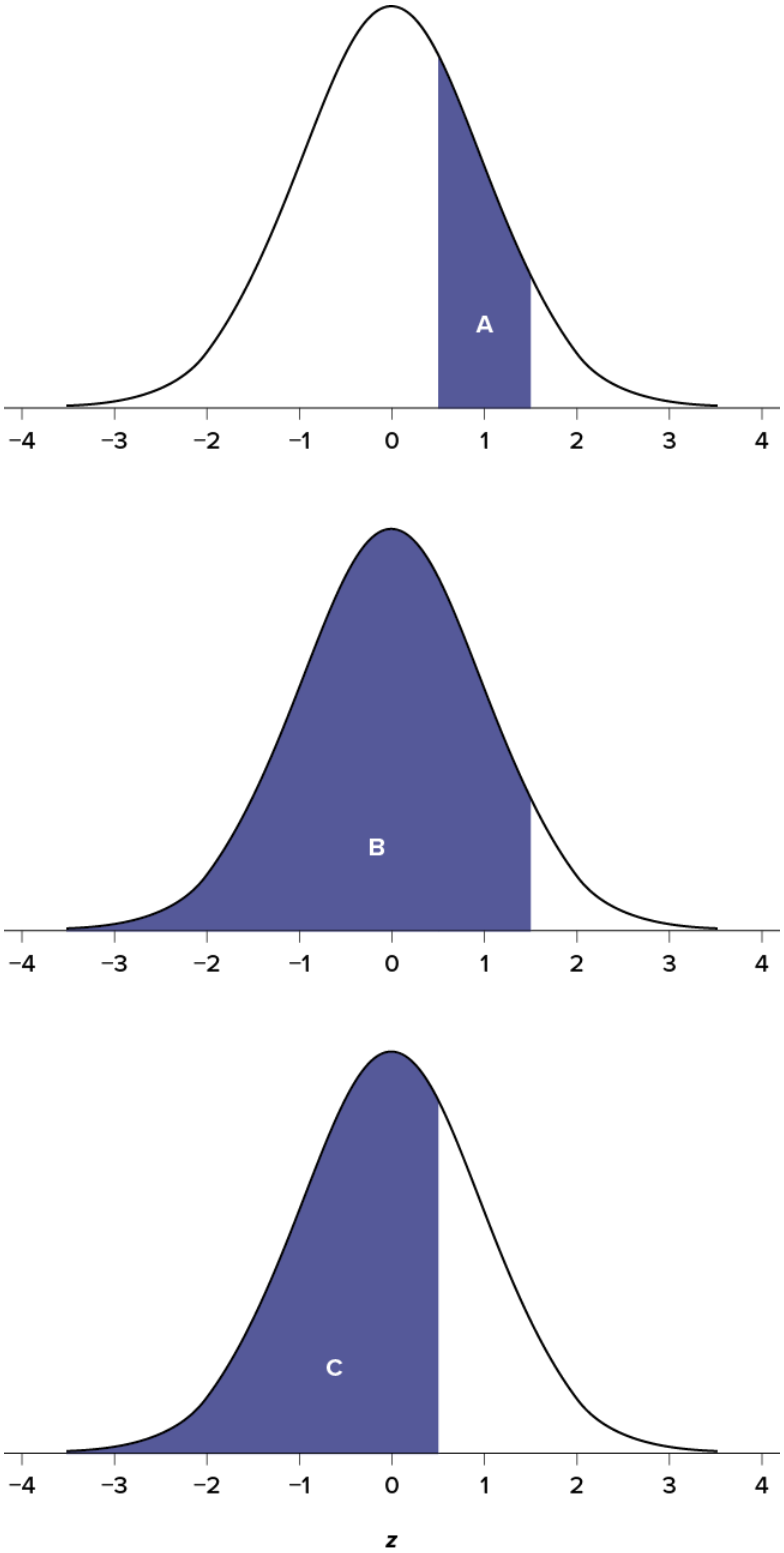


Figure **6.2.6**: Area between z = 0.50 and 1.50, along with the corresponding areas in the body

Image Credit: Judy Schmitt, from Cote et al, 2021.

This page titled 6.2: Probability in Graphs and Distributions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **5.2: Probability in Graphs and Distributions** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

6.3: The Bigger Picture

The concepts and ideas presented in this chapter are likely not intuitive at first. Probability is a tough topic for everyone, but the tools it gives us are incredibly powerful and enable us to do amazing things with data analysis. They are the heart of how inferential statistics work.

To summarize, the probability that an event happens is the number of outcomes that qualify as that event (i.e. the number of ways the event could happen) compared to the total number of outcomes (i.e. how many things are possible). This extends to graphs like a pie chart, where the biggest slices take up more of the area and are therefore more likely to be chosen at random. This idea then brings us back around to our normal distribution, which can also be broken up into regions or areas, each of which are bounded by one or two z-scores and correspond to all z-scores in that region. The probability of randomly getting one of those z-scores in the specified region can then be found on the Standard Normal Distribution Table. Thus, the larger the region, the more likely an event is, and vice versa. Because the tails of the distribution are, by definition, smaller and we go farther out into the tail, the likelihood or probability of finding a result out in the extremes becomes small.

This page titled 6.3: The Bigger Picture is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **5.3: The Bigger Picture** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

6.E: Probability (Exercises)

1. In your own words, what is probability?

Answer:

Your answer should include information about an event happening under certain conditions given certain criteria. You could also discuss the relation between probability and the area under the curve or the proportion of the area in a chart.

2. There is a bag with 5 red blocks, 2 yellow blocks, and 4 blue blocks. If you reach in and grab one block without looking, what is the probability it is red?
3. Under a normal distribution, which of the following is more likely? (Note: this question can be answered without any calculations if you draw out the distributions and shade properly)
   - Getting a z-score greater than z = 2.75
   - Getting a z-score less than z = -1.50

Answer:

Getting a z-score less than z = -1.50 is more likely. z = 2.75 is farther out into the right tail than z = -1.50 is into the left tail, therefore there are fewer more extreme scores beyond 2.75 than -1.50, regardless of the direction

4. The heights of women in the United States are normally distributed with a mean of 63.7 inches and a standard deviation of 2.7 inches. If you randomly select a woman in the United States, what is the probability that she will be between 65 and 67 inches tall?
5. The heights of men in the United States are normally distributed with a mean of 69.1 inches and a standard deviation of 2.9 inches. What proportion of men are taller than 6 feet (72 inches)?

Answer:

15.87% or 0.1587

6. You know you need to score at least 82 points on the final exam to pass your class. After the final, you find out that the average score on the exam was 78 with a standard deviation of 7. How likely is it that you pass the class?
7. What proportion of the area under the normal curve is greater than z = 1.65?

Answer:

4.95% or 0.0495

8. Find the z-score that bounds 25% of the lower tail of the distribution.
9. Find the z-score that bounds the top 9% of the distribution.

Answer:

z = 1.34 (the top 9% means 9% of the area is in the upper tail and 91% is in the body to the left; finding the value in the normal table closest to .9100 is .9099, which corresponds to z = 1.34)

10. In a distribution with a mean of 70 and standard deviation of 12, what proportion of scores are lower than 55?

This page titled 6.E: Probability (Exercises) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **5.E: Probability (Exercises)** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

7: Sampling Distributions

7.1: People, Samples, and Populations
7.2: The Sampling Distribution of Sample Means
7.3: Sampling Distribution, Probability and Inference
7.E: Sampling Distributions (Exercises)

This page titled 7: Sampling Distributions is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

7.1: People, Samples, and Populations

Most of what we have dealt with so far has concerned individual scores grouped into samples, with those samples being drawn from and, hopefully, representative of a population. We saw how we can understand the location of individual scores within a sample's distribution via z-scores, and how we can extend that to understand how likely it is to observe scores higher or lower than an individual score via probability.

Inherent in this work is the notion that an individual score will differ from the mean, which we quantify as a z-score. All of the individual scores will differ from the mean in different amounts and different directions, which is natural and expected. We quantify these differences as variance and standard deviation. Measures of spread and the idea of variability in observations is a key principle in inferential statistics. We know that any observation, whether it is a single score, a set of scores, or a particular descriptive statistic will differ from the center of whatever distribution it belongs in.

This is equally true of things outside of statistics and format data collection and analysis. Some days you hear your alarm and wake up easily, other days you need to hit snooze a few [dozen] times. Some days traffic is light, other days it is very heavy. Some classes you are able to focus, pay attention, and take good notes, but other days you find yourself zoning out the entire time. Each individual observation is an insight but is not, by itself, the entire story, and it takes an extreme deviation from what we expect for us to think that something strange is going on. Being a little sleepy is normal, but being completely unable to get out of bed might indicate that we are sick. Light traffic is a good thing, but almost no cars on the road might make us think we forgot it is Saturday. Zoning out occasionally is fine, but if we cannot focus at all, we might be in a stats class rather than a fun one.

All of these principles carry forward from scores within samples to samples within populations. Just like an individual score will differ from its mean, an individual sample mean will differ from the true population mean. We encountered this principle in earlier chapters: sampling error. As mentioned way back in chapter 1, sampling error is an incredibly important principle. We know ahead of time that if we collect data and compute a sample, the observed value of that sample will be at least slightly off from what we expect it to be based on our supposed population mean; this is natural and expected. However, if our sample mean is extremely different from what we expect based on the population mean, there may be something going on.

This page titled 7.1: People, Samples, and Populations is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **6.1: People, Samples, and Populations** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

7.2: The Sampling Distribution of Sample Means

To see how we use sampling error, we will learn about a new, theoretical distribution known as the sampling distribution. In the same way that we can gather a lot of individual scores and put them together to form a distribution with a center and spread, if we were to take many samples, all of the same size, and calculate the mean of each of those, we could put those means together to form a distribution. This new distribution is, intuitively, known as the distribution of sample means. It is one example of what we call a sampling distribution, we can be formed from a set of any statistic, such as a mean, a test statistic, or a correlation coefficient (more on the latter two in Units 2 and 3). For our purposes, understanding the distribution of sample means will be enough to see how all other sampling distributions work to enable and inform our inferential analyses, so these two terms will be used interchangeably from here on out. Let's take a deeper look at some of its characteristics.

The sampling distribution of sample means can be described by its shape, center, and spread, just like any of the other distributions we have worked with. The shape of our sampling distribution is normal: a bell-shaped curve with a single peak and two tails extending symmetrically in either direction, just like what we saw in previous chapters. The center of the sampling distribution of sample means – which is, itself, the mean or average of the means – is the true population mean, μ. This will sometimes be written as $\mu_{\overline{X}}$ to denote it as the mean of the sample means. The spread of the sampling distribution is called the standard error, the quantification of sampling error, denoted $\sigma_{\overline{X}}$. The formula for standard error is:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \tag{7.2.1}$$

Notice that the sample size is in this equation. As stated above, the sampling distribution refers to samples of a specific size. That is, all sample means must be calculated from samples of the same size n, such n = 10, n = 30, or n = 100. This sample size refers to how many people or observations are in each individual sample, not how many samples are used to form the sampling distribution. This is because the sampling distribution is a theoretical distribution, not one we will ever actually calculate or observe. Figure 7.2.1 displays the principles stated here in graphical form.

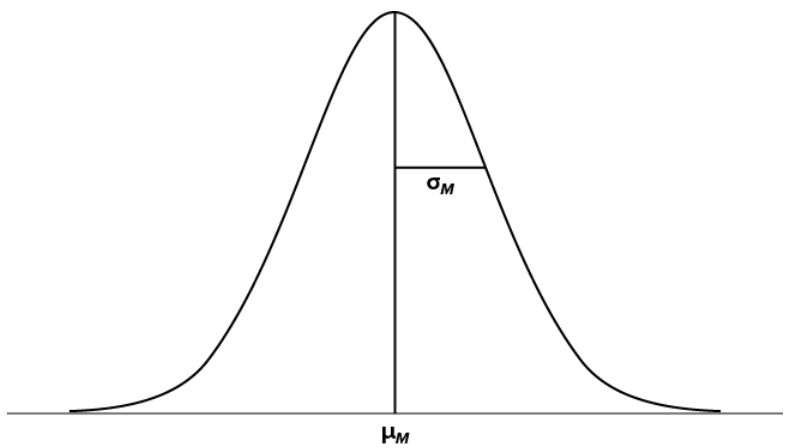


Figure 7.2.1: The sampling distribution of sample means

Image Credit: Judy Schmitt, from Cote et al, 2021.

Two Important Axioms

We just learned that the sampling distribution is theoretical: we never actually see it. If that is true, then how can we know it works? How can we use something that we don't see? The answer lies in two very important mathematical facts: the central limit theorem and the law of large numbers. We will not go into the math behind how these statements were derived, but knowing what they are and what they mean is important to understanding why inferential statistics work and how we can draw conclusions about a population based on information gained from a single sample.

Central Limit Theorem

The central limit theorem states:

Theorem 7.2.1

For samples of a single size n, drawn from a population with a given mean μ and variance σ^2, the sampling distribution of sample means will have a mean $\mu_{\overline{X}} = \mu$ and variance $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$. This distribution will approach normality as n increases.

From this, we are able to find the standard deviation of our sampling distribution, the standard error. As you can see, just like any other standard deviation, the standard error is simply the square root of the variance of the distribution.

The last sentence of the central limit theorem states that the sampling distribution will be normal as the sample size of the samples used to create it increases. What this means is that bigger samples will create a more normal distribution, so we are better able to use the techniques we developed for normal distributions and probabilities. So how large is large enough? In general, a sampling distribution will be normal if either of two characteristics is true:

1. the population from which the samples are drawn is normally distributed or
2. the sample size is equal to or greater than 30.

This second criteria is very important because it enables us to use methods developed for normal distributions even if the true population distribution is skewed.

Law of Large Numbers

The law of large numbers simply states that as our sample size increases, the probability that our sample mean is an accurate representation of the true population mean also increases. It is the formal mathematical way to state that larger samples are more accurate.

The law of large numbers is related to the central limit theorem, specifically the formulas for variance and standard error. Notice that the sample size appears in the denominators of those formulas. A larger denominator in any fraction means that the overall value of the fraction gets smaller (i.e 1/2 = 0.50, 1/3 – 0.33, 1/4 = 0.25, and so on). Thus, larger sample sizes will create smaller standard errors. We already know that standard error is the spread of the sampling distribution and that a smaller spread creates a narrower distribution. Therefore, larger sample sizes create narrower sampling distributions, which increases the probability that a sample mean will be close to the center and decreases the probability that it will be in the tails. This is illustrated in Figures 7.2.2 and 7.2.3.

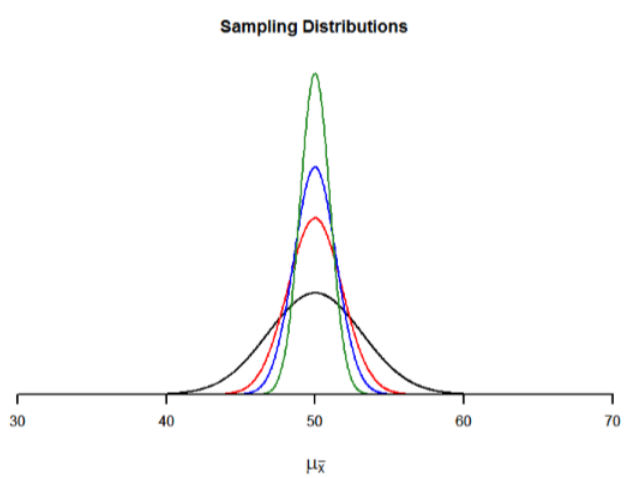


Figure 7.2.2: Sampling distributions from the same population with μ = 50 and σ = 10 but different sample sizes (N = 10, N = 30, N = 50, N = 100)

Image Credit: Judy Schmitt, from Cote et al, 2021.

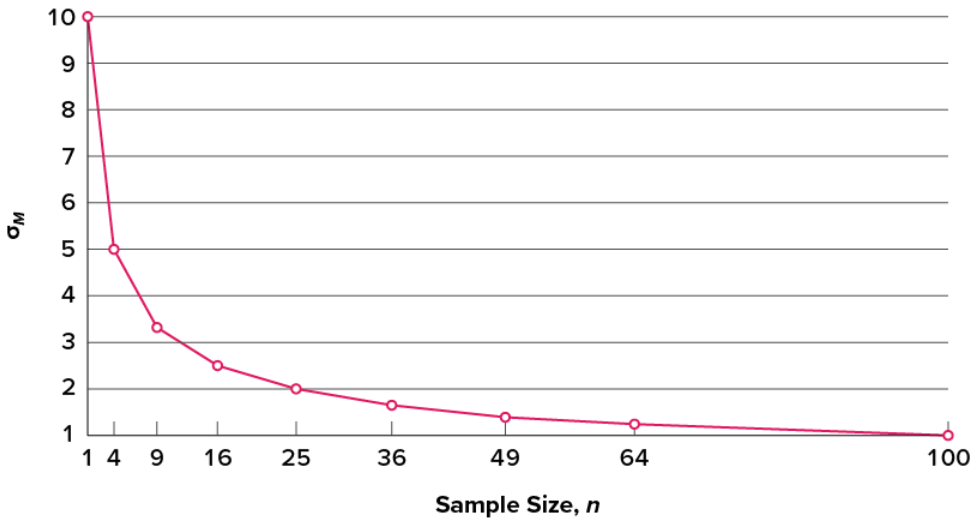


Figure 7.2.3: Relation between sample size and standard error for a constant $\sigma = 10$

Image Credit: Judy Schmitt, from Cote et al, 2021.

This page titled 7.2: The Sampling Distribution of Sample Means is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **6.2: The Sampling Distribution of Sample Means** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

7.3: Sampling Distribution, Probability and Inference

We've seen how we can use the standard error to determine probability based on our normal curve. We can think of the standard error as how much we would naturally expect our statistic – be it a mean or some other statistic) – to vary. In our formula for z based on a sample mean, the numerator $(\overline{X} - \mu)$ is what we call an observed effect. That is, it is what we observe in our sample mean versus what we expected based on the population from which that sample mean was calculated. Because the sample mean will naturally move around due to sampling error, our observed effect will also change naturally. In the context of our formula for z, then, our standard error is how much we would naturally expect the observed effect to change. Changing by a little is completely normal, but changing by a lot might indicate something is going on. This is the basis of inferential statistics and the logic behind hypothesis testing, the subject of Unit 2.

This page titled 7.3: Sampling Distribution, Probability and Inference is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **6.4: Sampling Distribution, Probability and Inference** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

7.E: Sampling Distributions (Exercises)

1. What is a sampling distribution?

Answer:

The sampling distribution (or sampling distribution of the sample means) is the distribution formed by combining many sample means taken from the same population and of a single, consistent sample size.

2. What are the two mathematical facts that describe how sampling distributions work?
3. What is the difference between a sampling distribution and a regular distribution?

Answer:

A sampling distribution is made of statistics (e.g. the mean) whereas a regular distribution is made of individual scores.

4. What effect does sample size have on the shape of a sampling distribution?
5. What is standard error?

Answer:

Standard error is the spread of the sampling distribution and is the quantification of sampling error. It is how much we expect the sample mean to naturally change based on random chance.

6. For a population with a mean of 75 and a standard deviation of 12, what proportion of sample means of size n = 16 fall above 82?
7. For a population with a mean of 100 and standard deviation of 16, what is the probability that a random sample of size 4 will have a mean between 110 and 130?

Answer:

10.46% or 0.1046

8. Find the z-score for the following means taken from a population with mean 10 and standard deviation 2:
   a. $\overline{X}$= 8, n = 12
   b. $\overline{X}$= 8, n = 30
   c. $\overline{X}$ = 20, n = 4
   d. $\overline{X}$ = 20, n = 16
9. As the sample size increases, what happens to the p-value associated with a given sample mean?

Answer:

As sample size increases, the p-value will decrease

10. For a population with a mean of 35 and standard deviation of 7, find the sample mean of size n = 20 that cuts off the top 5% of the sampling distribution.

This page titled 7.E: Sampling Distributions (Exercises) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **6.E: Sampling Distributions (Exercises)** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

8: Introduction to Hypothesis Testing

8.1: Logic and Purpose of Hypothesis Testing
8.2: The Probability Value
8.3: The Null Hypothesis
8.4: The Alternative Hypothesis
8.5: Critical values, p-values, and significance level
8.6: Steps of the Hypothesis Testing Process
8.7: Movie Popcorn
8.8: Effect Size
8.9: Office Temperature
8.10: Different Significance Level
8.11: Other Considerations in Hypothesis Testing
8.E: Introduction to Hypothesis Testing (Exercises)

This page titled 8: Introduction to Hypothesis Testing is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

8.1: Logic and Purpose of Hypothesis Testing

Let's consider a hypothetical experiment to determine whether James Bond can tell the difference between a shaken and a stirred martini. Suppose we gave Mr. Bond a series of 16 taste tests. In each test, we flipped a fair coin to determine whether to stir or shake the martini. Then we presented the martini to Mr. Bond and asked him to decide whether it was shaken or stirred. Let's say Mr. Bond was correct on 13 of the 16 taste tests. Does this prove that Mr. Bond has at least some ability to tell whether the martini was shaken or stirred?

This result does not prove that he does; it could be he was just lucky and guessed right 13 out of 16 times. But how plausible is the explanation that he was just lucky? To assess its plausibility, we determine the probability that someone who was just guessing would be correct 13/16 times or more. This probability can be computed to be 0.0106. This is a pretty low probability, and therefore someone would have to be very lucky to be correct 13 or more times out of 16 if they were just guessing. So either Mr. Bond was very lucky, or he can tell whether the drink was shaken or stirred. The hypothesis that he was guessing is not proven false, but considerable doubt is cast on it. Therefore, there is strong evidence that Mr. Bond can tell whether a drink was shaken or stirred.

Let's consider another example. The case study Physicians' Reactions sought to determine whether physicians spend less time with obese patients. Physicians were sampled randomly and each was shown a chart of a patient complaining of a migraine headache. They were then asked to estimate how long they would spend with the patient. The charts were identical except that for half the charts, the patient was obese and for the other half, the patient was of average weight. The chart a particular physician viewed was determined randomly. Thirty-three physicians viewed charts of average-weight patients and 38 physicians viewed charts of obese patients.

The mean time physicians reported that they would spend with obese patients was 24.7 minutes as compared to a mean of 31.4 minutes for normal-weight patients. How might this difference between means have occurred? One possibility is that physicians were influenced by the weight of the patients. On the other hand, perhaps by chance, the physicians who viewed charts of the obese patients tend to see patients for less time than the other physicians. Random assignment of charts does not ensure that the groups will be equal in all respects other than the chart they viewed. In fact, it is certain the groups differed in many ways by chance. The two groups could not have exactly the same mean age (if measured precisely enough such as in days). Perhaps a physician's age affects how long physicians see patients. There are innumerable differences between the groups that could affect how long they view patients. With this in mind, is it plausible that these chance differences are responsible for the difference in times?

To assess the plausibility of the hypothesis that the difference in mean times is due to chance, we compute the probability of getting a difference as large or larger than the observed difference (31.4 - 24.7 = 6.7 minutes) if the difference were, in fact, due solely to chance. Using methods presented in later chapters, this probability can be computed to be 0.0057. Since this is such a low probability, we have confidence that the difference in times is due to the patient's weight and is not due to chance.

This page titled 8.1: Logic and Purpose of Hypothesis Testing is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri’s Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **7.1: Logic and Purpose of Hypothesis Testing** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

8.2: The Probability Value

It is very important to understand precisely what the probability values mean. In the James Bond example, the computed probability of 0.0106 is the probability he would be correct on 13 or more taste tests (out of 16) if he were just guessing.

It is easy to mistake this probability of 0.0106 as the probability he cannot tell the difference. This is not at all what it means.

The probability of 0.0106 is the probability of a certain outcome (13 or more out of 16) assuming a certain state of the world (James Bond was only guessing). It is not the probability that a state of the world is true. Although this might seem like a distinction without a difference, consider the following example. An animal trainer claims that a trained bird can determine whether or not numbers are evenly divisible by 7. In an experiment assessing this claim, the bird is given a series of 16 test trials. On each trial, a number is displayed on a screen and the bird pecks at one of two keys to indicate its choice. The numbers are chosen in such a way that the probability of any number being evenly divisible by 7 is 0.50. The bird is correct on 9/16 choices. We can compute that the probability of being correct nine or more times out of 16 if one is only guessing is 0.40. Since a bird who is only guessing would do this well 40% of the time, these data do not provide convincing evidence that the bird can tell the difference between the two types of numbers. As a scientist, you would be very skeptical that the bird had this ability. Would you conclude that there is a 0.40 probability that the bird can tell the difference? Certainly not! You would think the probability is much lower than 0.0001.

To reiterate, the probability value is the probability of an outcome (9/16 or better) and not the probability of a particular state of the world (the bird was only guessing). In statistics, it is conventional to refer to possible states of the world as hypotheses since they are hypothesized states of the world. Using this terminology, the probability value is the probability of an outcome given the hypothesis. It is not the probability of the hypothesis given the outcome.

This is not to say that we ignore the probability of the hypothesis. If the probability of the outcome given the hypothesis is sufficiently low, we have evidence that the hypothesis is false. However, we do not compute the probability that the hypothesis is false. In the James Bond example, the hypothesis is that he cannot tell the difference between shaken and stirred martinis. The probability value is low (0.0106), thus providing evidence that he can tell the difference. However, we have not computed the probability that he can tell the difference.

This page titled 8.2: The Probability Value is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **7.2: The Probability Value** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

8.3: The Null Hypothesis

The hypothesis that an apparent effect is due to chance is called the null hypothesis, written H_0 ("H-naught"). In the Physicians' Reactions example, the null hypothesis is that in the population of physicians, the mean time expected to be spent with obese patients is equal to the mean time expected to be spent with average-weight patients. This null hypothesis can be written as:

$$\mathrm{H}_0 : \mu_{\text{obese}} - \mu_{\text{average}} = 0 \tag{8.3.1}$$

Although the null hypothesis is usually that the value of a parameter is 0, there are occasions in which the null hypothesis is a value other than 0. For example, if we are working with mothers in the U.S. whose children are at risk of low birth weight, we can use 7.47 pounds, the average birthweight in the US, as our null value and test for differences against that.

For now, we will focus on testing a value of a single mean against what we expect from the population. Using birthweight as an example, our null hypothesis takes the form:

$$\mathrm{H}_0 : \mu = 7.47$$

The number on the right hand side is our null hypothesis value that is informed by our research question. Notice that we are testing the value for μ, the population parameter, NOT the sample statistic $\overline{\mathrm{X}}$. This is for two reasons: 1) once we collect data, we know what the value of $\overline{\mathrm{X}}$ is – it's not a mystery or a question, it is observed and used for the second reason, which is 2) we are interested in understanding the population, not just our sample.

Keep in mind that the null hypothesis is typically the opposite of the researcher's hypothesis. In the Physicians' Reactions study, the researchers hypothesized that physicians would expect to spend less time with obese patients. The null hypothesis that the two types of patients are treated identically is put forward with the hope that it can be discredited and therefore rejected. If the null hypothesis were true, a difference as large or larger than the sample difference of 6.7 minutes would be very unlikely to occur. Therefore, the researchers rejected the null hypothesis of no difference and concluded that in the population, physicians intend to spend less time with obese patients.

In general, the null hypothesis is the idea that nothing is going on: there is no effect of our treatment, no relation between our variables, and no difference in our sample mean from what we expected about the population mean. This is always our baseline starting assumption, and it is what we seek to reject. If we are trying to treat depression, we want to find a difference in average symptoms between our treatment and control groups. If we are trying to predict job performance, we want to find a relation between conscientiousness and evaluation scores. However, until we have evidence against it, we must use the null hypothesis as our starting point.

This page titled 8.3: The Null Hypothesis is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **7.3: The Null Hypothesis** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

8.4: The Alternative Hypothesis

If the null hypothesis is rejected, then we will need some other explanation, which we call the alternative hypothesis, H_A or H_1. The alternative hypothesis is simply the reverse of the null hypothesis, and there are three options, depending on where we expect the difference to lie. Thus, our alternative hypothesis is the mathematical way of stating our research question. If we expect our obtained sample mean to be above or below the null hypothesis value, which we call a directional hypothesis, then our alternative hypothesis takes the form:

$$\mathrm{H_A}: \mu > 7.47 \quad \text{or} \quad \mathrm{H_A}: \mu < 7.47$$

based on the research question itself. We should only use a directional hypothesis if we have good reason, based on prior observations or research, to suspect a particular direction. When we do not know the direction, such as when we are entering a new area of research, we use a non-directional alternative:

$$\mathrm{H_A}: \mu \neq 7.47$$

We will set different criteria for rejecting the null hypothesis based on the directionality (greater than, less than, or not equal to) of the alternative. To understand why, we need to see where our criteria come from and how they relate to z-scores and distributions.

This page titled 8.4: The Alternative Hypothesis is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **7.4: The Alternative Hypothesis** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

8.5: Critical values, p-values, and significance level

A low probability value casts doubt on the null hypothesis. How low must the probability value be in order to conclude that the null hypothesis is false? Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than 0.05. More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01. When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis. The probability value below which the null hypothesis is rejected is called the α level or simply α ("alpha"). It is also called the significance level. If α is not explicitly specified, assume that α = 0.05.

The significance level is a threshold we set before collecting data in order to determine whether or not we should reject the null hypothesis. We set this value beforehand to avoid biasing ourselves by viewing our results and then determining what criteria we should use. If our data produce values that meet or exceed this threshold, then we have sufficient evidence to reject the null hypothesis; if not, we fail to reject the null (we never "accept" the null).

There are two criteria we use to assess whether our data meet the thresholds established by our chosen significance level, and they both have to do with our discussions of probability and distributions. Recall that probability refers to the likelihood of an event, given some situation or set of conditions. In hypothesis testing, that situation is the assumption that the null hypothesis value is the correct value, or that there is no effect. The value laid out in H0 is our condition under which we interpret our results. To reject this assumption, and thereby reject the null hypothesis, we need results that would be very unlikely if the null was true. Now recall that values of z which fall in the tails of the standard normal distribution represent unlikely values. That is, the proportion of the area under the curve as or more extreme than z is very small as we get into the tails of the distribution. Our significance level corresponds to the area under the tail that is exactly equal to α: if we use our normal criterion of α = .05, then 5% of the area under the curve becomes what we call the rejection region (also called the critical region) of the distribution. This is illustrated in Figure 8.5.1.

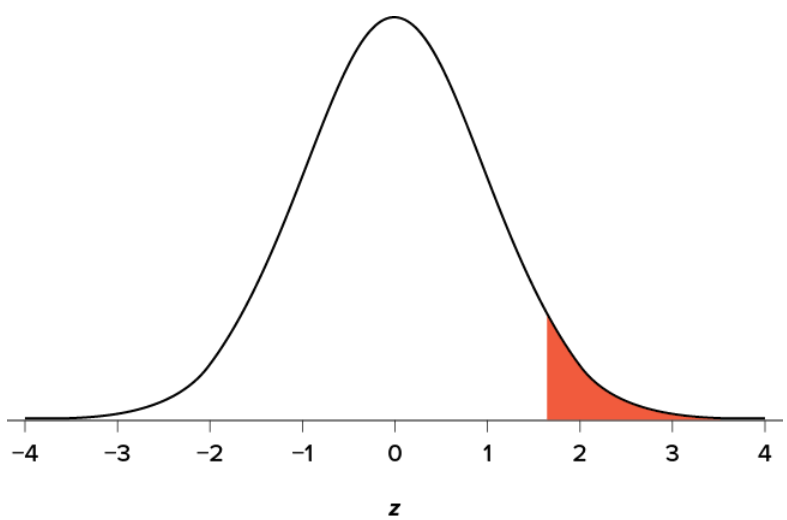


Figure 8.5.1: The rejection region for a one-tailed test.

Image Credit: Judy Schmitt, from Cote et al, 2021

The shaded rejection region takes us 5% of the area under the curve. Any result which falls in that region is sufficient evidence to reject the null hypothesis.

The rejection region is bounded by a specific z-value, as is any area under the curve. In hypothesis testing, the value corresponding to a specific rejection region is called the critical value, z_{crit} ("z-crit") or $z*$ (hence the other name "critical region"). Finding the critical value works exactly the same as finding the z-score corresponding to any area under the curve like we did in Unit 1. If we go to the normal table, we will find that the z-score corresponding to 5% of the area under the curve is equal to 1.645 (z = 1.64 corresponds to 0.0405 and z = 1.65 corresponds to 0.0495, so .05 is exactly in between them) if we go to the right and -1.645 if we go to the left. The direction must be determined by your alternative hypothesis, and drawing then shading the distribution is helpful for keeping directionality straight.

Suppose, however, that we want to do a non-directional test. We need to put the critical region in both tails, but we don't want to increase the overall size of the rejection region (for reasons we will see later). To do this, we simply split it in half so that an equal proportion of the area under the curve falls in each tail's rejection region. For α = .05, this means 2.5% of the area is in each tail, which, based on the z-table, corresponds to critical values of $z*$ = ±1.96. This is shown in Figure 8.5.2.

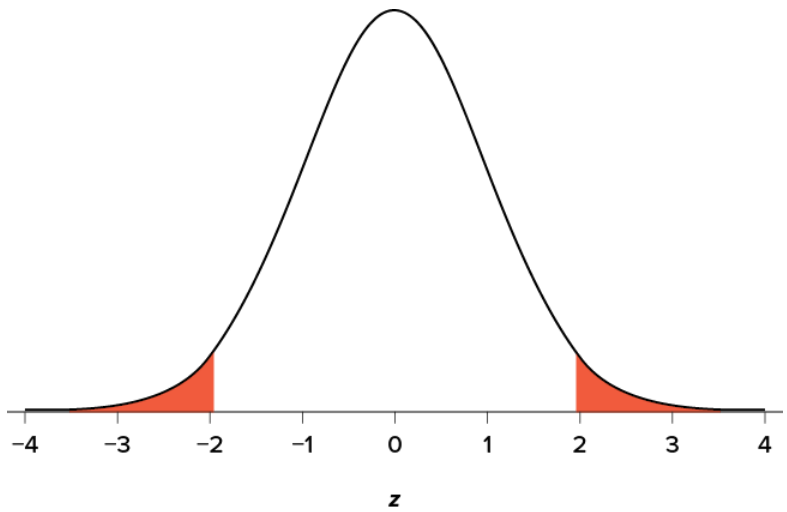


Figure 8.5.2: Two-tailed rejection region.

Image Credit: Judy Schmitt, from Cote et al, 2021

Thus, any z-score falling outside ±1.96 (greater than 1.96 in absolute value) falls in the rejection region. When we use z-scores in this way, the obtained value of z (sometimes called z-obtained) is something known as a test statistic, which is simply an inferential statistic used to test a null hypothesis. The formula for our z-statistic has not changed:

$$z = \frac{\overline{\mathrm{X}} - \mu}{\bar{\sigma}/\sqrt{\mathrm{n}}} \tag{8.5.1}$$

To formally test our hypothesis, we compare our obtained z-statistic to our critical z-value. If $\mathrm{Z}_{\mathrm{obt}} > \mathrm{Z}_{\mathrm{crit}}$, that means it falls in the rejection region (to see why, draw a line for z = 2.5 on Figure 8.5.1 or Figure 8.5.2) and so we reject H_0. If $\mathrm{Z}_{\mathrm{obt}} < \mathrm{Z}_{\mathrm{crit}}$, we fail to reject. Remember that as z gets larger, the corresponding area under the curve beyond z gets smaller. Thus, the proportion, or p-value, will be smaller than the area for α, and if the area is smaller, the probability gets smaller. Specifically, the probability of obtaining that result, or a more extreme result, under the condition that the null hypothesis is true gets smaller.

The z-statistic is very useful when we are doing our calculations by hand. However, when we use computer software, it will report to us a p-value, which is simply the proportion of the area under the curve in the tails beyond our obtained z-statistic. We can directly compare this p-value to α to test our null hypothesis: if $p < \alpha$, we reject H_0, but if $p > \alpha$, we fail to reject. Note also that the reverse is always true: if we use critical values to test our hypothesis, we will always know if p is greater than or less than α. If we reject, we know that $p < \alpha$ because the obtained z-statistic falls farther out into the tail than the critical z-value that corresponds to α, so the proportion (p-value) for that z-statistic will be smaller. Conversely, if we fail to reject, we know that the proportion will be larger than α because the z-statistic will not be as far into the tail. This is illustrated for a one-tailed test in Figure 8.5.3.

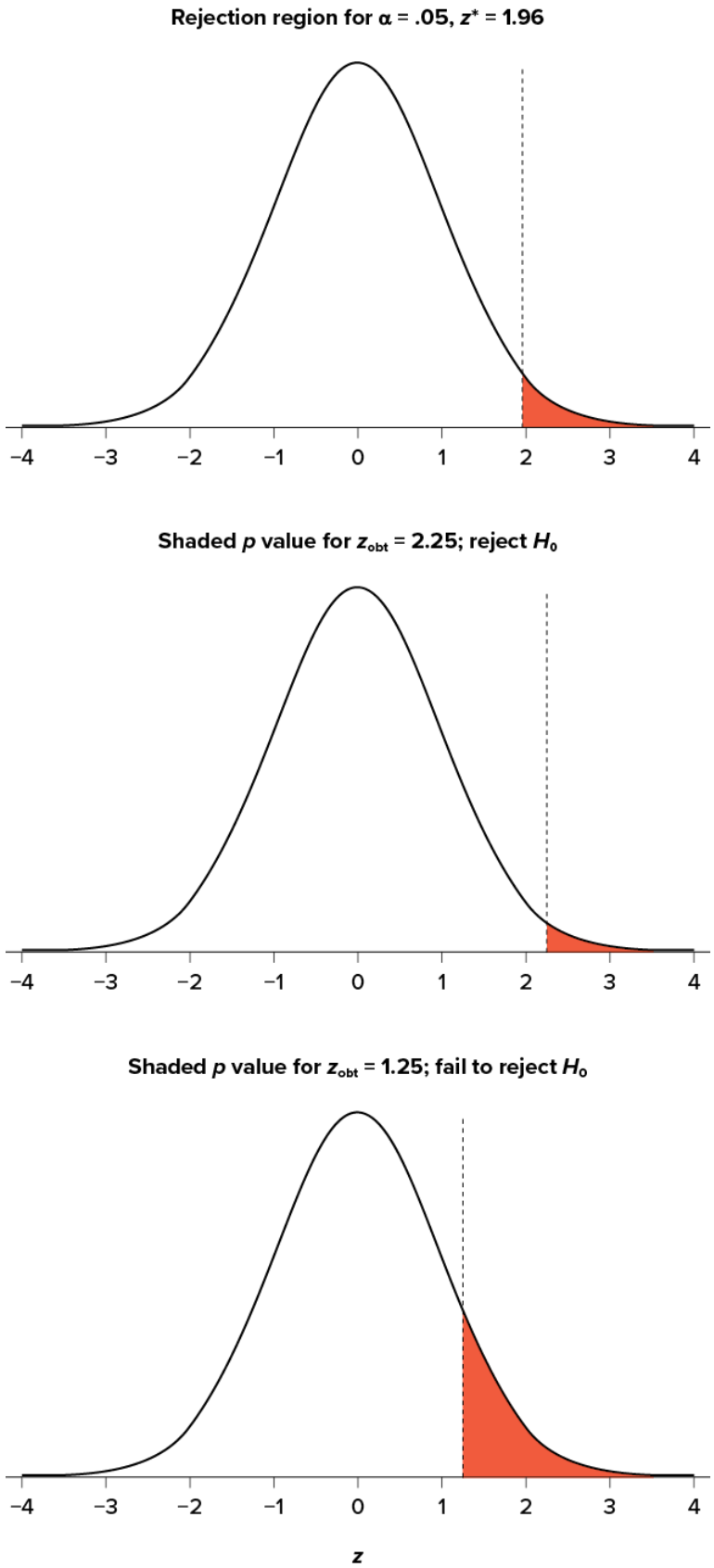

Rejection region for α = .05, z* = 1.96
−4 −3 −2 −1 0 1 2 3 4
Shaded p value for z_obt = 2.25; reject H_0
−4 −3 −2 −1 0 1 2 3 4
Shaded p value for z_obt = 1.25; fail to reject H_0
−4 −3 −2 −1 0 1 2 3 4
z

Figure **8.5.3**: Relation between α, z_{obt}, and p

Image Credit: Judy Schmitt, from Cote et al, 2021

When the null hypothesis is rejected, the effect is said to be statistically significant. For example, in the Physicians Reactions case study, the probability value is 0.0057. Therefore, the effect of obesity is statistically significant and the null hypothesis that obesity makes no difference is rejected. It is very important to keep in mind that statistical significance means only that the null hypothesis of exactly no effect is rejected; it does not mean that the effect is important, which is what "significant" usually means. When an effect is significant, you can have confidence the effect is not exactly zero. Finding that an effect is significant does not tell you about how large or important the effect is. Do not confuse statistical significance with practical significance. A small effect can be highly significant if the sample size is large enough. Why does the word "significant" in the phrase "statistically significant" mean something so different from other uses of the word? Interestingly, this is because the meaning of "significant" in everyday language has changed. It turns out that when the procedures for hypothesis testing were developed, something was "significant" if it signified something. Thus, finding that an effect is statistically significant signifies that the effect is real and not due to chance. Over the years, the meaning of "significant" changed, leading to the potential misinterpretation.

This page titled 8.5: Critical values, p-values, and significance level is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **7.5: Critical values, p-values, and significance level** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

8.6: Steps of the Hypothesis Testing Process

The process of testing hypotheses follows a simple four-step procedure. This process will be what we use for the remained of the textbook and course, and though the hypothesis and statistics we use will change, this process will not.

Step 1: State the Hypotheses Your hypotheses are the first thing you need to lay out. Otherwise, there is nothing to test! You have to state the null hypothesis (which is what we test) and the alternative hypothesis (which is what we expect). These should be stated mathematically as they were presented above AND in words, explaining in normal English what each one means in terms of the research question.

Step 2: Find the Critical Values Next, we formally lay out the criteria we will use to test our hypotheses. There are two pieces of information that inform our critical values: α, which determines how much of the area under the curve composes our rejection region, and the directionality of the test, which determines where the region will be.

Step 3: Compute the Test Statistic Once we have our hypotheses and the standards we use to test them, we can collect data and calculate our test statistic, in this case z. This step is where the vast majority of differences in future chapters will arise: different tests used for different data are calculated in different ways, but the way we use and interpret them remains the same.

Step 4: Make the Decision Finally, once we have our obtained test statistic, we can compare it to our critical value and decide whether we should reject or fail to reject the null hypothesis. When we do this, we must interpret the decision in relation to our research question, stating what we concluded, what we based our conclusion on, and the specific statistics we obtained.

This page titled 8.6: Steps of the Hypothesis Testing Process is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **7.6: Steps of the Hypothesis Testing Process** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

8.7: Movie Popcorn

Let's see how hypothesis testing works in action by working through an example. Say that a movie theater owner likes to keep a very close eye on how much popcorn goes into each bag sold, so he knows that the average bag has 8 cups of popcorn and that this varies a little bit, about half a cup. That is, the known population mean is μ = 8.00 and the known population standard deviation is σ = 0.50. The owner wants to make sure that the newest employee is filling bags correctly, so over the course of a week he randomly assesses 25 bags filled by the employee to test for a difference (N = 25). He doesn't want bags overfilled or under filled, so he looks for differences in both directions. This scenario has all of the information we need to begin our hypothesis testing procedure.

Step 1: State the Hypotheses Our manager is looking for a difference in the mean weight of popcorn bags compared to the population mean of 8. We will need both a null and an alternative hypothesis written both mathematically and in words. We'll always start with the null hypothesis:

H_0: There is no difference in the weight of popcorn bags from this employee

H_0: $\mu = 8.00$

Notice that we phrase the hypothesis in terms of the population parameter μ, which in this case would be the true average weight of bags filled by the new employee. Our assumption of no difference, the null hypothesis, is that this mean is exactly the same as the known population mean value we want it to match, 8.00. Now let's do the alternative:

H_A: There is a difference in the weight of popcorn bags from this employee

H_A: $\mu \neq 8.00$

In this case, we don't know if the bags will be too full or not full enough, so we do a two-tailed alternative hypothesis that there is a difference.

Step 2: Find the Critical Values Our critical values are based on two things: the directionality of the test and the level of significance. We decided in step 1 that a two-tailed test is the appropriate directionality. We were given no information about the level of significance, so we assume that α = 0.05 is what we will use. As stated earlier in the chapter, the critical values for a two-tailed z-test at α = 0.05 are $z*$ = ±1.96. This will be the criteria we use to test our hypothesis. We can now draw out our distribution so we can visualize the rejection region and make sure it makes sense.

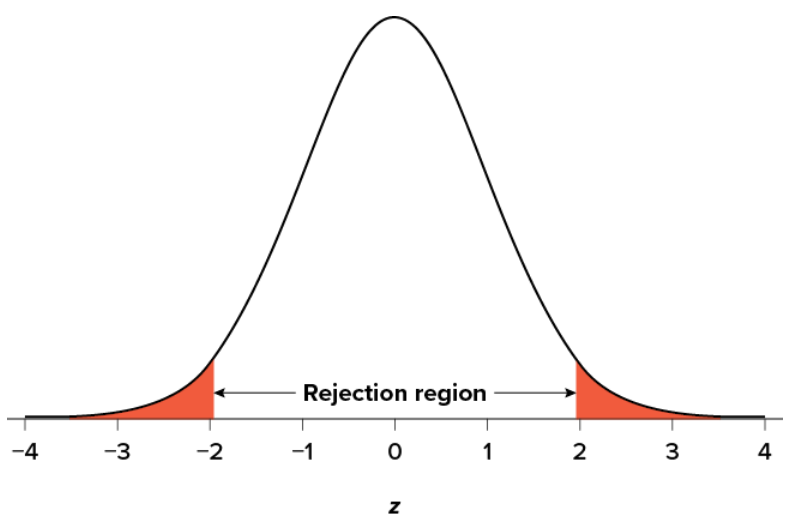


Figure 8.7.1: Rejection region for $z*$ = ±1.96

Image Credit: Judy Schmitt, from Cote et al, 2021

Step 3: Calculate the Test Statistic Now we come to our formal calculations. Let's say that the manager collects data and finds that the average weight of this employee's popcorn bags is $\overline{X}$= 7.75 cups. We can now plug this value, along with the values presented in the original problem, into our equation for z:

$$z = \frac{7.75 - 8.00}{0.50/\sqrt{25}} = \frac{-0.25}{0.10} = -2.50$$

So our test statistic is z = -2.50, which we can draw onto our rejection region distribution:

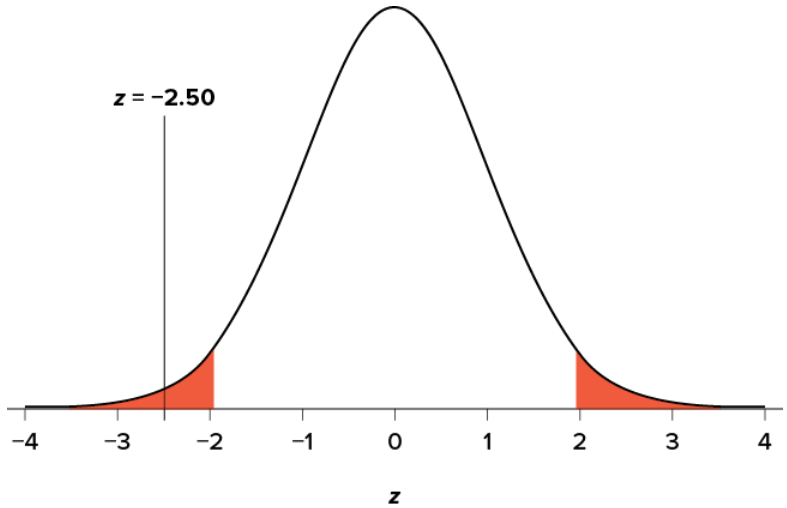


Figure 8.7.2: Test statistic location

Image Credit: Judy Schmitt, from Cote et al, 2021

Step 4: Make the Decision Looking at Figure 8.7.2, we can see that our obtained z-statistic falls in the rejection region. We can also directly compare it to our critical value: in terms of absolute value, -2.50 > -1.96, so we reject the null hypothesis. We can now write our conclusion:

> Reject H_0. Based on the sample of 25 bags, we can conclude that the average popcorn bag from this employee is smaller ($\overline{X}$= 7.75 cups) than the average weight of popcorn bags at this movie theater, z = 2.50, p < 0.05.

When we write our conclusion, we write out the words to communicate what it actually means, but we also include the average sample size we calculated (the exact location doesn't matter, just somewhere that flows naturally and makes sense) and the z-statistic and p-value. We don't know the exact p-value, but we do know that because we rejected the null, it must be less than α.

This page titled 8.7: Movie Popcorn is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **7.7: Movie Popcorn** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

8.8: Effect Size

When we reject the null hypothesis, we are stating that the difference we found was statistically significant, but we have mentioned several times that this tells us nothing about practical significance. To get an idea of the actual size of what we found, we can compute a new statistic called an effect size. Effect sizes give us an idea of how large, important, or meaningful a statistically significant effect is. For mean differences like we calculated here, our effect size is Cohen's d:

$$d = \frac{\overline{X} - \mu}{\sigma} \tag{8.8.1}$$

This is very similar to our formula for z, but we no longer take into account the sample size (since overly large samples can make it too easy to reject the null). Cohen's d is interpreted in units of standard deviations, just like z. For our example:

$$d = \frac{7.75 - 8.00}{0.50} = \frac{-0.25}{0.50} = 0.50$$

Cohen's d is interpreted as small, moderate, or large. Specifically, d = 0.20 is small, d = 0.50 is moderate, and d = 0.80 is large. Obviously values can fall in between these guidelines, so we should use our best judgment and the context of the problem to make our final interpretation of size. Our effect size happened to be exactly equal to one of these, so we say that there was a moderate effect.

Effect sizes are incredibly useful and provide important information and clarification that overcomes some of the weakness of hypothesis testing. Whenever you find a significant result, you should always calculate an effect size.

This page titled 8.8: Effect Size is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **7.8: Effect Size** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

8.9: Office Temperature

Let's do another example to solidify our understanding. Let's say that the office building you work in is supposed to be kept at 74 degree Fahrenheit but is allowed to vary by 1 degree in either direction. You suspect that, as a cost saving measure, the temperature was secretly set higher. You set up a formal way to test your hypothesis.

Step 1: State the Hypotheses You start by laying out the null hypothesis:

H_0: There is no difference in the average building temperature

$H_0 : \mu = 74$

Next you state the alternative hypothesis. You have reason to suspect a specific direction of change, so you make a one-tailed test:

H_A: The average building temperature is higher than claimed

$\mathrm{H_A} : \mu > 74$

Step 2: Find the Critical Values You know that the most common level of significance is α = 0.05, so you keep that the same and know that the critical value for a one-tailed z-test is $z*$ = 1.645. To keep track of the directionality of the test and rejection region, you draw out your distribution:

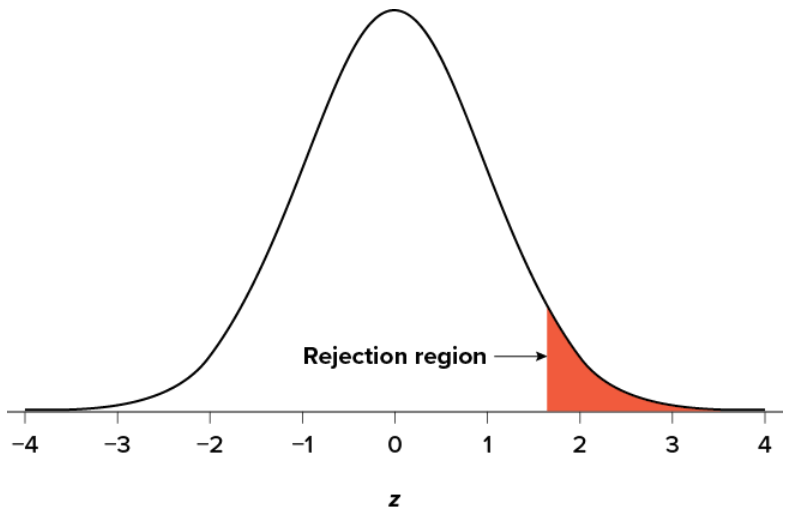


Figure 8.9.1: Rejection region

Image Credit: Judy Schmitt, from Cote et al, 2021

Step 3: Calculate the Test Statistic Now that you have everything set up, you spend one week collecting temperature data:

Table 8.9.1: Temperature data for a week

Day	Temperature
Monday	77
Tuesday	76
Wednesday	74
Thursday	78
Friday	78

You calculate the average of these scores to be $\overline{\mathrm{X}}$= 76.6 degrees. You use this to calculate the test statistic, using μ = 74 (the supposed average temperature), σ = 1.00 (how much the temperature should vary), and n = 5 (how many data points you collected):

$$z = \frac{76.60 - 74.00}{1.00/\sqrt{5}} = \frac{2.60}{0.45} = 5.78$$

This value falls so far into the tail that it cannot even be plotted on the distribution!

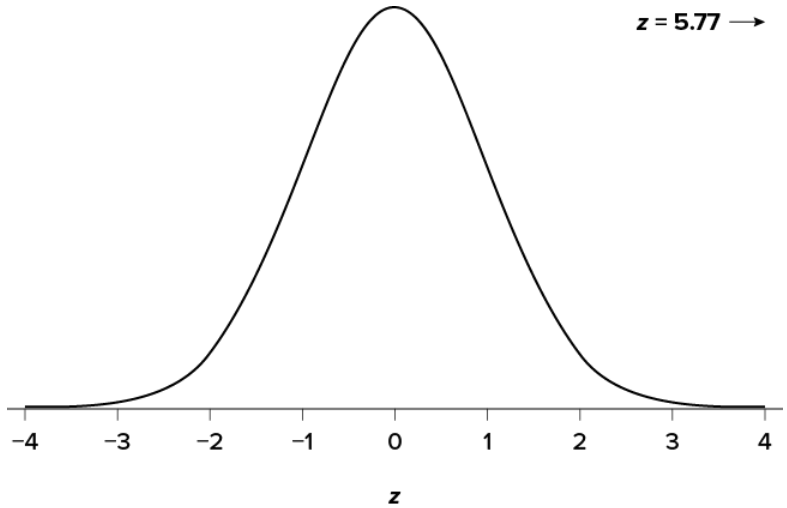


Figure 8.9.2: Obtained z-statistic

Image Credit: Judy Schmitt, from Cote et al, 2021

Step 4: Make the Decision You compare your obtained z-statistic, z = 5.77, to the critical value, $z*$ = 1.645, and find that $z > z*$. Therefore you reject the null hypothesis, concluding:

Based on 5 observations, the average temperature ($\overline{X}$= 76.6 degrees) is statistically significantly higher than it is supposed to be, z = 5.77, p < .05.

Because the result is significant, you also calculate an effect size:

$$d = \frac{76.60 - 74.00}{1.00} = \frac{2.60}{1.00} = 2.60$$

The effect size you calculate is definitely large, meaning someone has some explaining to do!

This page titled 8.9: Office Temperature is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **7.9: Office Temperature** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

8.10: Different Significance Level

Finally, let's take a look at an example phrased in generic terms, rather than in the context of a specific research question, to see the individual pieces one more time. This time, however, we will use a stricter significance level, α = 0.01, to test the hypothesis.

Step 1: State the Hypotheses We will use 60 as an arbitrary null hypothesis value:

H_0: The average score does not differ from the population

$H_0 : \mu = 50$

We will assume a two-tailed test:

H_A: The average score does differ

$H_A : \mu \neq 50$

Step 2: Find the Critical Values We have seen the critical values for z-tests at α = 0.05 levels of significance several times. To find the values for α = 0.01, we will go to the standard normal table and find the z-score cutting of 0.005 (0.01 divided by 2 for a two-tailed test) of the area in the tail, which is $z*$ = ±2.575. Notice that this cutoff is much higher than it was for α = 0.05. This is because we need much less of the area in the tail, so we need to go very far out to find the cutoff. As a result, this will require a much larger effect or much larger sample size in order to reject the null hypothesis.

Step 3: Calculate the Test Statistic We can now calculate our test statistic. We will use σ = 10 as our known population standard deviation and the following data to calculate our sample mean:

61	62
65	61
58	59
54	61
60	63

The average of these scores is $\overline{X}$= 60.40. From this we calculate our z-statistic as:

$$z = \frac{60.40 - 60.00}{10.00/\sqrt{10}} = \frac{0.40}{3.16} = 0.13$$

Step 4: Make the Decision Our obtained z-statistic, z = 0.13, is very small. It is much less than our critical value of 2.575. Thus, this time, we fail to reject the null hypothesis. Our conclusion would look something like:

Based on the sample of 10 scores, we cannot conclude that there is no effect causing the mean ($\overline{X}$= 60.40) to be statistically significantly different from 60.00, z = 0.13, p > 0.01.

Notice two things about the end of the conclusion. First, we wrote that p is greater than instead of p is less than, like we did in the previous two examples. This is because we failed to reject the null hypothesis. We don't know exactly what the p-value is, but we know it must be larger than the α level we used to test our hypothesis. Second, we used 0.01 instead of the usual 0.05, because this time we tested at a different level. The number you compare to the p-value should always be the significance level you test at.

Finally, because we did not detect a statistically significant effect, we do not need to calculate an effect size.

This page titled 8.10: Different Significance Level is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **7.10: Different Significance Level** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

8.11: Other Considerations in Hypothesis Testing

There are several other considerations we need to keep in mind when performing hypothesis testing.

Errors in Hypothesis Testing

In the Physicians' Reactions case study, the probability value associated with the significance test is 0.0057. Therefore, the null hypothesis was rejected, and it was concluded that physicians intend to spend less time with obese patients. Despite the low probability value, it is possible that the null hypothesis of no true difference between obese and average-weight patients is true and that the large difference between sample means occurred by chance. If this is the case, then the conclusion that physicians intend to spend less time with obese patients is in error. This type of error is called a Type I error. More generally, a Type I error occurs when a significance test results in the rejection of a true null hypothesis.

By one common convention, if the probability value is below 0.05 then the null hypothesis is rejected. Another convention, although slightly less common, is to reject the null hypothesis if the probability value is below 0.01. The threshold for rejecting the null hypothesis is called the α level or simply α. It is also called the significance level. As discussed in the introduction to hypothesis testing, it is better to interpret the probability value as an indication of the weight of evidence against the null hypothesis than as part of a decision rule for making a reject or do-not-reject decision. Therefore, keep in mind that rejecting the null hypothesis is not an all-or-nothing decision.

The Type I error rate is affected by the α level: the lower the α level the lower the Type I error rate. It might seem that α is the probability of a Type I error. However, this is not correct. Instead, α is the probability of a Type I error given that the null hypothesis is true. If the null hypothesis is false, then it is impossible to make a Type I error.

The second type of error that can be made in significance testing is failing to reject a false null hypothesis. This kind of error is called a Type II error. Unlike a Type I error, a Type II error is not really an error. When a statistical test is not significant, it means that the data do not provide strong evidence that the null hypothesis is false. Lack of significance does not support the conclusion that the null hypothesis is true. Therefore, a researcher should not make the mistake of incorrectly concluding that the null hypothesis is true when a statistical test was not significant. Instead, the researcher should consider the test inconclusive. Contrast this with a Type I error in which the researcher erroneously concludes that the null hypothesis is false when, in fact, it is true.

A Type II error can only occur if the null hypothesis is false. If the null hypothesis is false, then the probability of a Type II error is called β (beta). The probability of correctly rejecting a false null hypothesis equals 1- β and is called power. Power is simply our ability to correctly detect an effect that exists. It is influenced by the size of the effect (larger effects are easier to detect), the significance level we set (making it easier to reject the null makes it easier to detect an effect, but increases the likelihood of a Type I Error), and the sample size used (larger samples make it easier to reject the null).

Misconceptions in Hypothesis Testing

Misconceptions about significance testing are common. This section lists three important ones.

1. Misconception: The probability value is the probability that the null hypothesis is false.
   - Proper interpretation: The probability value is the probability of a result as extreme or more extreme given that the null hypothesis is true. It is the probability of the data given the null hypothesis. It is not the probability that the null hypothesis is false.
2. Misconception: A low probability value indicates a large effect.
   - Proper interpretation: A low probability value indicates that the sample outcome (or one more extreme) would be very unlikely if the null hypothesis were true. A low probability value can occur with small effect sizes, particularly if the sample size is large.
3. Misconception: A non-significant outcome means that the null hypothesis is probably true.
   - Proper interpretation: A non-significant outcome means that the data do not conclusively demonstrate that the null hypothesis is false.

This page titled 8.11: Other Considerations in Hypothesis Testing is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **7.11: Other Considerations in Hypothesis Testing** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

8.E: Introduction to Hypothesis Testing (Exercises)

1. In your own words, explain what the null hypothesis is.

Answer:

Your answer should include mention of the baseline assumption of no difference between the sample and the population.

2. What are Type I and Type II Errors?
3. What is α?

Answer:

Alpha is the significance level. It is the criteria we use when decided to reject or fail to reject the null hypothesis, corresponding to a given proportion of the area under the normal distribution and a probability of finding extreme scores assuming the null hypothesis is true.

4. Why do we phrase null and alternative hypotheses with population parameters and not sample means?
5. If our null hypothesis is "$H_0 : \mu = 40$", what are the three possible alternative hypotheses?

Answer:

$H_A : \mu \neq 40$, $H_A : \mu > 40$, $H_A : \mu < 40$

6. Why do we state our hypotheses and decision criteria before we collect our data?
7. When and why do you calculate an effect size?

Answer:

We calculate an effect size when we find a statistically significant result to see if our result is practically meaningful or important

8. Determine whether you would reject or fail to reject the null hypothesis in the following situations:
   a. z= 1.99, two-tailed test at α = 0.05
   b. z = 0.34, $z*$ = 1.645
   c. p = 0.03, α = 0.05
   d. p = 0.015, α = 0.01

This page titled 8.E: Introduction to Hypothesis Testing (Exercises) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **7.E: Introduction to Hypothesis Testing (Exercises)** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

9: Introduction to t-tests

9.1: The t-statistic
9.2: Hypothesis Testing with t
9.3: Confidence Intervals
9.E: Introduction to t-tests (Exercises)

This page titled 9: Introduction to t-tests is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

9.1: The t-statistic

Last chapter, we were introduced to hypothesis testing using the z-statistic for sample means that we learned in Unit 1. This was a useful way to link the material and ease us into the new way to looking at data, but it isn't a very common test because it relies on knowing the populations standard deviation, σ, which is rarely going to be the case. Instead, we will estimate that parameter σ using the sample statistic s in the same way that we estimate μ using $\bar{X}$ (μ will still appear in our formulas because we suspect something about its value and that is what we are testing). Our new statistic is called t, and for testing one population mean using a single sample (called a 1-sample t-test) it takes the form:

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \tag{9.1.1}$$

Notice that t looks almost identical to z; this is because they test the exact same thing: the value of a sample mean compared to what we expect of the population. The only difference is that the standard error is now denoted $s_{\bar{X}}$ to indicate that we use the sample statistic for standard deviation, s, instead of the population parameter σ. The process of using and interpreting the standard error and the full test statistic remain exactly the same.

In chapter 3 we learned that the formulae for sample standard deviation and population standard deviation differ by one key factor: the denominator for the parameter is N but the denominator for the statistic is n–1, also known as degrees of freedom, df. Because we are using a new measure of spread, we can no longer use the standard normal distribution and the z-table to find our critical values. For t-tests, we will use the t-distribution and t-table to find these values.

The t-distribution, like the standard normal distribution, is symmetric and normally distributed with a mean of 0 and standard error (as the measure of standard deviation for sampling distributions) of 1. However, because the calculation of standard error uses degrees of freedom, there will be a different t-distribution for every degree of freedom. Luckily, they all work exactly the same, so in practice this difference is minor.

Figure 9.1.1 shows four curves: a normal distribution curve labeled z, and three t-distribution curves for 2, 10, and 30 degrees of freedom. Two things should stand out: First, for lower degrees of freedom (e.g. 2), the tails of the distribution are much fatter, meaning the a larger proportion of the area under the curve falls in the tail. This means that we will have to go farther out into the tail to cut off the portion corresponding to 5% or α = 0.05, which will in turn lead to higher critical values. Second, as the degrees of freedom increase, we get closer and closer to the z curve. Even the distribution with df = 30, corresponding to a sample size of just 31 people, is nearly indistinguishable from z. In fact, a t-distribution with infinite degrees of freedom (theoretically, of course) is exactly the standard normal distribution. Because of this, the bottom row of the t-table also includes the critical values for z-tests at the specific significance levels. Even though these curves are very close, it is still important to use the correct table and critical values, because small differences can add up quickly.

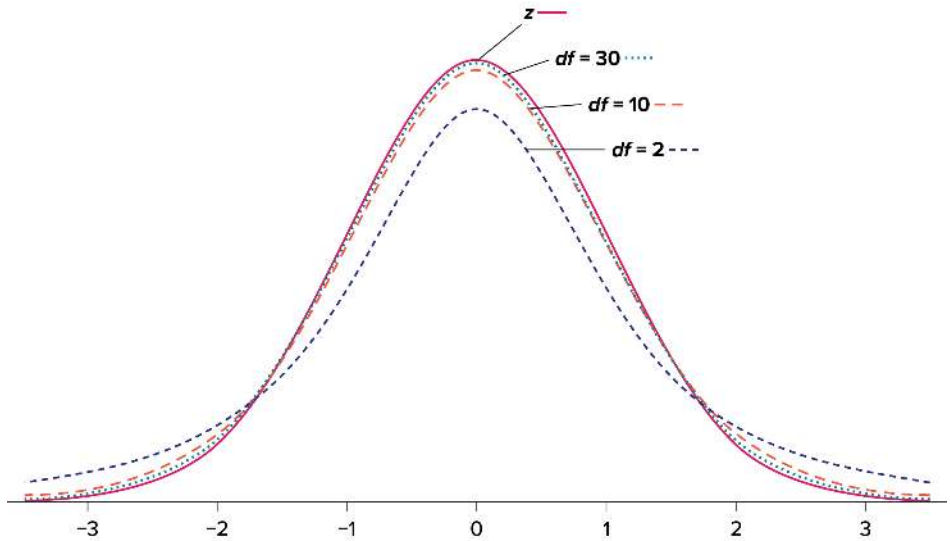


Figure **9.1.1**: Distributions comparing effects of degrees of freedom

Image Credit: Judy Schmitt, from Cote et al, 2021.

The t-distribution table lists critical values for one- and two-tailed tests at several levels of significance arranged into columns. The rows of the t-table list degrees of freedom up to df = 100 in order to use the appropriate distribution curve. It does not, however, list all possible degrees of freedom in this range, because that would take too many rows. Above df = 40, the rows jump in increments of 10. If a problem requires you to find critical values and the exact degrees of freedom is not listed, you always round down to the next smallest number. For example, if you have 48 people in your sample, the degrees of freedom are n – 1 = 48 – 1 = 47; however, 47 doesn't appear on our table, so we round down and use the critical values for df = 40, even though 50 is closer. We do this because it avoids inflating Type I Error (false positives, see chapter 7) by using criteria that are too lax.

This page titled 9.1: The t-statistic is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **8.1: The t-statistic** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

9.2: Hypothesis Testing with t

Hypothesis testing with the t-statistic works exactly the same way as z-tests did, following the four-step process of

1. Stating the Hypothesis
2. Finding the Critical Values
3. Computing the Test Statistic
4. Making the Decision.

We will work though an example: let's say that you move to a new city and find a an auto shop to change your oil. Your old mechanic did the job in about 30 minutes (though you never paid close enough attention to know how much that varied), and you suspect that your new shop takes much longer. After 4 oil changes, you think you have enough evidence to demonstrate this.

Step 1: State the Hypotheses Our hypotheses for 1-sample t-tests are identical to those we used for z-tests. We still state the null and alternative hypotheses mathematically in terms of the population parameter and written out in readable English. For our example:

H_0: There is no difference in the average time to change a car's oil

$$H_0 : \mu = 30$$

H_A: This shop takes longer to change oil than your old mechanic

$$H_A : \mu > 30$$

Step 2: Find the Critical Values As noted above, our critical values still delineate the area in the tails under the curve corresponding to our chosen level of significance. Because we have no reason to change significance levels, we will use α = 0.05, and because we suspect a direction of effect, we have a one-tailed test. To find our critical values for t, we need to add one more piece of information: the degrees of freedom. For this example:

$$df = N–1 = 4–1 = 3$$

Going to our t-table, we find the column corresponding to our one-tailed significance level and find where it intersects with the row for 3 degrees of freedom. As shown in Figure 9.2.1: our critical value is t_* = 2.353

t-distribution Table

df	0.05	0.025	0.01	0.005	1-tailed α
	0.10	0.05	0.02	0.01	2-tailed α
1	6.314	12.706	31.821	63.657	
2	2.920	4.303	6.965	9.925	
3	2.353	3.182	4.541	5.841	
4	2.132	2.776	3.747	4.604	
5	2.015	2.571	3.365	4.032	
6	1.943	2.447	3.143	3.707	

Figure 9.2.1: t-table

We can then shade this region on our t-distribution to visualize our rejection region

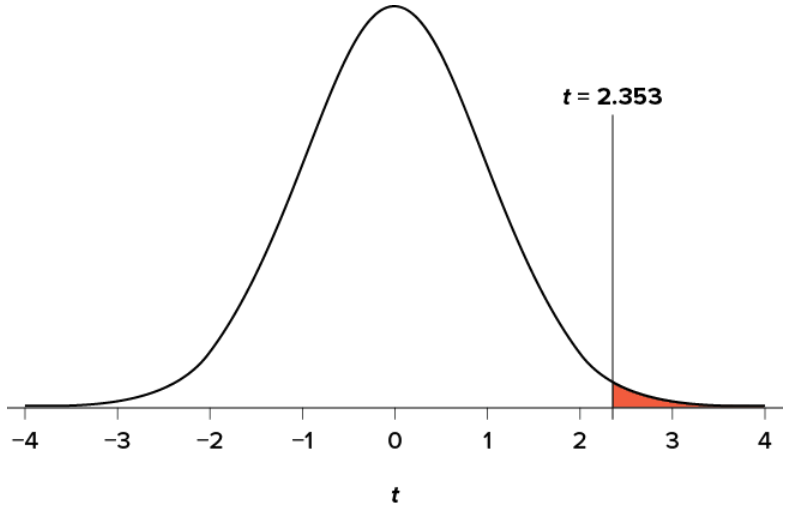


Figure 9.2.2: Rejection Region

Image Credit: Judy Schmitt, from Cote et al, 2021.

Step 3: Compute the Test Statistic The four wait times you experienced for your oil changes are the new shop were 46 minutes, 58 minutes, 40 minutes, and 71 minutes. We will use these to calculate $\overline{X}$ and s by first filling in the sum of squares table in Table 9.2.1:

Table 9.2.1: Sum of Squares Table

$\overline{X}$	$X-\overline{X}$	$(X-\overline{X})^2$
46	-7.75	60.06
58	4.25	18.06
40	-13.75	189.06
71	17.25	297.56
Σ=215	Σ=0	Σ=564.74

After filling in the first row to get Σ=215, we find that the mean is $\overline{X}$ = 53.75 (215 divided by sample size 4), which allows us to fill in the rest of the table to get our sum of squares SS = 564.74, which we then plug in to the formula for standard deviation from chapter 3:

$$s=\sqrt{\frac{\sum(X-\overline{X})^2}{N-1}}=\sqrt{\frac{SS}{df}}=\sqrt{\frac{564.74}{3}}=13.72$$

Next, we take this value and plug it in to the formula for standard error:

$$s_{\overline{X}}=\frac{s}{\sqrt{n}}=\frac{13.72}{2}=6.86$$

And, finally, we put the standard error, sample mean, and null hypothesis value into the formula for our test statistic t:

$$t=\frac{\overline{X}-\mu}{s_{\overline{X}}}=\frac{53.75-30}{6.86}=\frac{23.75}{6.68}=3.46$$

This may seem like a lot of steps, but it is really just taking our raw data to calculate one value at a time and carrying that value forward into the next equation: data sample size/degrees of freedom mean sum of squares standard deviation standard error test statistic. At each step, we simply match the symbols of what we just calculated to where they appear in the next formula to make sure we are plugging everything in correctly.

Step 4: Make the Decision Now that we have our critical value and test statistic, we can make our decision using the same criteria we used for a z-test. Our obtained t-statistic was t = 3.46 and our critical value was $t* = 2.353 : t > t*$, so we reject the null hypothesis and conclude:

Based on our four oil changes, the new mechanic takes longer on average ($\overline{X}$ = 53.75) to change oil than our old mechanic, $t(3)$ = 3.46, p < .05.

Notice that we also include the degrees of freedom in parentheses next to t. And because we found a significant result, we need to calculate an effect size, which is still Cohen's d, but now we use s in place of σ:

$$d = \frac{\overline{X} - \mu}{s} = \frac{53.75 - 30.00}{13.72} = 1.73$$

This is a large effect. It should also be noted that for some things, like the minutes in our current example, we can also interpret the magnitude of the difference we observed (23 minutes and 45 seconds) as an indicator of importance since time is a familiar metric.

This page titled 9.2: Hypothesis Testing with t is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **8.2: Hypothesis Testing with t** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

9.3: Confidence Intervals

Up to this point, we have learned how to estimate the population parameter for the mean using sample data and a sample statistic. From one point of view, this makes sense: we have one value for our parameter so we use a single value (called a point estimate) to estimate it. However, we have seen that all statistics have sampling error and that the value we find for the sample mean will bounce around based on the people in our sample, simply due to random chance. Thinking about estimation from this perspective, it would make more sense to take that error into account rather than relying just on our point estimate. To do this, we calculate what is known as a confidence interval.

A confidence interval starts with our point estimate then creates a range of scores considered plausible based on our standard deviation, our sample size, and the level of confidence with which we would like to estimate the parameter. This range, which extends equally in both directions away from the point estimate, is called the margin of error. We calculate the margin of error by multiplying our two-tailed critical value by our standard error:

$$\text{Margin of Error } = t^*(s/\sqrt{n}) \tag{9.3.1}$$

One important consideration when calculating the margin of error is that it can only be calculated using the critical value for a two-tailed test. This is because the margin of error moves away from the point estimate in both directions, so a one-tailed value does not make sense.

The critical value we use will be based on a chosen level of confidence, which is equal to 1 – α. Thus, a 95% level of confidence corresponds to α = 0.05. Thus, at the 0.05 level of significance, we create a 95% Confidence Interval. How to interpret that is discussed further on.

Once we have our margin of error calculated, we add it to our point estimate for the mean to get an upper bound to the confidence interval and subtract it from the point estimate for the mean to get a lower bound for the confidence interval:

$$\begin{aligned} \text{Upper Bound} &= \bar{X} + \text{Margin of Error} \\ \text{Lower Bound } &= \bar{X} - \text{Margin of Error} \end{aligned} \tag{9.3.2}$$

Or simply:

$$\text{Confidence Interval } = \overline{X} \pm t^*(s/\sqrt{n}) \tag{9.3.3}$$

To write out a confidence interval, we always use soft brackets and put the lower bound, a comma, and the upper bound:

$$\text{Confidence Interval } = \text{ (Lower Bound, Upper Bound)} \tag{9.3.4}$$

Let's see what this looks like with some actual numbers by taking our oil change data and using it to create a 95% confidence interval estimating the average length of time it takes at the new mechanic. We already found that our average was $\overline{X}$= 53.75 and our standard error was $s_{\overline{X}}$ = 6.86. We also found a critical value to test our hypothesis, but remember that we were testing a one-tailed hypothesis, so that critical value won't work. To see why that is, look at the column headers on the t-table. The column for one-tailed α = 0.05 is the same as a two-tailed α = 0.10. If we used the old critical value, we'd actually be creating a 90% confidence interval (1.00-0.10 = 0.90, or 90%). To find the correct value, we use the column for two-tailed α = 0.05 and, again, the row for 3 degrees of freedom, to find $t*$ = 3.182.

Now we have all the pieces we need to construct our confidence interval:

$$95\%CI = 53.75 \pm 3.182(6.86)$$

$$\begin{aligned} \text{Upper Bound} &= 53.75 + 3.182(6.86) \\ UB &= 53.75 + 21.83 \\ UB &= 75.58 \end{aligned}$$

$$\begin{aligned} \text{Lower Bound} &= 53.75 - 3.182(6.86) \\ LB &= 53.75 - 21.83 \\ LB &= 31.92 \end{aligned}$$

$$95\%CI = (31.92, 75.58)$$

So we find that our 95% confidence interval runs from 31.92 minutes to 75.58 minutes, but what does that actually mean? The range (31.92, 75.58) represents values of the mean that we consider reasonable or plausible based on our observed data. It includes our point estimate of the mean, $\overline{X}$= 53.75, in the center, but it also has a range of values that could also have been the case based on what we know about how much these scores vary (i.e. our standard error).

It is very tempting to also interpret this interval by saying that we are 95% confident that the true population mean falls within the range (31.92, 75.58), but this is not true. The reason it is not true is that phrasing our interpretation this way suggests that we have firmly established an interval and the population mean does or does not fall into it, suggesting that our interval is firm and the population mean will move around. However, the population mean is an absolute that does not change; it is our interval that will vary from data collection to data collection, even taking into account our standard error. The correct interpretation, then, is that we are 95% confident that the range (31.92, 75.58) brackets the true population mean. This is a very subtle difference, but it is an important one.

This page titled 9.3: Confidence Intervals is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **8.3: Confidence Intervals** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

9.E: Introduction to t-tests (Exercises)

1. What is the difference between a z-test and a 1-sample t-test?

Answer:

A z-test uses population standard deviation for calculating standard error and gets critical values based on the standard normal distribution. A t-test uses sample standard deviation as an estimate when calculating standard error and gets critical values from the t-distribution based on degrees of freedom.

2. What does a confidence interval represent?
3. What is the relationship between a chosen level of confidence for a confidence interval and how wide that interval is? For instance, if you move from a 95% CI to a 90% CI, what happens? Hint: look at the t-table to see how critical values change when you change levels of significance.

Answer:

As the level of confidence gets higher, the interval gets wider. In order to speak with more confidence about having found the population mean, you need to cast a wider net. This happens because critical values for higher confidence levels are larger, which creates a wider margin of error.

4. Construct a confidence interval around the sample mean $\overline{X}$= 25 for the following conditions:
   a. N = 25, s = 15, 95% confidence level
   b. N = 25, s = 15, 90% confidence level
   c. $s_{\overline{X}}$ = 4.5, α = 0.05, df = 20
   d. s = 12, df = 16 (yes, that is all the information you need)
5. True or False: a confidence interval represents the most likely location of the true population mean.

Answer:

False: a confidence interval is a range of plausible scores that may or may not bracket the true population mean.

6. You hear that college campuses may differ from the general population in terms of political affiliation, and you want to use hypothesis testing to see if this is true and, if so, how big the difference is. You know that the average political affiliation in the nation is μ = 4.00 on a scale of 1.00 to 7.00, so you gather data from 150 college students across the nation to see if there is a difference. You find that the average score is 3.76 with a standard deviation of 1.52. Use a 1-sample t-test to see if there is a difference at the α = 0.05 level.
7. You hear a lot of talk about increasing global temperature, so you decide to see for yourself if there has been an actual change in recent years. You know that the average land temperature from 1951-1980 was 8.79 degrees Celsius. You find annual average temperature data from 1981-2017 and decide to construct a 99% confidence interval (because you want to be as sure as possible and look for differences in both directions, not just one) using this data to test for a difference from the previous average.

Year	Temp	Year	Temp	Year	Temp	Year	Temp
1981	9.301	1991	9.336	2001	9.542	2011	9.65
1982	8.788	1992	8.974	2002	9.695	2012	9.635
1983	9.173	1993	9.008	2003	9.649	2013	9.753
1984	8.824	1994	9.175	2004	9.451	2014	9.714
1985	8.799	1995	9.484	2005	9.829	2015	9.962
1986	8.985	1996	9.168	2006	9.662	2016	10.16
1987	9.141	1997	9.326	2007	9.876	2017	10.049
1988	9.345	1998	9.66	2008	9.581		
1989	9.076	1999	9.406	2009	9.657		
1990	9.378	2000	9.332	2010	9.828		

Answer:

$\overline{X}$= 9.44, s = 0.35, $s_{\overline{X}}$= 0.06, df = 36, $t*$ = 2.719, 99% CI = (9.28, 9.60); CI does not bracket μ, reject null hypothesis. d = 1.83

8. Determine whether you would reject or fail to reject the null hypothesis in the following situations:
   a. t = 2.58, N = 21, two-tailed test at α = 0.05

b. t = 1.99, N = 49, one-tailed test at α = 0.01
c. μ = 47.82, 99% CI = (48.71, 49.28)
d. μ = 0, 95% CI = (-0.15, 0.20)

9. You are curious about how people feel about craft beer, so you gather data from 55 people in the city on whether or not they like it. You code your data so that 0 is neutral, positive scores indicate liking craft beer, and negative scores indicate disliking craft beer. You find that the average opinion was $\overline{X}$= 1.10 and the spread was s = 0.40, and you test for a difference from 0 at the α = 0.05 level.

Answer:

Step 1: $H_0 \neq 0$ "The average person has a neutral opinion towards craft beer", $H_A \neq 0$ "Overall people will have an opinion about craft beer, either good or bad."

Step 2: Two-tailed test, df = 54, $t*$ = 2.009.

Step 3: $\overline{X}$= 1.10, $s_{\overline{X}}$= 0.05, t = 22.00.

Step 4: $t > t*$, Reject H_0.

Based on opinions from 55 people, we can conclude that the average opinion of craft beer ($\overline{X}$= 1.10) is positive, $t(54)$ = 22.00, p < .05. Since the result is significant, we need an effect size: Cohen's d = 2.75, which is a large effect.

10. You want to know if college students have more stress in their daily lives than the general population (μ = 12), so you gather data from 25 people to test your hypothesis. Your sample has an average stress score of $\overline{X}$= 13.11 and a standard deviation of s = 3.89. Use a 1-sample t-test to see if there is a difference.

This page titled 9.E: Introduction to t-tests (Exercises) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **8.E: Introduction to t-tests (Exercises)** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

10: Repeated Measures

10.1: Change and Differences
10.2: Hypotheses of Change and Differences
10.3: Increasing Satisfaction at Work
10.4: Bad Press
10.E: Repeated Measures (Exercises)

This page titled 10: Repeated Measures is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

10.1: Change and Differences

Researchers are often interested in change over time. Sometimes we want to see if change occurs naturally, and other times we are hoping for change in response to some manipulation. In each of these cases, we measure a single variable at different times, and what we are looking for is whether or not we get the same score at time 2 as we did at time 1. The absolute value of our measurements does not matter – all that matters is the change. Let's look at an example:

Table **10.1.1**: Raw and difference scores before and after training.

Before	After	Improvement
6	9	3
7	7	0
4	10	6
1	3	2
8	10	2

Table **10.1.1** shows scores on a quiz that five employees received before they took a training course and after they took the course. The difference between these scores (i.e. the score after minus the score before) represents improvement in the employees' ability. This third column is what we look at when assessing whether or not our training was effective. We want to see positive scores, which indicate that the employees' performance went up. What we are not interested in is how good they were before they took the training or after the training. Notice that the lowest scoring employee before the training (with a score of 1) improved just as much as the highest scoring employee before the training (with a score of 8), regardless of how far apart they were to begin with. There's also one improvement score of 0, meaning that the training did not help this employee. An important factor in this is that the participants received the same assessment at both time points. To calculate improvement or any other difference score, we must measure only a single variable.

When looking at change scores like the ones in Table **10.1.1**, we calculate our difference scores by taking the time 2 score and subtracting the time 1 score. That is:

$$\mathbf{X_d = X_{T2} - X_{T1}} \tag{10.1.1}$$

Where $\mathbf{X_d}$ is the difference score, $\mathbf{X_{T1}}$ is the score on the variable at time 1, and $\mathbf{X_{T2}}$ is the score on the variable at time 2. The difference score, $\mathbf{X_d}$, will be the data we use to test for improvement or change. We subtract time 2 minus time 1 for ease of interpretation; if scores get better, then the difference score will be positive. Similarly, if we're measuring something like reaction time or depression symptoms that we are trying to reduce, then better outcomes (lower scores) will yield negative difference scores.

We can also test to see if people who are matched or paired in some way agree on a specific topic. For example, we can see if a parent and a child agree on the quality of home life, or we can see if two romantic partners agree on how serious and committed their relationship is. In these situations, we also subtract one score from the other to get a difference score. This time, however, it doesn't matter which score we subtract from the other because what we are concerned with is the agreement.

In both of these types of data, what we have are multiple scores on a single variable. That is, a single observation or data point is comprised of two measurements that are put together into one difference score. This is what makes the analysis of change unique – our ability to link these measurements in a meaningful way. This type of analysis would not work if we had two separate samples of people that weren't related at the individual level, such as samples of people from different states that we gathered independently. Such datasets and analyses are the subject of the following chapter.

A rose by any other name…

It is important to point out that this form of t-test has been called many different things by many different people over the years: "matched pairs", "paired samples", "repeated measures", "dependent measures", "dependent samples", and many others. What all of these names have in common is that they describe the analysis of two scores that are related in a systematic way within people or

within pairs, which is what each of the datasets usable in this analysis have in common. As such, all of these names are equally appropriate, and the choice of which one to use comes down to preference. In this text, we will refer to paired samples, though the appearance of any of the other names throughout this chapter should not be taken to refer to a different analysis: they are all the same thing.

Now that we have an understanding of what difference scores are and know how to calculate them, we can use them to test hypotheses. As we will see, this works exactly the same way as testing hypotheses about one sample mean with a tstatistic. The only difference is in the format of the null and alternative hypotheses.

This page titled 10.1: Change and Differences is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **9.1: Change and Differences** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

10.2: Hypotheses of Change and Differences

When we work with difference scores, our research questions have to do with change. Did scores improve? Did symptoms get better? Did prevalence go up or down? Our hypotheses will reflect this. Remember that the null hypothesis is the idea that there is nothing interesting, notable, or impactful represented in our dataset. In a paired samples t-test, that takes the form of 'no change'. There is no improvement in scores or decrease in symptoms. Thus, our null hypothesis is:

H_0: There is no change or difference

$$H_0 : \mu D = 0$$

As with our other null hypotheses, we express the null hypothesis for paired samples t-tests in both words and mathematical notation. The exact wording of the written-out version should be changed to match whatever research question we are addressing (e.g. " There is no change in ability scores after training"). However, the mathematical version of the null hypothesis is always exactly the same: the average change score is equal to zero. Our population parameter for the average is still μ, but it now has a subscript D to denote the fact that it is the average change score and not the average raw observation before or after our manipulation. Obviously individual difference scores can go up or down, but the null hypothesis states that these positive or negative change values are just random chance and that the true average change score across all people is 0.

Our alternative hypotheses will also follow the same format that they did before: they can be directional if we suspect a change or difference in a specific direction, or we can use an inequality sign to test for any change:

H_A: There is a change or difference

$$H_A : \mu D \neq 0$$

H_A: The average score increases

$$H_A : \mu D > 0$$

H_A: The average score decreases

$$H_A : \mu D < 0$$

As before, you choice of which alternative hypothesis to use should be specified before you collect data based on your research question and any evidence you might have that would indicate a specific directional (or non-directional) change.

Critical Values and Decision Criteria

As with before, once we have our hypotheses laid out, we need to find our critical values that will serve as our decision criteria. This step has not changed at all from the last chapter. Our critical values are based on our level of significance (still usually α = 0.05), the directionality of our test (one-tailed or two-tailed), and the degrees of freedom, which are still calculated as $df = n–1$. Because this is a t-test like the last chapter, we will find our critical values on the same t-table using the same process of identifying the correct column based on our significance level and directionality and the correct row based on our degrees of freedom or the next lowest value if our exact degrees of freedom are not presented. After we calculate our test statistic, our decision criteria are the same as well: $p < \alpha$ or $t_{obt} > t*$.

Test Statistic

Our test statistic for our change scores follows exactly the same format as it did for our 1-sample t-test. In fact, the only difference is in the data that we use. For our change test, we first calculate a difference score as shown above. Then, we use those scores as the raw data in the same mean calculation, standard error formula, and t-statistic. Let's look at each of these.

The mean difference score is calculated in the same way as any other mean: sum each of the individual difference scores and divide by the sample size.

$$\overline{X_D} = \frac{\Sigma X_D}{n} \tag{10.2.1}$$

Here we are using the subscript D to keep track of that fact that these are difference scores instead of raw scores; it has no actual effect on our calculation. Using this, we calculate the standard deviation of the difference scores the same way as well:

$$s_D = \sqrt{\frac{\sum\left(X_D - \overline{X_D}\right)^2}{n-1}} = \sqrt{\frac{SS}{df}} \tag{10.2.2}$$

We will find the numerator, the Sum of Squares, using the same table format that we learned in chapter 3. Once we have our standard deviation, we can find the standard error:

$$s_{\overline{X}_D} = {}^{S_D}/\sqrt{n} \tag{10.2.3}$$

Finally, our test statistic t has the same structure as well:

$$t = \frac{\overline{X_D} - \mu_D}{s_{\overline{X}_D}} \tag{10.2.4}$$

As we can see, once we calculate our difference scores from our raw measurements, everything else is exactly the same. Let's see an example.

This page titled 10.2: Hypotheses of Change and Differences is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **9.2: Hypotheses of Change and Differences** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

10.3: Increasing Satisfaction at Work

Workers at a local company have been complaining that working conditions have gotten very poor, hours are too long, and they don't feel supported by the management. The company hires a consultant to come in and help fix the situation before it gets so bad that the employees start to quit. The consultant first assesses 40 of the employee's level of job satisfaction as part of focus groups used to identify specific changes that might help. The company institutes some of these changes, and six months later the consultant returns to measure job satisfaction again. Knowing that some interventions miss the mark and can actually make things worse, the consultant tests for a difference in either direction (i.e. and increase or a decreased in average job satisfaction) at the α = 0.05 level of significance.

Step 1: State the Hypotheses First, we state our null and alternative hypotheses:

H_0: There is no change in average job satisfaction

$$H_0 : \mu D = 0$$

H_A: There is an increase in average job satisfaction

$$H_A : \mu D > 0$$

In this case, we are hoping that the changes we made will improve employee satisfaction, and, because we based the changes on employee recommendations, we have good reason to believe that they will. Thus, we will use a one-directional alternative hypothesis.

Step 2: Find the Critical Values Our critical values will once again be based on our level of significance, which we know is α = 0.05, the directionality of our test, which is one-tailed to the right, and our degrees of freedom. For our dependent-samples t-test, the degrees of freedom are still given as $df = n–1$. For this problem, we have 40 people, so our degrees of freedom are 39. Going to our t-table, we find that the critical value is $t*$ = 1.685 as shown in Figure **10.3.1**.

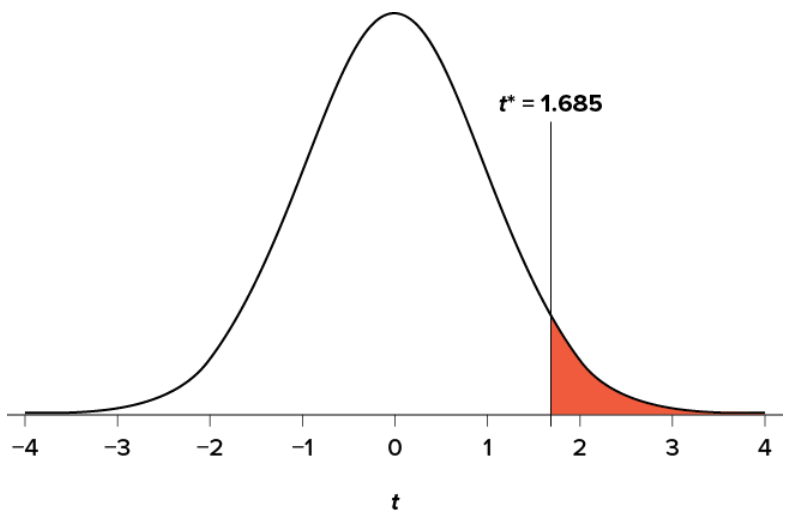


Figure **10.3.1**: Critical region for one-tailed t-test at α = 0.05

Image Credit: Judy Schmitt, from Cote et al, 2021.

Step 3: Calculate the Test Statistic Now that the criteria are set, it is time to calculate the test statistic. The data obtained by the consultant found that the difference scores from time 1 to time 2 had a mean of $\overline{X_D}$ = 2.96 and a standard deviation of s_D = 2.85. Using this information, plus the size of the sample (N = 40), we first calculate the standard error:

$$s_{\overline{x_D}} = s_{D/\sqrt{n}} = 2.85/\sqrt{40} = 2.85/6.32 = 0.46$$

Now, we can put that value, along with our sample mean and null hypothesis value, into the formula for t and calculate the test statistic:

$$t = \frac{\overline{X_D} - \mu_D}{s_{\overline{X}_D}} = \frac{2.96 - 0}{0.46} = 6.43$$

Notice that, because the null hypothesis value of a dependent samples t-test is always 0, we can simply divide our obtained sample mean by the standard error.

Step 4: Make the Decision We have obtained a test statistic of t = 6.43 that we can compare to our previously established critical value of $t*$ = 1.685. 6.43 is larger than 1.685, so $t > t*$ and we reject the null hypothesis:

> Reject H_0. Based on the sample data from 40 workers, we can say that the intervention statistically significantly improved job satisfaction ($\overline{X_D}$= 2.96) among the workers, $t(39) = 6.43$, $p < 0.05$.

Because this result was statistically significant, we will want to calculate Cohen's d as an effect size using the same format as we did for the last t-test:

$$d = \frac{\overline{X_D} - \mu_D}{s_D} = \frac{2.96}{2.85} = 1.04$$

This is a large effect size. Notice again that we can omit the null hypothesis value here because it is always equal to 0.

Hopefully the above example made it clear that running a dependent samples t-test to look for differences before and after some treatment works exactly the same way as a regular 1-sample t-test does, which was just a small change in how z-tests were performed in chapter 7. At this point, this process should feel familiar, and we will continue to make small adjustments to this familiar process as we encounter new types of data to test new types of research questions.

This page titled 10.3: Increasing Satisfaction at Work is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **9.3: Increasing Satisfaction at Work** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

10.4: Bad Press

Let's say that a bank wants to make sure that their new commercial will make them look good to the public, so they recruit 7 people to view the commercial as a focus group. The focus group members fill out a short questionnaire about how they view the company, then watch the commercial and fill out the same questionnaire a second time. The bank really wants to find significant results, so they test for a change at α = 0.10. However, they use a 2-tailed test since they know that past commercials have not gone over well with the public, and they want to make sure the new one does not backfire. They decide to test their hypothesis using a confidence interval to see just how spread out the opinions are. As we will see, confidence intervals work the same way as they did before, just like with the test statistic.

Step 1: State the Hypotheses As always, we start with hypotheses:

H_0: There is no change in how people view the bank

$$H_0 : \mu D = 0$$

H_A: There is a change in how people view the bank

$$H_A : \mu D \neq 0$$

Step 2: Find the Critical Values Just like with our regular hypothesis testing procedure, we will need critical values from the appropriate level of significance and degrees of freedom in order to form our confidence interval. Because we have 7 participants, our degrees of freedom are df = 6. From our t-table, we find that the critical value corresponding to this df at this level of significance is $t*$ = 1.943.

Step 3: Calculate the Confidence Interval The data collected before (time 1) and after (time 2) the participants viewed the commercial is presented in Table 10.4.1. In order to build our confidence interval, we will first have to calculate the mean and standard deviation of the difference scores, which are also in Table 10.4.1. As a reminder, the difference scores are calculated as Time 2 – Time 1.

Table 10.4.1: Opinions of the bank

Time 1	Time 2	X_D
3	2	-1
3	6	3
5	3	-2
8	4	-4
3	9	6
1	2	1
4	5	1

The mean of the difference scores is:

$$\overline{X_D} = \frac{\sum X_D}{n} = \frac{4}{7} = 0.57$$

The standard deviation will be solved by first using the Sum of Squares Table:

Table 10.4.2: Sum of Squares

X_D	$X_D - \overline{X_D}$	$(X_D - \overline{X_D})^2$
-1	-1.57	2.46
3	2.43	5.90
-2	-2.57	6.60
-4	-4.57	20.88
6	5.43	29.48
1	0.43	0.18
1	0.43	0.18
$\Sigma = 4$	$\Sigma = 0$	$\Sigma = 65.68$

$$s_D = \sqrt{\frac{SS}{df}} = \sqrt{\frac{65.68}{6}} = \sqrt{10.95} = 3.31$$

Finally, we find the standard error:

$$s_{\overline{X_D}} = {}^{S_D} / \sqrt{n} = 3.31/\sqrt{7} = 1.25$$

We now have all the pieces needed to compute our confidence interval:

$$95\%CI = \overline{X_D} \pm t^* \left(s_{\overline{X_D}}\right)$$
$$95\%CI = 0.57 \pm 1.943(1.25)$$

$$\text{Upper Bound} = 0.57 + 1.943(1.25)$$
$$UB = 0.57 + 2.43$$
$$UB = 3.00$$

$$\text{Lower Bound} = 0.57 - 1.943(1.25)$$
$$LB = 0.57 - 2.43$$
$$LB = -1.86$$

$$95\%CI = (-1.86, 3.00)$$

Step 4: Make the Decision Remember that the confidence interval represents a range of values that seem plausible or reasonable based on our observed data. The interval spans -1.86 to 3.00, which includes 0, our null hypothesis value. Because the null hypothesis value is in the interval, it is considered a reasonable value, and because it is a reasonable value, we have no evidence against it. We fail to reject the null hypothesis.

Fail to Reject H_0. Based on our focus group of 7 people, we cannot say that the average change in opinion ($\overline{X_D}$ = 0.57) was any better or worse after viewing the commercial, CI: (-1.86, 3.00).

As with before, we only report the confidence interval to indicate how we performed the test.

This page titled 10.4: Bad Press is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **9.4: Bad Press** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

10.E: Repeated Measures (Exercises)

1. What is the difference between a 1-sample t-test and a dependent-samples t-test? How are they alike?

Answer:

A 1-sample t-test uses raw scores to compare an average to a specific value. A dependent samples t-test uses two raw scores from each person to calculate difference scores and test for an average difference score that is equal to zero. The calculations, steps, and interpretation is exactly the same for each.

2. Name 3 research questions that could be addressed using a dependent samples t-test.
3. What are difference scores and why do we calculate them?

Answer:

Difference scores indicate change or discrepancy relative to a single person or pair of people. We calculate them to eliminate individual differences in our study of change or agreement.

4. Why is the null hypothesis for a dependent-samples t-test always $\mu_D = 0$?
5. A researcher is interested in testing whether explaining the processes of statistics helps increase trust in computer algorithms. He wants to test for a difference at the α = 0.05 level and knows that some people may trust the algorithms less after the training, so he uses a two-tailed test. He gathers pre-post data from 35 people and finds that the average difference score is $\overline{X_D}$ = 12.10 with a standard deviation of s_D= 17.39. Conduct a hypothesis test to answer the research question.

Answer:

Step 1: $H_0 : \mu = 0$ "The average change in trust of algorithms is 0", $H_A : \mu \neq 0$ "People's opinions of how much they trust algorithms changes."

Step 2: Two-tailed test, df = 34, $t*$ = 2.032.

Step 3: $\overline{X_D}$ = 12.10, $s_{\overline{X_D}}$= 2.94, t = 4.12.

Step 4: $t > t*$, Reject H_0. Based on opinions from 35 people, we can conclude that people trust algorithms more ($\overline{X_D}$= 12.10) after learning statistics, $t(34) = 4.12, p < .05$. Since the result is significant, we need an effect size: Cohen's d = 0.70, which is a moderate to large effect.

6. Decide whether you would reject or fail to reject the null hypothesis in the following situations:
   a. $\overline{X_D}$ = 3.50, s_D = 1.10, n = 12, α = 0.05, two-tailed test
   b. 95% CI= (0.20,1.85)
   c. t = 2.98, $t*$ = -2.36, one-tailed test to the left
   d. 90% CI = (-1.12, 4.36)
7. Calculate difference scores for the following data:

Time 1	Time 2	X_D
61	83	
75	89	
91	98	
83	92	
74	80	
82	88	
98	98	
82	77	
69	88	
76	79	
91	91	
70	80	

Answer:

Time 1	Time 2	X_D
61	83	22
75	89	14
91	98	7
83	92	9
74	80	6
82	88	6
98	98	0
82	77	-5
69	88	19
76	79	3
91	91	0
70	80	10

8. You want to know if an employee's opinion about an organization is the same as the opinion of that employee's boss. You collect data from 18 employee-supervisor pairs and code the difference scores so that positive scores indicate that the employee has a higher opinion and negative scores indicate that the boss has a higher opinion (meaning that difference scores of 0 indicate no difference and complete agreement). You find that the mean difference score is $\overline{X_D}$= -3.15 with a standard deviation of s_D = 1.97. Test this hypothesis at the α = 0.01 level.
9. Construct confidence intervals from a mean of $\overline{X_D}$ = 1.25, standard error of $s_{\overline{X_D}}$= 0.45, and df = 10 at the 90%, 95%, and 99% confidence level. Describe what happens as confidence changes and whether to reject H_0.

Answer:

At the 90% confidence level, $t*$ = 1.812 and CI = (0.43, 2.07) so we reject H_0. At the 95% confidence level, $t*$ = 2.228 and CI = (0.25, 2.25) so we reject H_0. At the 99% confidence level, $t*$ = 3.169 and CI = (-0.18, 2.68) so we fail to reject H_0. As the confidence level goes up, our interval gets wider (which is why we have higher confidence), and eventually we do not reject the null hypothesis because the interval is so wide that it contains 0.

10. A professor wants to see how much students learn over the course of a semester. A pre-test is given before the class begins to see what students know ahead of time, and the same test is given at the end of the semester to see what students know at the end. The data are below. Test for an improvement at the α = 0.05 level. Did scores increase? How much did scores increase?

Pretest	Posttest	X_D
90	89	
60	66	
95	99	
93	91	
95	100	
67	64	
89	91	
90	95	
94	95	
83	89	
75	82	
87	92	
82	83	
82	85	
88	93	
66	69	
90	90	
93	100	
86	95	
91	96	

This page titled 10.E: Repeated Measures (Exercises) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **9.E: Repeated Measures (Exercises)** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

11: Independent Samples

11.1: Difference of Means
11.2: Research Questions about Independent Means
11.3: Hypotheses and Decision Criteria
11.4: Independent Samples t-statistic
11.5: Standard Error and Pooled Variance
11.6: Movies and Mood
11.7: Effect Sizes and Confidence Intervals
11.8: Homogeneity of Variance
11.E: Independent Samples (Exercises)

This page titled 11: Independent Samples is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

11.1: Difference of Means

Last chapter, we learned about mean differences, that is, the average value of difference scores. Those difference scores came from ONE group and TWO time points (or two perspectives). Now, we will deal with the difference of the means, that is, the average values of separate groups that are represented by separate descriptive statistics. This analysis involves TWO groups and ONE time point. As with all of our other tests as well, both of these analyses are concerned with a single variable.

It is very important to keep these two tests separate and understand the distinctions between them because they assess very different questions and require different approaches to the data. When in doubt, think about how the data were collected and where they came from. If they came from two time points with the same people (sometimes referred to as "longitudinal" data), you know you are working with repeated measures data (the measurement literally was repeated) and will use a repeated measures/dependent samples t-test. If it came from a single time point that used separate groups, you need to look at the nature of those groups and if they are related. Can individuals in one group being meaningfully matched up with one and only one individual from the other group? For example, are they a romantic couple? If so, we call those data matched and we use a matched pairs/dependent samples t-test. However, if there's no logical or meaningful way to link individuals across groups, or if there is no overlap between the groups, then we say the groups are independent and use the independent samples t-test, the subject of this chapter.

This page titled 11.1: Difference of Means is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **10.1: Difference of Means** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

11.2: Research Questions about Independent Means

Many research ideas in the behavioral sciences and other areas of research are concerned with whether or not two means are the same or different. Logically, we therefore say that these research questions are concerned with group mean differences. That is, on average, do we expect a person from Group A to be higher or lower on some variable that a person from Group B. In any time of research design looking at group mean differences, there are some key criteria we must consider: the groups must be mutually exclusive (i.e. you can only be part of one group at any given time) and the groups have to be measured on the same variable (i.e. you can't compare personality in one group to reaction time in another group since those values would not be the same anyway).

Let's look at one of the most common and logical examples: testing a new medication. When a new medication is developed, the researchers who created it need to demonstrate that it effectively treats the symptoms they are trying to alleviate. The simplest design that will answer this question involves two groups: one group that receives the new medication (the "treatment" group) and one group that receives a placebo (the "control" group). Participants are randomly assigned to one of the two groups (remember that random assignment is the hallmark of a true experiment), and the researchers test the symptoms in each person in each group after they received either the medication or the placebo. They then calculate the average symptoms in each group and compare them to see if the treatment group did better (i.e. had fewer or less severe symptoms) than the control group.

In this example, we had two groups: treatment and control. Membership in these two groups was mutually exclusive: each individual participant received either the experimental medication or the placebo. No one in the experiment received both, so there was no overlap between the two groups. Additionally, each group could be measured on the same variable: symptoms related to the disease or ailment being treated. Because each group was measured on the same variable, the average scores in each group could be meaningfully compared. If the treatment was ineffective, we would expect that the average symptoms of someone receiving the treatment would be the same as the average symptoms of someone receiving the placebo (i.e. there is no difference between the groups). However, if the treatment WAS effective, we would expect fewer symptoms from the treatment group, leading to a lower group average.

Now let's look at an example using groups that already exist. A common, and perhaps salient, question is how students feel about their job prospects after graduation. Suppose that we have narrowed our potential choice of college down to two universities and, in the course of trying to decide between the two, we come across a survey that has data from each university on how students at those universities feel about their future job prospects. As with our last example, we have two groups: University A and University B, and each participant is in only one of the two groups (assuming there are no transfer students who were somehow able to rate both universities). Because students at each university completed the same survey, they are measuring the same thing, so we can use a t-test to compare the average perceptions of students at each university to see if they are the same. If they are the same, then we should continue looking for other things about each university to help us decide on where to go. But, if they are different, we can use that information in favor of the university with higher job prospects.

As we can see, the grouping variable we use for an independent samples t-test can be a set of groups we create (as in the experimental medication example) or groups that already exist naturally (as in the university example). There are countless other examples of research questions relating to two group means, making the independent samples t-test one of the most widely used analyses around.

This page titled 11.2: Research Questions about Independent Means is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **10.2: Research Questions about Independent Means** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

11.3: Hypotheses and Decision Criteria

The process of testing hypotheses using an independent samples t-test is the same as it was in the last three chapters, and it starts with stating our hypotheses and laying out the criteria we will use to test them.

Our null hypothesis for an independent samples t-test is the same as all others: there is no difference. The means of the two groups are the same under the null hypothesis, no matter how those groups were formed. Mathematically, this takes on two equivalent forms:

$$H_0 : \mu_1 = \mu_2$$

or

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

Both of these formulations of the null hypothesis tell us exactly the same thing: that the numerical value of the means is the same in both groups. This is more clear in the first formulation, but the second formulation also makes sense (any number minus itself is always zero) and helps us out a little when we get to the math of the test statistic. Either one is acceptable and you only need to report one. The English interpretation of both of them is also the same:

$$H_0 : \text{ There is no difference between the means of the two groups}$$

Our alternative hypotheses are also unchanged: we simply replace the equal sign (=) with one of the three inequalities (>, <, ≠):

$$H_A : \mu_1 > \mu_2$$
$$H_A : \mu_1 < \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

Or

$$H_A : \mu_1 - \mu_2 > 0$$
$$H_A : \mu_1 - \mu_2 < 0$$
$$H_A : \mu_1 - \mu_2 \neq 0$$

Whichever formulation you chose for the null hypothesis should be the one you use for the alternative hypothesis (be consistent), and the interpretation of them is always the same:

$$H_A : \text{There is a difference between the means of the two groups}$$

Notice that we are now dealing with two means instead of just one, so it will be very important to keep track of which mean goes with which population and, by extension, which dataset and sample data. We use subscripts to differentiate between the populations, so make sure to keep track of which is which. If it is helpful, you can also use more descriptive subscripts. To use the experimental medication example:

H_0: There is no difference between the means of the treatment and control groups

$$H_0 : \mu_{\text{treatment}} = \mu_{\text{control}}$$

H_A: There is a difference between the means of the treatment and control groups

$$H_A : \mu_{\text{treatment}} \neq \mu_{\text{control}}$$

Once we have our hypotheses laid out, we can set our criteria to test them using the same three pieces of information as before: significance level (α), directionality (left, right, or two-tailed), and degrees of freedom, which for an independent samples t-test are:

$$df = n_1 + n_2 - 2$$

This looks different than before, but it is just adding the individual degrees of freedom from each group (n–1) together. Notice that the sample sizes, n, also get subscripts so we can tell them apart.

For an independent samples t-test, it is often the case that our two groups will have slightly different sample sizes, either due to chance or some characteristic of the groups themselves. Generally, this is not as issue, so long as one group is not massively larger

than the other group. What is of greater concern is keeping track of which is which using the subscripts.

This page titled 11.3: Hypotheses and Decision Criteria is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **10.3: Hypotheses and Decision Criteria** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

11.4: Independent Samples t-statistic

The test statistic for our independent samples t-test takes on the same logical structure and format as our other t-tests: our observed effect minus our null hypothesis value, all divided by the standard error:

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{s_{\overline{X_1} - \overline{X_2}}} \tag{11.4.1}$$

This looks like more work to calculate, but remember that our null hypothesis states that the quantity $\mu_1 - \mu_2 = 0$, so we can drop that out of the equation and are left with:

$$t = \frac{(\overline{X_1} - \overline{X_2})}{s_{\overline{X_1} - \overline{X_2}}} \tag{11.4.2}$$

Our standard error in the denomination is still standard deviation (s) with a subscript denoting what it is the standard error of. Because we are dealing with the difference between two separate means, rather than a single mean or single mean of difference scores, we put both means in the subscript. Calculating our standard error, as we will see next, is where the biggest differences between this t-test and other t-tests appears. However, once we do calculate it and use it in our test statistic, everything else goes back to normal. Our decision criteria is still comparing our obtained test statistic to our critical value, and our interpretation based on whether or not we reject the null hypothesis is unchanged as well.

This page titled 11.4: Independent Samples t-statistic is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **10.4: Independent Samples t-statistic** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

11.5: Standard Error and Pooled Variance

Recall that the standard error is the average distance between any given sample mean and the center of its corresponding sampling distribution, and it is a function of the standard deviation of the population (either given or estimated) and the sample size. This definition and interpretation hold true for our independent samples t-test as well, but because we are working with two samples drawn from two populations, we have to first combine their estimates of standard deviation – or, more accurately, their estimates of variance – into a single value that we can then use to calculate our standard error.

The combined estimate of variance using the information from each sample is called the pooled variance and is denoted s_p^2; the subscript p serves as a reminder indicating that it is the pooled variance. The term "pooled variance" is a literal name because we are simply pooling or combining the information on variance – the Sum of Squares and Degrees of Freedom – from both of our samples into a single number. The result is a weighted average of the observed sample variances, the weight for each being determined by the sample size, and will always fall between the two observed variances. The computational formula for the pooled variance is:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} \tag{11.5.1}$$

This formula can look daunting at first, but it is in fact just a weighted average. Even more conveniently, some simple algebra can be employed to greatly reduce the complexity of the calculation. The simpler and more appropriate formula to use when calculating pooled variance is:

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} \tag{11.5.2}$$

Using this formula, it's very simple to see that we are just adding together the same pieces of information we have been calculating since chapter 3. Thus, when we use this formula, the pooled variance is not nearly as intimidating as it might have originally seemed.

Once we have our pooled variance calculated, we can drop it into the equation for our standard error:

$$S_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \tag{11.5.3}$$

Once again, although this formula may seem different than it was before, in reality it is just a different way of writing the same thing. An alternative but mathematically equivalent way of writing our old standard error is:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}} \tag{11.5.4}$$

Looking at that, we can now see that, once again, we are simply adding together two pieces of information: no new logic or interpretation required. Once the standard error is calculated, it goes in the denominator of our test statistic, as shown above and as was the case in all previous chapters. Thus, the only additional step to calculating an independent samples t-statistic is computing the pooled variance. Let's see an example in action.

This page titled 11.5: Standard Error and Pooled Variance is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **10.5: Standard Error and Pooled Variance** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

11.6: Movies and Mood

We are interested in whether the type of movie someone sees at the theater affects their mood when they leave. We decide to ask people about their mood as they leave one of two movies: a comedy (group 1, n = 35) or a horror film (group 2, n = 29). Our data are coded so that higher scores indicate a more positive mood. We have good reason to believe that people leaving the comedy will be in a better mood, so we use a one-tailed test at α = 0.05 to test our hypothesis.

Step 1: State the Hypotheses As always, we start with hypotheses:

H_0: There is no difference in average mood between the two movie types

$$\mathrm{H}_0 : \mu_1 - \mu_2 = 0$$

or

$$\mathrm{H}_0 : \mu_1 = \mu_2$$

H_A: The comedy film will give a better average mood than the horror film

$$\mathrm{H}_\mathrm{A} : \mu_1 - \mu_2 > 0$$

or

$$\mathrm{H}_\mathrm{A} : \mu_1 > \mu_2$$

Notice that in the first formulation of the alternative hypothesis we say that the first mean minus the second mean will be greater than zero. This is based on how we code the data (higher is better), so we suspect that the mean of the first group will be higher. Thus, we will have a larger number minus a smaller number, which will be greater than zero. Be sure to pay attention to which group is which and how your data are coded (higher is almost always used as better outcomes) to make sure your hypothesis makes sense!

Step 2: Find the Critical Values Just like before, we will need critical values, which come from out t-table. In this example, we have a one-tailed test at α = 0.05 and expect a positive answer (because we expect the difference between the means to be greater than zero). Our degrees of freedom for our independent samples t-test is just the degrees of freedom from each group added together: 35 + 29 – 2 = 62. From our t-table, we find that our critical value is $t*$ = 1.671. Note that because 62 does not appear on the table, we use the next lowest value, which in this case is 60.

Step 3: Compute the Test Statistic The data from our two groups are presented in the tables below. Table **11.6.1** shows the values for the Comedy group, and Table **11.6.2** shows the values for the Horror group. Values for both have already been placed in the Sum of Squares tables since we will need to use them for our further calculations. As always, the column on the left is our raw data.

Table **11.6.1**: Raw scores and Sum of Squares for Group 1

x	$x - \bar{x}$	$(x - \bar{x})^2$
Group 1: Comedy Film		
39.10	15.10	228.01
38.00	14.00	196.00
14.90	-9.10	82.81
20.70	-3.30	10.89
19.50	-4.50	20.25
32.20	8.20	67.24
11.00	-13.00	169.00
20.70	-3.30	10.89
26.40	2.40	5.76
35.70	11.70	136.89
26.40	2.40	5.76
28.80	4.80	23.04
33.40	9.40	88.36
13.70	-10.30	106.09
46.10	22.10	488.41
13.70	-10.30	106.09
23.00	-1.00	1.00
20.70	-3.30	10.89
19.50	-4.50	20.25
11.40	-12.60	158.76
24.10	0.10	0.01
17.20	-6.80	46.24
38.00	14.00	196.00
10.30	-13.70	187.69
35.70	11.70	136.89
41.50	17.50	306.25
18.40	-5.60	31.36
36.80	12.80	163.84
54.10	30.10	906.01
11.40	-12.60	158.76
8.70	-15.30	234.09
23.00	-1.00	1.00

X	$(X-\overline{X})$	$(X-\overline{X})^2$
14.30	-9.70	94.09
5.30	-18.70	349.69
6.30	-17.70	313.29
$\Sigma = 840$	$\Sigma = 0$	$\Sigma = 5061.60$

Table 11.6.2: Raw scores and Sum of Squares for Group 2

X	$(X-\overline{X})$	$(X-\overline{X})^2$
	Group 2: Horror Film	
24.00	7.50	56.25
17.00	0.50	0.25
35.80	19.30	372.49
18.00	1.50	2.25
-1.70	-18.20	331.24
11.10	-5.40	29.16
10.10	-6.40	40.96
16.10	-0.40	0.16
-0.70	-17.20	295.84
14.10	-2.40	5.76
25.90	9.40	88.36
23.00	6.50	42.25
20.00	3.50	12.25
14.10	-2.40	5.76
-1.70	-18.20	331.24
19.00	2.50	6.25
20.00	3.50	12.25
30.90	14.40	207.36
30.90	14.40	207.36
22.00	5.50	30.25
6.20	-10.30	106.09
27.90	11.40	129.96
14.10	-2.40	5.76
33.80	17.30	299.29
26.90	10.40	108.16
5.20	-11.30	127.69
13.10	-3.40	11.56
19.00	2.50	6.25
-15.50	-32.00	1024.00
$\Sigma = 478.6$	$\Sigma = 0.10$	$\Sigma = 3896.45$

Using the sum of the first column for each table, we can calculate the mean for each group:

$$\overline{X_1} = \frac{840}{35} = 24.00$$

And

$$\overline{X_2} = \frac{478.60}{29} = 16.50$$

These values were used to calculate the middle rows of each table, which sum to zero as they should (the middle column for group 2 sums to a very small value instead of zero due to rounding error – the exact mean is 16.50344827586207, but that's far more than we need for our purposes). Squaring each of the deviation scores in the middle columns gives us the values in the third columns, which sum to our next important value: the Sum of Squares for each group: SS_1 = 5061.60 and SS_2 = 3896.45. These values have all been calculated and take on the same interpretation as they have since chapter 3 – no new computations yet. Before we move on to the pooled variance that will allow us to calculate standard error, let's compute our standard deviation for each group; even though we will not use them in our calculation of the test statistic, they are still important descriptors of our data:

$$s_1 = \sqrt{\frac{5061.60}{34}} = 12.20$$

And

$$s_2 = \sqrt{\frac{3896.45}{28}} = 11.80$$

Now we can move on to our new calculation, the pooled variance, which is just the Sums of Squares that we calculated from our table and the degrees of freedom, which is just n–1 for each group:

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{5061.60 + 3896.45}{34 + 28} = \frac{8958.05}{62} = 144.48$$

As you can see, if you follow the regular process of calculating standard deviation using the Sum of Squares table, finding the pooled variance is very easy. Now we can use that value to calculate our standard error, the last step before we can find our test statistic:

$$s_{\overline{X_1}-\overline{X_2}} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{144.48}{35} + \frac{144.48}{29}} = \sqrt{4.13 + 4.98} = \sqrt{9.11} = 3.02$$

Finally, we can use our standard error and the means we calculated earlier to compute our test statistic. Because the null hypothesis value of μ_1–μ_2 is 0.00, we will leave that portion out of the equation for simplicity:

$$t = \frac{(\overline{X_1} - \overline{X_2})}{s_{\overline{X_1}-\overline{X_2}}} = \frac{24.00 - 16.50}{3.02} = 2.48$$

The process of calculating our obtained test statistic t = 2.48 followed the same sequence of steps as before: use raw data to compute the mean and sum of squares (this time for two groups instead of one), use the sum of squares and degrees of freedom to calculate standard error (this time using pooled variance instead of standard deviation), and use that standard error and the observed means to get t. Now we can move on to the final step of the hypothesis testing procedure.

Step 4: Make the Decision Our test statistic has a value of t = 2.48, and in step 2 we found that the critical value is $t*$ = 1.671. 2.48 > 1.671, so we reject the null hypothesis:

Reject H_0. Based on our sample data from people who watched different kinds of movies, we can say that the average mood after a comedy movie ($\overline{X_1} = 24.00$) is better than the average mood after a horror movie ($\overline{X_2} = 16.50$), $t(62) = 2.48, p < .05$.

This page titled 11.6: Movies and Mood is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **10.6: Movies and Mood** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

11.7: Effect Sizes and Confidence Intervals

We have seen in previous chapters that even a statistically significant effect needs to be interpreted along with an effect size to see if it is practically meaningful. We have also seen that our sample means, as a point estimate, are not perfect and would be better represented by a range of values that we call a confidence interval. As with all other topics, this is also true of our independent samples t-tests.

Our effect size for the independent samples t-test is still Cohen's d, and it is still just our observed effect divided by the standard deviation. Remember that standard deviation is just the square root of the variance, and because we work with pooled variance in our test statistic, we will use the square root of the pooled variance as our denominator in the formula for Cohen's d. This gives us:

$$d = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{s_p^2}} \tag{11.7.1}$$

For our example above, we can calculate the effect size to be:

$$d = \frac{24.00 - 16.50}{\sqrt{144.48}} = \frac{7.50}{12.02} = 0.62$$

We interpret this using the same guidelines as before, so we would consider this a moderate or moderately large effect.

Our confidence intervals also take on the same form and interpretation as they have in the past. The value we are interested in is the difference between the two means, so our point estimate is the value of one mean minus the other, or xbar1 minus xbar2. Just like before, this is our observed effect and is the same value as the one we place in the numerator of our test statistic. We calculate this value then place the margin of error – still our critical value times our standard error – above and below it. That is:

$$\text{Confidence Interval } = (\overline{X_1} - \overline{X_2}) \pm t^* \left(s_{\overline{X_1} - \overline{X_2}}\right) \tag{11.7.2}$$

Because our hypothesis testing example used a one-tailed test, it would be inappropriate to calculate a confidence interval on those data (remember that we can only calculate a confidence interval for a two-tailed test because the interval extends in both directions). Let's say we find summary statistics on the average life satisfaction of people from two different towns and want to create a confidence interval to see if the difference between the two might actually be zero.

Our sample data are $\overline{X_1} = 28.65$ $s_1 = 12.40$ $n_1 = 40$ and $\overline{X_2} = 25.40 s_2 = 15.68 n_2 = 42$. At face value, it looks like the people from the first town have higher life satisfaction (28.65 vs. 25.40), but it will take a confidence interval (or complete hypothesis testing process) to see if that is true or just due to random chance. First, we want to calculate the difference between our sample means, which is 28.65 – 25.40 = 3.25. Next, we need a critical value from our t-table. If we want to test at the normal 95% level of confidence, then our sample sizes will yield degrees of freedom equal to 40 + 42 – 2 = 80. From our table, that gives us a critical value of $t*$ = 1.990. Finally, we need our standard error. Recall that our standard error for an independent samples t-test uses pooled variance, which requires the Sum of Squares and degrees of freedom. Up to this point, we have calculated the Sum of Squares using raw data, but in this situation, we do not have access to it. So, what are we to do?

If we have summary data like standard deviation and sample size, it is very easy to calculate the pooled variance, and the key lies in rearranging the formulas to work backwards through them. We need the Sum of Squares and degrees of freedom to calculate our pooled variance. Degrees of freedom is very simple: we just take the sample size minus 1.00 for each group. Getting the Sum of Squares is also easy: remember that variance is standard deviation squared and is the Sum of Squares divided by the degrees of freedom. That is:

$$s^2 = (s)^2 = \frac{SS}{df} \tag{11.7.3}$$

To get the Sum of Squares, we just multiply both sides of the above equation to get:

$$s^2 * df = SS \tag{11.7.4}$$

Which is the squared standard deviation multiplied by the degrees of freedom $(n - 1)$ equals the Sum of Squares.

Using our example data:

$$(s_1)^2 * df_1 = SS_1$$
$$(12.40)^2 * (40-1) = 5996.64$$

$$(s_2)^2 * df_2 = SS_2$$
$$(15.68)^2 * (42-1) = 10080.36$$

And thus our pooled variance equals:

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{5996.64 + 10080.36}{39 + 41} = \frac{16077}{80} = 200.96$$

And our standard error equals:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{200.96}{40} + \frac{200.96}{42}} = \sqrt{5.02 + 4.78} = \sqrt{9.89} = 3.14$$

All of these steps are just slightly different ways of using the same formulae, numbers, and ideas we have worked with up to this point. Once we get out standard error, it's time to build our confidence interval.

$$95\% CI = 3.25 \pm 1.990(3.14)$$

$$\text{Upper Bound} = 3.25 + 1.990(3.14)$$
$$UB = 3.25 + 6.25$$
$$UB = 9.50$$

$$\text{Lower Bound } = 3.25 - 1.990(3.14)$$
$$LB = 3.25 - 6.25$$
$$LB = -3.00$$

$$95\% CI = (-3.00, 9.50)$$

Our confidence interval, as always, represents a range of values that would be considered reasonable or plausible based on our observed data. In this instance, our interval (-3.00, 9.50) does contain zero. Thus, even though the means look a little bit different, it may very well be the case that the life satisfaction in both of these towns is the same. Proving otherwise would require more data.

This page titled 11.7: Effect Sizes and Confidence Intervals is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **10.7: Effect Sizes and Confidence Intervals** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

11.8: Homogeneity of Variance

Before wrapping up the coverage of independent samples t-tests, there is one other important topic to cover. Using the pooled variance to calculate the test statistic relies on an assumption known as homogeneity of variance. In statistics, an assumption is some characteristic that we assume is true about our data, and our ability to use our inferential statistics accurately and correctly relies on these assumptions being true. If these assumptions are not true, then our analyses are at best ineffective (e.g. low power to detect effects) and at worst inappropriate (e.g. too many Type I errors). A detailed coverage of assumptions is beyond the scope of this course, but it is important to know that they exist for all analyses.

For the current analysis, one important assumption is homogeneity of variance. This is fancy statistical talk for the idea that the true population variance for each group is the same and any difference in the observed sample variances is due to random chance (if this sounds eerily similar to the idea of testing the null hypothesis that the true population means are equal, that's because it is exactly the same!) This notion allows us to compute a single pooled variance that uses our easily calculated degrees of freedom. If the assumption is shown to not be true, then we have to use a very complicated formula to estimate the proper degrees of freedom. There are formal tests to assess whether or not this assumption is met, but we will not discuss them here.

Many statistical programs incorporate the test of homogeneity of variance automatically and can report the results of the analysis assuming it is true or assuming it has been violated. You can easily tell which is which by the degrees of freedom: the corrected degrees of freedom (which is used when the assumption of homogeneity of variance is violated) will have decimal places. Fortunately, the independent samples t-test is very robust to violations of this assumption (an analysis is "robust" if it works well even when its assumptions are not met), which is why we do not bother going through the tedious work of testing and estimating new degrees of freedom by hand.

This page titled 11.8: Homogeneity of Variance is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **10.8: Homogeneity of Variance** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

11.E: Independent Samples (Exercises)

1. What is meant by “the difference of the means” when talking about an independent samples t-test? How does it differ from the “mean of the differences” in a repeated measures t-test?

Answer:

The difference of the means is one mean, calculated from a set of scores, compared to another mean which is calculated from a different set of scores; the independent samples t-test looks for whether the two separate values are different from one another. This is different than the “mean of the differences” because the latter is a single mean computed on a single set of difference scores that come from one data collection of matched pairs. So, the difference of the means deals with two numbers but the mean of the differences is only one number.

2. Describe three research questions that could be tested using an independent samples t-test.
3. Calculate pooled variance from the following raw data:

Group 1	Group 2
16	4
11	10
9	15
7	13
5	12
4	9
12	8

Answer:

$\mathrm{SS}_1 = 106.86, \mathrm{SS}_2 = 78.86, s_p^2 = 15.48$

4. Calculate the standard error from the following descriptive statistics
   a. s_1 = 24, s_2 = 21, n_1 = 36, n_2 = 49
   b. s_1 = 15.40, s_2 = 14.80, n_1 = 20, n_2 = 23
   c. s_1 = 12, s_2 = 10, n_1 = 25, n_2 = 25
5. Determine whether to reject or fail to reject the null hypothesis in the following situations:
   a. $t(40) = 2.49$, $\alpha = 0.01$, one-tailed test to the right
   b. $\overline{X_1} = 64, \overline{X_2} = 54, n_1 = 14, n_2 = 12, s_{\overline{X_1}-\overline{X_2}} = 9.75, \alpha = 0.05$, two-tailed test
   c. 95% Confidence Interval: (0.50, 2.10)

Answer:

a. Reject
b. Fail to Reject
c. Reject

6. A professor is interest in whether or not the type of software program used in a statistics lab affects how well students learn the material. The professor teaches the same lecture material to two classes but has one class use a point-and-click software program in lab and has the other class use a basic programming language. The professor tests for a difference between the two classes on their final exam scores.

Point-and-Click	Programming
83	86
83	79
63	100
77	74
86	70
84	67
78	83
61	85
65	74
75	86
100	87
60	61
90	76
66	100
54	

7. A researcher wants to know if there is a difference in how busy someone is based on whether that person identifies as an early bird or a night owl. The researcher gathers data from people in each group, coding the data so that higher scores represent higher levels of being busy, and tests for a difference between the two at the .05 level of significance.

Early Bird	Night Owl
23	26
28	10
27	20
33	19
26	26
30	18
22	12
25	25
26	

Answer:

Step 1: $H_0 : \mu_1 - \mu_2 = 0$ "There is not difference in the average business of early birds versus night owls", $H_A : \mu_1 - \mu_2 \neq 0$ "There is a difference in the average business of early birds versus night owls."

Step 2: Two-tailed test, df = 15, $t*$ = 2.131.

Step 3: $\overline{X_1} = 26.67, \overline{X_2} = 19.50, s_p^2 = 23.73, s_{X_1} - \overline{X_2} = 2.37$

Step 4: $t > t*$, Reject H_0. Based on our data of early birds and night owls, we can conclude that early birds are busier ($\overline{X_1} = 26.67$) than night owls ($\overline{X_2} = 19.50$), $t(15) = 3.03$, $p < .05$. Since the result is significant, we need an effect size: Cohen's $d = 1.47$, which is a large effect.

8. Lots of people claim that having a pet helps lower their stress level. Use the following summary data to test the claim that there is a lower average stress level among pet owners (group 1) than among non-owners (group 2) at the .05 level of significance.

$$\overline{X_1} = 16.25, \overline{X_2} = 20.95, s_1 = 4.00, s_2 = 5.10, n_1 = 29, n_2 = 25$$

9. Administrators at a university want to know if students in different majors are more or less extroverted than others. They provide you with descriptive statistics they have for English majors (coded as 1) and History majors (coded as 2) and ask you to create a confidence interval of the difference between them. Does this confidence interval suggest that the students from the majors differ?

$$\overline{X_1} = 3.78, \overline{X_2} = 2.23, s_1 = 2.60, s_2 = 1.15, n_1 = 45, n_2 = 40$$

Answer:

$\overline{X_1} - \overline{X_2} = 1.55, \mathrm{t}^* = 1.990, s_{\overline{X_1} - \overline{X_2}} = 0.45$, CI = (0.66, 2.44). This confidence interval does not contain zero, so it does suggest that there is a difference between the extroversion of English majors and History majors.

10. Researchers want to know if people's awareness of environmental issues varies as a function of where they live. The researchers have the following summary data from two states, Alaska and Hawaii, that they want to use to test for a difference.

$$\overline{\mathrm{X}_H} = 47.50, \overline{\mathrm{X}_A} = 45.70, \mathrm{s_H} = 14.65, \mathrm{s_A} = 13.20, \mathrm{n_H} = 139, \mathrm{n_A} = 150$$

This page titled 11.E: Independent Samples (Exercises) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **10.E: Independent Samples (Exercises)** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

12: Analysis of Variance

12.1: Observing and Interpreting Variability
12.2: Sources of Variance
12.3: ANOVA Table
12.4: ANOVA and Type I Error
12.5: Hypotheses in ANOVA
12.6: Scores on Job Application Tests
12.7: Variance Explained
12.8: Post Hoc Tests
12.9: Other ANOVA Designs
12.10: Analysis of Variance (Exercises)

This page titled 12: Analysis of Variance is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

12.1: Observing and Interpreting Variability

We have seen time and again that scores, be they individual data or group means, will differ naturally. Sometimes this is due to random chance, and other times it is due to actual differences. Our job as scientists, researchers, and data analysts is to determine if the observed differences are systematic and meaningful (via a hypothesis test) and, if so, what is causing those differences. Through this, it becomes clear that, although we are usually interested in the mean or average score, it is the variability in the scores that is key.

Take a look at Figure 12.1.1, which shows scores for many people on a test of skill used as part of a job application. The x-axis has each individual person, in no particular order, and the y-axis contains the score each person received on the test. As we can see, the job applicants differed quite a bit in their performance, and understanding why that is the case would be extremely useful information. However, there's no interpretable pattern in the data, especially because we only have information on the test, not on any other variable (remember that the x-axis here only shows individual people and is not ordered or interpretable).

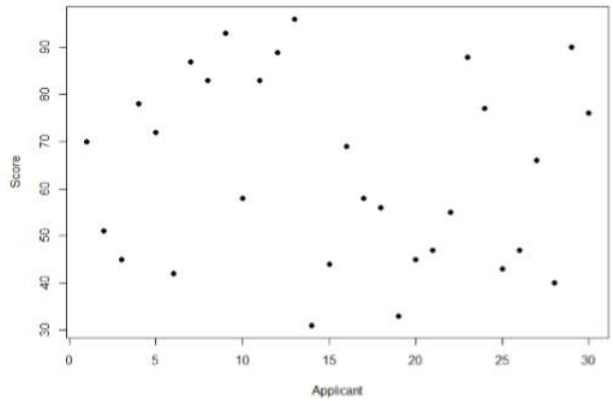


Figure 12.1.1: Scores on a job test

Our goal is to explain this variability that we are seeing in the dataset. Let's assume that as part of the job application procedure we also collected data on the highest degree each applicant earned. With knowledge of what the job requires, we could sort our applicants into three groups: those applicants who have a college degree related to the job, those applicants who have a college degree that is not related to the job, and those applicants who did not earn a college degree. This is a common way that job applicants are sorted, and we can use ANOVA to test if these groups are actually different. Figure 12.1.2 presents the same job applicant scores, but now they are color coded by group membership (i.e. which group they belong in). Now that we can differentiate between applicants this way, a pattern starts to emerge: those applicants with a relevant degree (coded red) tend to be near the top, those applicants with no college degree (coded black) tend to be near the bottom, and the applicants with an unrelated degree (coded green) tend to fall into the middle. However, even within these groups, there is still some variability, as shown in Figure 12.1.2.

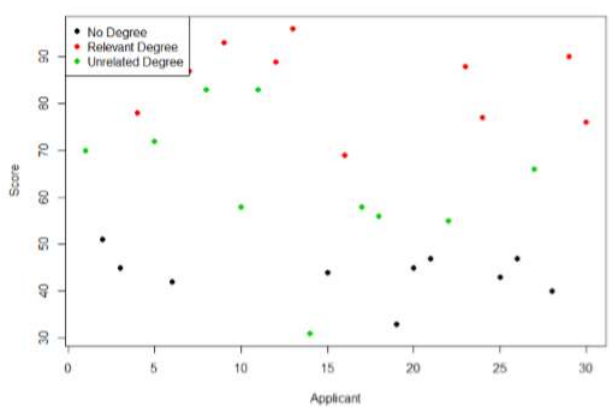


Figure 12.1.2: Applicant scores coded by degree earned

This pattern is even easier to see when the applicants are sorted and organized into their respective groups, as shown in Figure 12.1.3.

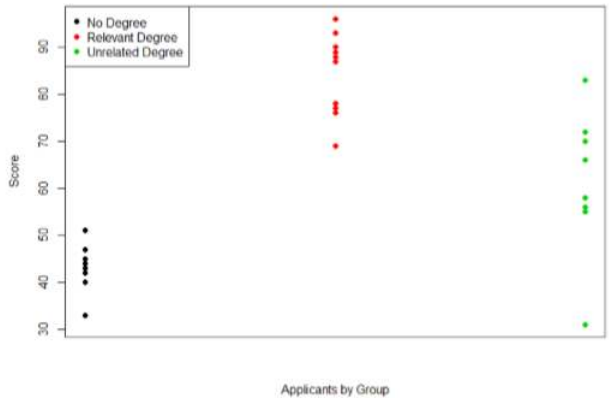


Figure 12.1.3: Applicant scores by group

Now that we have our data visualized into an easily interpretable format, we can clearly see that our applicants' scores differ largely along group lines. Those applicants who do not have a college degree received the lowest scores, those who had a degree relevant to the job received the highest scores, and those who did have a degree but one that is not related to the job tended to fall somewhere in the middle. Thus, we have systematic variance between our groups.

We can also clearly see that within each group, our applicants' scores differed from one another. Those applicants without a degree tended to score very similarly, since the scores are clustered close together. Our group of applicants with relevant degrees varied a little but more than that, and our group of applicants with unrelated degrees varied quite a bit. It may be that there are other factors that cause the observed score differences within each group, or they could just be due to random chance. Because we do not have any other explanatory data in our dataset, the variability we observe within our groups is considered random error, with any deviations between a person and that person's group mean caused only by chance. Thus, we have unsystematic (random) variance within our groups.

The process and analyses used in ANOVA will take these two sources of variance (systematic variance between groups and random error within groups, or how much groups differ from each other and how much people differ within each group) and compare them to one another to determine if the groups have any explanatory value in our outcome variable. By doing this, we will test for statistically significant differences between the group means, just like we did for ***t***-tests. We will go step by step to break down the math to see how ANOVA actually works.

This page titled 12.1: Observing and Interpreting Variability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **11.1: Observing and Interpreting Variability** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

12.2: Sources of Variance

ANOVA is all about looking at the different sources of variance (i.e. the reasons that scores differ from one another) in a dataset. Fortunately, the way we calculate these sources of variance takes a very familiar form: the Sum of Squares. Before we get into the calculations themselves, we must first lay out some important terminology and notation.

In ANOVA, we are working with two variables, a grouping or explanatory variable and a continuous outcome variable. The grouping variable is our predictor (it predicts or explains the values in the outcome variable) or, in experimental terms, our independent variable, and it made up of k groups, with k being any whole number 2 or greater. That is, ANOVA requires two or more groups to work, and it is usually conducted with three or more. In ANOVA, we refer to groups as "levels", so the number of levels is just the number of groups, which again is k. In the above example, our grouping variable was education, which had 3 levels, so k = 3. When we report any descriptive value (e.g. mean, sample size, standard deviation) for a specific group, we will use a subscript 1…k to denote which group it refers to. For example, if we have three groups and want to report the standard deviation s for each group, we would report them as s_1, s_2, and s_3.

Our second variable is our outcome variable. This is the variable on which people differ, and we are trying to explain or account for those differences based on group membership. In the example above, our outcome was the score each person earned on the test. Our outcome variable will still use X for scores as before. When describing the outcome variable using means, we will use subscripts to refer to specific group means. So if we have k = 3 groups, our means will be $\overline{X_1}$, $\overline{X_2}$, and $\overline{X_3}$. We will also have a single mean representing the average of all participants across all groups. This is known as the grand mean, and we use the symbol $\overline{X_G}$. These different means – the individual group means and the overall grand mean – will be how we calculate our sums of squares.

Finally, we now have to differentiate between several different sample sizes. Our data will now have sample sizes for each group, and we will denote these with a lower case "n" and a subscript, just like with our other descriptive statistics: n_1, n_2, and n_3. We also have the overall sample size in our dataset, and we will denote this with a capital N. The total sample size is just the group sample sizes added together.

Between Groups Sum of Squares

One source of variability we can identify in 11.1.3 of the above example was differences or variability between the groups. That is, the groups clearly had different average levels. The variability arising from these differences is known as the between groups variability, and it is quantified using Between Groups Sum of Squares.

Our calculations for sums of squares in ANOVA will take on the same form as it did for regular calculations of variance. Each observation, in this case the group means, is compared to the overall mean, in this case the grand mean, to calculate a deviation score. These deviation scores are squared so that they do not cancel each other out and sum to zero. The squared deviations are then added up, or summed. There is, however, one small difference. Because each group mean represents a group composed of multiple people, before we sum the deviation scores we must multiple them by the number of people within that group. Incorporating this, we find our equation for Between Groups Sum of Squares to be:

$$SS_B = \sum n_j \left(\overline{X}_J - \overline{X_G}\right)^2 \tag{12.2.1}$$

The subscript j refers to the "j^{th}" group where j = 1…k to keep track of which group mean and sample size we are working with. As you can see, the only difference between this equation and the familiar sum of squares for variance is that we are adding in the sample size. Everything else logically fits together in the same way.

Within Groups Sum of Squares

The other source of variability in the figures comes from differences that occur within each group. That is, each individual deviates a little bit from their respective group mean, just like the group means differed from the grand mean. We therefore label this source the Within Groups Sum of Squares. Because we are trying to account for variance based on group-level means, any deviation from the group means indicates an inaccuracy or error. Thus, our within groups variability represents our error in ANOVA.

The formula for this sum of squares is again going to take on the same form and logic. What we are looking for is the distance between each individual person and the mean of the group to which they belong. We calculate this deviation score, square it so that they can be added together, then sum all of them into one overall value:

$$SS_W = \sum \left(X_{ij} - \overline{X}_j\right)^2 \tag{12.2.2}$$

In this instance, because we are calculating this deviation score for each individual person, there is no need to multiple by how many people we have. The subscript j again represents a group and the subscript i refers to a specific person. So, X_{ij} is read as "the ith person of the jth group." It is important to remember that the deviation score for each person is only calculated relative to their group mean: do not calculate these scores relative to the other group means.

Total Sum of Squares

The Between Groups and Within Groups Sums of Squares represent all variability in our dataset. We also refer to the total variability as the Total Sum of Squares, representing the overall variability with a single number. The calculation for this score is exactly the same as it would be if we were calculating the overall variance in the dataset (because that's what we are interested in explaining) without worrying about or even knowing about the groups into which our scores fall:

$$SS_T = \sum \left(X_i - \overline{X_G}\right)^2 \tag{12.2.3}$$

We can see that our Total Sum of Squares is just each individual score minus the grand mean. As with our Within Groups Sum of Squares, we are calculating a deviation score for each individual person, so we do not need to multiply anything by the sample size; that is only done for Between Groups Sum of Squares.

An important feature of the sums of squares in ANOVA is that they all fit together. We could work through the algebra to demonstrate that if we added together the formulas for SS_B and SS_W, we would end up with the formula for SS_T. That is:

$$SS_T = SS_B + SS_W \tag{12.2.4}$$

This will prove to be very convenient, because if we know the values of any two of our sums of squares, it is very quick and easy to find the value of the third. It is also a good way to check calculations: if you calculate each SS by hand, you can make sure that they all fit together as shown above, and if not, you know that you made a math mistake somewhere.

We can see from the above formulas that calculating an ANOVA by hand from raw data can take a very, very long time. For this reason, you will not be required to calculate the SS values by hand, but you should still take the time to understand how they fit together and what each one represents to ensure you understand the analysis itself.

This page titled 12.2: Sources of Variance is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **11.2: Sources of Variance** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

12.3: ANOVA Table

All of our sources of variability fit together in meaningful, interpretable ways as we saw above, and the easiest way to do this is to organize them into a table. The ANOVA table, shown in Table 12.3.1, is how we calculate our test statistic.

Table 12.3.1: ANOVA Table

Source	SS	df	MS	F
Between	SS_B	$k-1$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_W}$
Within	SS_W	$N-k$	$\frac{SS_W}{df_W}$	
Total	SS_T	$N-1$		

The first column of the ANOVA table, labeled "Source", indicates which of our sources of variability we are using: between groups, within groups, or total. The second column, labeled "SS", contains our values for the sums of squares that we learned to calculate above. As noted previously, calculating these by hand takes too long, and so the formulas are not presented in Table 12.3.1. However, remember that the Total is the sum of the other two, in case you are only given two SS values and need to calculate the third.

The next column, labeled "df", is our degrees of freedom. As with the sums of squares, there is a different df for each group, and the formulas are presented in the table. Notice that the total degrees of freedom, N–1, is the same as it was for our regular variance. This matches the SS_T formulation to again indicate that we are simply taking our familiar variance term and breaking it up into difference sources. Also remember that the capital N in the df calculations refers to the overall sample size, not a specific group sample size. Notice that the total row for degrees of freedom, just like for sums of squares, is just the Between and Within rows added together. If you take N–k + k–1, then the "–k" and "+k" portions will cancel out, and you are left with N–1. This is a convenient way to quickly check your calculations.

The third column, labeled "MS", is our Mean Squares for each source of variance. A "mean square" is just another way to say variability. Each mean square is calculated by dividing the sum of squares by its corresponding degrees of freedom. Notice that we do this for the Between row and the Within row, but not for the Total row. There are two reasons for this. First, our Total Mean Square would just be the variance in the full dataset (put together the formulas to see this for yourself), so it would not be new information. Second, the Mean Square values for Between and Within would not add up to equal the Mean Square Total because they are divided by different denominators. This is in contrast to the first two columns, where the Total row was both the conceptual total (i.e. the overall variance and degrees of freedom) and the literal total of the other two rows.

The final column in the ANOVA table, labeled "F", is our test statistic for ANOVA. The F statistic, just like a t- or z-statistic, is compared to a critical value to see whether we can reject for fail to reject a null hypothesis. Thus, although the calculations look different for ANOVA, we are still doing the same thing that we did in all of Unit 2. We are simply using a new type of data to test our hypotheses. We will see what these hypotheses look like shortly, but first, we must take a moment to address why we are doing our calculations this way.

This page titled 12.3: ANOVA Table is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **11.3: ANOVA Table** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

12.4: ANOVA and Type I Error

You may be wondering why we do not just use another t-test to test our hypotheses about three or more groups the way we did in Unit 2. After all, we are still just looking at group mean differences. The reason is that our t-statistic formula can only handle up to two groups, one minus the other. With only two groups, we can move our population parameters for the group means around in our null hypothesis and still get the same interpretation: the means are equal, which can also be concluded if one mean minus the other mean is equal to zero. However, if we tried adding a third mean, we would no longer be able to do this. So, in order to use t-tests to compare three or more means, we would have to run a series of individual group comparisons.

For only three groups, we would have three t-tests: group 1 vs group 2, group 1 vs group 3, and group 2 vs group 3. This may not sound like a lot, especially with the advances in technology that have made running an analysis very fast, but it quickly scales up. With just one additional group, bringing our total to four, we would have six comparisons: group 1 vs group 2, group 1 vs group 3, group 1 vs group 4, group 2 vs group 3, group 2 vs group 4, and group 3 vs group 4. This makes for a logistical and computation nightmare for five or more groups.

A bigger issue, however, is our probability of committing a Type I Error. Remember that a Type I error is a false positive, and the chance of committing a Type I error is equal to our significance level, α. This is true if we are only running a single analysis (such as a t-test with only two groups) on a single dataset. However, when we start running multiple analyses on the same dataset, our Type I error rate increases, raising the probability that we are capitalizing on random chance and rejecting a null hypothesis when we should not. ANOVA, by comparing all groups simultaneously with a single analysis, averts this issue and keeps our error rate at the α we set.

This page titled 12.4: ANOVA and Type I Error is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **11.4: ANOVA and Type I Error** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

12.5: Hypotheses in ANOVA

So far we have seen what ANOVA is used for, why we use it, and how we use it. Now we can turn to the formal hypotheses we will be testing. As with before, we have a null and an alternative hypothesis to lay out. Our null hypothesis is still the idea of "no difference" in our data. Because we have multiple group means, we simply list them out as equal to each other:

$$\mathrm{H}_0: \text{ There is no difference in the group means}$$
$$\mathrm{H}_0: \mu_1 = \mu_2 = \mu_3$$

We list as many μ parameters as groups we have. In the example above, we have three groups to test, so we have three parameters in our null hypothesis. If we had more groups, say, four, we would simply add another μ to the list and give it the appropriate subscript, giving us:

$$\mathrm{H}_0: \text{ There is no difference in the group means}$$
$$\mathrm{H}_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Notice that we do not say that the means are all equal to zero, we only say that they are equal to one another; it does not matter what the actual value is, so long as it holds for all groups equally.

Our alternative hypothesis for ANOVA is a little bit different. Let's take a look at it and then dive deeper into what it means:

$$\mathrm{H}_A: \text{ At least one mean is different}$$

The first difference in obvious: there is no mathematical statement of the alternative hypothesis in ANOVA. This is due to the second difference: we are not saying which group is going to be different, only that at least one will be. Because we do not hypothesize about which mean will be different, there is no way to write it mathematically. Related to this, we do not have directional hypotheses (greater than or less than) like we did in Unit 2. Due to this, our alternative hypothesis is always exactly the same: at least one mean is different.

In Unit 2, we saw that, if we reject the null hypothesis, we can adopt the alternative, and this made it easy to understand what the differences looked like. In ANOVA, we will still adopt the alternative hypothesis as the best explanation of our data if we reject the null hypothesis. However, when we look at the alternative hypothesis, we can see that it does not give us much information. We will know that a difference exists somewhere, but we will not know where that difference is. Is only group 1 different but groups 2 and 3 the same? Is it only group 2? Are all three of them different? Based on just our alternative hypothesis, there is no way to be sure. We will come back to this issue later and see how to find out specific differences. For now, just remember that we are testing for any difference in group means, and it does not matter where that difference occurs.

Now that we have our hypotheses for ANOVA, let's work through an example. We will continue to use the data from Figures 11.1.1 through 11.1.3 for continuity.

This page titled 12.5: Hypotheses in ANOVA is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **11.5: Hypotheses in ANOVA** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

12.6: Scores on Job Application Tests

Our data come from three groups of 10 people each, all of whom applied for a single job opening: those with no college degree, those with a college degree that is not related to the job opening, and those with a college degree from a relevant field. We want to know if we can use this group membership to account for our observed variability and, by doing so, test if there is a difference between our three group means. We will start, as always, with our hypotheses.

Step 1: State the Hypotheses

Our hypotheses are concerned with the means of groups based on education level, so:

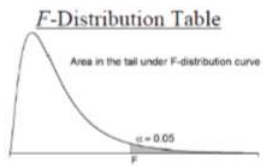
$$\mathrm{H}_0: \text{ There is no difference between the means of the education groups}$$
$$\mathrm{H}_0: \mu_1 = \mu_2 = \mu_3$$

$$\mathrm{H}_A: \text{ At least one mean is different}$$

Again, we phrase our null hypothesis in terms of what we are actually looking for, and we use a number of population parameters equal to our number of groups. Our alternative hypothesis is always exactly the same.

Step 2: Find the Critical Values

Our test statistic for ANOVA, as we saw above, is F. Because we are using a new test statistic, we will get a new table: the F distribution table, the top of which is shown in Figure 12.6.1:



df denom.	Degrees of Freedom: Numerator 1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.97	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.10	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.97	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.56	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.33	3.47	3.07	2.84	2.69	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.38	2.32	2.28
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.26
25	4.24	3.39	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20

Figure 12.6.1: F distribution table.

The F table only displays critical values for α = 0.05. This is because other significance levels are uncommon and so it is not worth it to use up the space to present them. There are now two degrees of freedom we must use to find our critical value: Numerator and Denominator. These correspond to the numerator and denominator of our test statistic, which, if you look at the ANOVA table presented earlier, are our Between Groups and Within Groups rows, respectively. The df_B is the "Degrees of Freedom: Numerator" because it is the degrees of freedom value used to calculate the Mean Square Between, which in turn was the numerator of our F statistic. Likewise, the df_W is the "df denom." (short for denominator) because it is the degrees of freedom value used to calculate the Mean Square Within, which was our denominator for F.

The formula for df_B is k–1, and remember that k is the number of groups we are assessing. In this example, $k = 3$ so our df_B = 2. This tells us that we will use the second column, the one labeled 2, to find our critical value. To find the proper row, we simply

calculate the df_W, which was $N-k$. The original prompt told us that we have "three groups of 10 people each," so our total sample size is 30. This makes our value for df_W = 27. If we follow the second column down to the row for 27, we find that our critical value is 3.35. We use this critical value the same way as we did before: it is our criterion against which we will compare our obtained test statistic to determine statistical significance.

Step 3: Calculate the Test Statistic

Now that we have our hypotheses and the criterion we will use to test them, we can calculate our test statistic. To do this, we will fill in the ANOVA table. When we do so, we will work our way from left to right, filling in each cell to get our final answer. We will assume that we are given the SS values as shown below:

Table 12.6.1: ANOVA Table

Source	SS	df	MS	F
Between	8246			
Within	3020			
Total				

These may seem like random numbers, but remember that they are based on the distances between the groups themselves and within each group. Figure 12.6.2 shows the plot of the data with the group means and grand mean included. If we wanted to, we could use this information, combined with our earlier information that each group has 10 people, to calculate the Between Groups Sum of Squares by hand. However, doing so would take some time, and without the specific values of the data points, we would not be able to calculate our Within Groups Sum of Squares, so we will trust that these values are the correct ones.

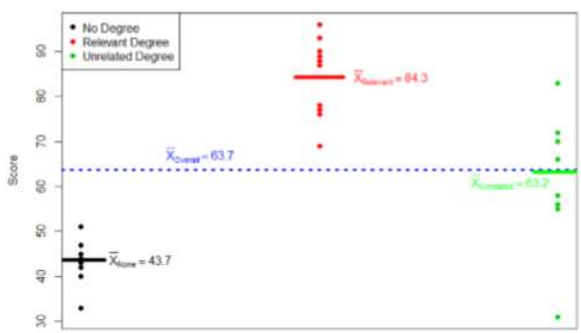


Figure 12.6.2: Means

We were given the sums of squares values for our first two rows, so we can use those to calculate the Total Sum of Squares.

Table 12.6.2: Total Sum of Squares

Source	SS	df	MS	F
Between	8246			
Within	3020			
Total	11266			

We also calculated our degrees of freedom earlier, so we can fill in those values. Additionally, we know that the total degrees of freedom is $N-1$, which is 29. This value of 29 is also the sum of the other two degrees of freedom, so everything checks out.

Table 12.6.3: Total Sum of Squares

Source	SS	df	MS	F
Between	8246	2		
Within	3020	27		
Total	11266	29		

Now we have everything we need to calculate our mean squares. Our MS values for each row are just the SS divided by the df for that row, giving us:

Table 12.6.4: Total Sum of Squares

Source	SS	df	MS	F
Between	8246	2	4123	
Within	3020	27	111.85	
Total	11266	29		

Remember that we do not calculate a Total Mean Square, so we leave that cell blank. Finally, we have the information we need to calculate our test statistic. F is our MS_B divided by MS_W.

Table 12.6.5: Total Sum of Squares

Source	SS	df	MS	F
Between	8246	2	4123	36.86
Within	3020	27	111.85	
Total	11266	29		

So, working our way through the table given only two SS values and the sample size and group size given before, we calculate our test statistic to be $F_{obt} = 36.86$, which we will compare to the critical value in step 4.

Step 4: Make the Decision

Our obtained test statistic was calculated to be $F_{obt} = 36.86$ and our critical value was found to be $F^* = 3.35$. Our obtained statistic is larger than our critical value, so we can reject the null hypothesis.

Reject H_0. Based on our 3 groups of 10 people, we can conclude that job test scores are statistically significantly different based on education level, $F(2, 27) = 36.86, p < .05$.

Notice that when we report F, we include both degrees of freedom. We always report the numerator then the denominator, separated by a comma. We must also note that, because we were only testing for any difference, we cannot yet conclude which groups are different from the others. We will do so shortly, but first, because we found a statistically significant result, we need to calculate an effect size to see how big of an effect we found.

This page titled 12.6: Scores on Job Application Tests is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **11.6: Scores on Job Application Tests** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

12.7: Variance Explained

Recall that the purpose of ANOVA is to take observed variability and see if we can explain those differences based on group membership. To that end, our effect size will be just that: the variance explained. You can think of variance explained as the proportion or percent of the differences we are able to account for based on our groups. We know that the overall observed differences are quantified as the Total Sum of Squares, and that our observed effect of group membership is the Between Groups Sum of Squares. Our effect size, therefore, is the ratio of these to sums of squares. Specifically:

$$\eta^2 = \frac{SS_B}{SS_T} \tag{12.7.1}$$

The effect size η^2 is called "eta-squared" and represents variance explained. For our example, our values give an effect size of:

$$\eta^2 = \frac{8246}{11266} = 0.73$$

So, we are able to explain 73% of the variance in job test scores based on education. This is, in fact, a huge effect size, and most of the time we will not explain nearly that much variance. Our guidelines for the size of our effects are:

Table 12.7.1: Guidelines for the size of our effects

η^2	Size
0.01	Small
0.09	Medium
0.25	Large

So, we found that not only do we have a statistically significant result, but that our observed effect was very large! However, we still do not know specifically which groups are different from each other. It could be that they are all different, or that only those who have a relevant degree are different from the others, or that only those who have no degree are different from the others. To find out which is true, we need to do a special analysis called a post hoc test.

This page titled 12.7: Variance Explained is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **11.7: Variance Explained** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

12.8: Post Hoc Tests

A post hoc test is used only after we find a statistically significant result and need to determine where our differences truly came from. The term "post hoc" comes from the Latin for "after the event". There are many different post hoc tests that have been developed, and most of them will give us similar answers. We will only focus here on the most commonly used ones. We will also only discuss the concepts behind each and will not worry about calculations.

Bonferroni Test

A Bonferroni test is perhaps the simplest post hoc analysis. A Bonferroni test is a series of t-tests performed on each pair of groups. As we discussed earlier, the number of groups quickly grows the number of comparisons, which inflates Type I error rates. To avoid this, a Bonferroni test divides our significance level α by the number of comparisons we are making so that when they are all run, they sum back up to our original Type I error rate. Once we have our new significance level, we simply run independent samples t-tests to look for difference between our pairs of groups. This adjustment is sometimes called a Bonferroni Correction, and it is easy to do by hand if we want to compare obtained p-values to our new corrected α level, but it is more difficult to do when using critical values like we do for our analyses so we will leave our discussion of it to that.

Tukey's Honest Significant Difference

Tukey's Honest Significant Difference (HSD) is a very popular post hoc analysis. This analysis, like Bonferroni's, makes adjustments based on the number of comparisons, but it makes adjustments to the test statistic when running the comparisons of two groups. These comparisons give us an estimate of the difference between the groups and a confidence interval for the estimate. We use this confidence interval in the same way that we use a confidence interval for a regular independent samples t-test: if it contains 0.00, the groups are not different, but if it does not contain 0.00 then the groups are different.

Below are the differences between the group means and the Tukey's HSD confidence intervals for the differences:

Table 12.8.1: Differences between the group means and the Tukey's HSD confidence intervals

Comparison	Difference	Tukey's HSD CI
None vs Relevant	40.60	(28.87, 52.33)
None vs Unrelated	19.50	(7.77, 31.23)
Relevant vs Unrelated	21.10	(9.37, 32.83)

As we can see, none of these intervals contain 0.00, so we can conclude that all three groups are different from one another.

Scheffe's Test

Another common post hoc test is Scheffe's Test. Like Tukey's HSD, Scheffe's test adjusts the test statistic for how many comparisons are made, but it does so in a slightly different way. The result is a test that is "conservative," which means that it is less likely to commit a Type I Error, but this comes at the cost of less power to detect effects. We can see this by looking at the confidence intervals that Scheffe's test gives us:

Table 12.8.2: Confidence intervals given by Scheffe's test

Comparison	Difference	Tukey's HSD CI
None vs Relevant	40.60	(28.35, 52.85)
None vs Unrelated	19.50	(7.25, 31.75)
Relevant vs Unrelated	21.10	(8.85, 33.35)

As we can see, these are slightly wider than the intervals we got from Tukey's HSD. This means that, all other things being equal, they are more likely to contain zero. In our case, however, the results are the same, and we again conclude that all three groups

differ from one another.

There are many more post hoc tests than just these three, and they all approach the task in different ways, with some being more conservative and others being more powerful. In general, though, they will give highly similar answers. What is important here is to be able to interpret a post hoc analysis. If you are given post hoc analysis confidence intervals, like the ones seen above, read them the same way we read confidence intervals in chapter 10: if they contain zero, there is no difference; if they do not contain zero, there is a difference.

This page titled 12.8: Post Hoc Tests is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **11.8: Post Hoc Tests** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

12.9: Other ANOVA Designs

We have only just scratched the surface on ANOVA in this chapter. There are many other variations available for the one-way ANOVA presented here. There are also other types of ANOVAs that you are likely to encounter. The first is called a factorial ANOVA. Factorial ANOVAs use multiple grouping variables, not just one, to look for group mean differences. Just as there is no limit to the number of groups in a one-way ANOVA, there is no limit to the number of grouping variables in a Factorial ANOVA, but it becomes very difficult to find and interpret significant results with many factors, so usually they are limited to two or three grouping variables with only a small number of groups in each. Another ANOVA is called a Repeated Measures ANOVA. This is an extension of a repeated measures or matched pairs t-test, but in this case we are measuring each person three or more times to look for a change. We can even combine both of these advanced ANOVAs into mixed designs to test very specific and valuable questions. These topics are far beyond the scope of this text, but you should know about their existence. Our treatment of ANOVA here is a small first step into a much larger world!

This page titled 12.9: Other ANOVA Designs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **11.9: Other ANOVA Designs** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

12.10: Analysis of Variance (Exercises)

1. What are the three pieces of variance analyzed in ANOVA?

Answer:

Variance between groups (SSB), variance within groups (SSW) and total variance (SST).

2. What does rejecting the null hypothesis in ANOVA tell us? What does it not tell us?
3. What is the purpose of post hoc tests?

Answer:

Post hoc tests are run if we reject the null hypothesis in ANOVA; they tell us which specific group differences are significant.

4. Based on the ANOVA table below, do you reject or fail to reject the null hypothesis? What is the effect size?

Source	SS	df	MS	F
Between	60.72	3	20.24	3.88
Within	213.61	41	5.21	
Total	274.33	44		

5. Finish filling out the following ANOVA tables:

a. $K = 4$

Source	SS	df	MS	F
Between	87.40			
Within				
Total	199.22	33		

b. $N = 14$

Source	SS	df	MS	F
Between		2	14.10	
Within				
Total	64.65			

c.

Source	SS	df	MS	F
Between		2		42.36
Within		54	2.48	
Total				

Answer:

a. $K = 4$

Source	SS	df	MS	F
Between	87.40	3	29.13	7.81
Within	111.82	30	3.73	
Total	199.22	33		

b. $N = 14$

Source	SS	df	MS	F
Between	28.20	2	14.10	4.26
Within	36.45	11	3.31	
Total	64.65	13		

c.

Source	SS	df	MS	F
Between	210.10	2	105.05	42.36
Within	133.92	54	2.48	
Total	344.02			

6. You know that stores tend to charge different prices for similar or identical products, and you want to test whether or not these differences are, on average, statistically significantly different. You go online and collect data from 3 different stores, gathering information on 15 products at each store. You find that the average prices at each store are: Store 1 xbar = $27.82, Store 2 xbar = $38.96, and Store 3 xbar = $24.53. Based on the overall variability in the products and the variability within each store, you find the following values for the Sums of Squares: SST = 683.22, SSW = 441.19. Complete the ANOVA table and use the 4 step hypothesis testing procedure to see if there are systematic price differences between the stores.
7. You and your friend are debating which type of candy is the best. You find data on the average rating for hard candy (e.g. jolly ranchers, $\overline{X}$= 3.60), chewable candy (e.g. starburst, $\overline{X}$ = 4.20), and chocolate (e.g. snickers, $\overline{X}$= 4.40); each type of candy was rated by 30 people. Test for differences in average candy rating using SSB = 16.18 and SSW = 28.74.

Answer:

Step 1: $H_0 : \mu_1 = \mu_2 = \mu_3$ "There is no difference in average rating of candy quality", H_A: "At least one mean is different."

Step 2: 3 groups and 90 total observations yields $df_{num} = 2$ and $df_{den} = 87$, $\alpha = 0.05$, $F^* = 3.11$.

Step 3: based on the given SSB and SSW and the computed df from step 2, is:

Source	SS	df	MS	F
Between	16.18	2	8.09	24.52
Within	28.74	87	0.33	
Total	44.92	89		

Step 4: $F > F^*$, reject H_0. Based on the data in our 3 groups, we can say that there is a statistically significant difference in the quality of different types of candy, $F(2, 87) = 24.52, p < .05$. Since the result is significant, we need an effect size: $\eta^2 = 16.18/44.92 = .36$, which is a large effect

8. Administrators at a university want to know if students in different majors are more or less extroverted than others. They provide you with data they have for English majors ($\overline{X}$ = 3.78, n = 45), History majors ($\overline{X}$= 2.23, n = 40), Psychology majors ($\overline{X}$ = 4.41, n = 51), and Math majors ($\overline{X}$ = 1.15, n = 28). You find the $SSB = 75.80$ and $SSW = 47.40$ and test at $\alpha = 0.05$.
9. You are assigned to run a study comparing a new medication ($\overline{X}$= 17.47, n = 19), an existing medication ($\overline{X}$= 17.94, n = 18), and a placebo ($\overline{X}$= 13.70, n = 20), with higher scores reflecting better outcomes. Use $SSB = 210.10$ and $SSW = 133.90$ to test for differences.

Answer:

Step 1: $H_0 : \mu_1 = \mu_2 = \mu_3$ "There is no difference in average outcome based on treatment", H_A: "At least one mean is different."

Step 2: 3 groups and 57 total participants yields $df_{num} = 2$ and $df_{den} = 54$, $\alpha = 0.05, F^* = 3.18$.

Step 3: based on the given SSB and SSW and the computed df from step 2, is:

Source	SS	df	MS	F
Between	210.10	2	105.02	42.36
Within	133.90	54	2.48	
Total	344.00	56		

Step 4: $F > F^*$, reject H_0. Based on the data in our 3 groups, we can say that there is a statistically significant difference in the effectiveness of the treatments, $F(2, 54) = 42.36, p < .05$. Since the result is significant, we need an effect size: $\eta^2 = 210.10/344.00 = .61$, which is a large effect.

10. You are in charge of assessing different training methods for effectiveness. You have data on 4 methods: Method 1 ($\overline{X}$ = 87, n = 12), Method 2 ($\overline{X}$= 92, n = 14), Method 3 ($\overline{X}$ = 88, n = 15), and Method 4 ($\overline{X}$= 75, n = 11). Test for differences among these means, assuming $SSB = 64.81$ and $SST = 399.45$.

This page titled 12.10: Analysis of Variance (Exercises) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **11.1E: Analysis of Variance (Exercises)** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

13: Two-Factor ANOVAs (by Dr. Alisa Beyer)

13.1: Two-Factor ANOVAs
13.2: Conducting a Two Factor ANOVA
13.3: Graphing the Results of Factorial Experiments
13.E: Exercises

13: Two-Factor ANOVAs (by Dr. Alisa Beyer) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

13.1: Two-Factor ANOVAs

A single factor ANOVA is the statistical analysis appropriate when we are analyzing the results of an experiment in which we have one factor and are looking for differences in the response variable among three or more groups, each of which is receiving different levels or amounts of the factor. In chapter 12, we learned about the single factor ANOVA, also known as the one-way. We will now conceptually review a multi-factor ANOVA. We will keep it on the simpler side and use 2-factors (two independent/predictor variables) using a between-subjects design.

Logic of a 2 Factor ANOVA

A two factor ANOVA is used when we believe that more than one factor may affect a particular response (dependent) variable. For example, believe that the age of an adolescent will have an impact on number of phone calls made to the opposite sex and I also suspect that gender of the adolescent will have an impact on the number of phone calls made to the opposite sex.

To test my hypothesis that Age and Gender of adolescent will impact the number of phone calls made to the opposite sex in the past week. In this case, we have a between-subjects design for both age and gender. I have 2 conditions/levels/groups for each factor/variable. I will have to collect data for these for 4 samples of subjects:

	Gender	
Age	Teen Males	Teen Females
	Older Males	Older Females

Table 1. Example of 2×2 ANOVA

A 2×2 ANOVA gives you 4 conditions. *Note:* one way to identify the total conditions in a factorial study is to multiply the conditions for each factor. Thus, a 2×2 design is 2 times 2 giving us 4 total conditions for the study. We will discuss this more in a moment.

Remember that there are different types of ANOVAs based on design. In this case, we have a between-subjects design. An individual can only be in 1 condition for gender and 1 condition for age. So among the 4 total conditions/levels/groups between the 2 factors, an individual is

only in 1 of the samples. For a between-subjects design, there are 4 different samples. Two Factor ANOVA data is commonly organized like the table above and is referred to a **matrix**. When the data is organized in a matrix it is very easy to see the factors, as well as the separate levels of the factors.

- **Factorial designs** like the 2-Factor ANOVA allow a researcher to examine more than one independent variable on the dependent variable
  - Individually for each factor, reporting out a F for each
  - Collectively where the collective influence of the factors is referred to as an **interaction.** An interaction is the result of the two independent variables combining to produce a result that is different from a result that is produced by either variable alone.

- A 2-Factor ANOVA allows a researcher to assess the main effects (the independent variables) and the interaction yielding three outcomes (3 Fs), a F for factor 1, a F for factor 2 and an interaction between factor 1 and 2.

Let's go back to our example:

- Main Effect of Factor A
  - Is there a significant effect of age of teen (Factor A) on number of phone calls made to the opposite sex (response variable).

- Main Effect of Factor B
  - Is there a significant effect of sex of the teen (Factor B) on number of phone calls made to the opposite sex (response variable).

- Interaction of AxB
  - Does the effect of age of teen (Factor A) on the number of phone calls made to the opposite sex (response variable) depend on the sex of the teen (Factor B)?

13.1: Two-Factor ANOVAs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

13.2: Conducting a Two Factor ANOVA

Conducting a Two Factor ANOVA

Before we begin the process of calculating a 2-Factor ANOVA we need to review several key elements of the study:

- **Factors:** the independent variables/predictors
- **Levels of each factor:** how many conditions/groups/treatments a factor has
- **Response variable:** this is the dependent variable/outcome variable/measurement taken
- **Total number of condition in the experiment:** this is identified by multiplying out the number of levels for each factor
- **Number of subjects per condition, n:** how many participants are in each level/group/treatment
- **Total number of experiment participants, N:** this will be determined by type of factor for each. In a between-group design, there will be four different conditions of participants. In a complete repeated measures design, all participants are in all conditions. In mixed design, it will vary by the study design for each factor. For this chapter, we are focused on a between subjects design.

Remember that in experiments that are designed to test for a cause and effect relationship between two variables (experimental designs) the factor is the variable hypothesized to cause something to happen. The **response variable** is the variable we believe will be affected (changed) by the factor.

Level of each factor refers to the categories of a factor represented in the experiment. In our example of age and gender the number of levels was 2 x 2 – we refer to design by its *levels* (can also call them conditions/groups/treatments).

	Gender	
Age	Teen Males	Teen Females
	Older Males	Older Females

Our example from Table 1 was a 2 x 2 design because there were two levels of the age variable (i.e., younger and older) and two levels of gender (i.e., male and female).

Total number of groups in the experiment equals the number of levels in Factor A multiplied by the number of levels for Factor B. For our example, there were four conditions. Another way to think about the number of groups or conditions is the number of cells in the matrix.

In a factorial design like the 2-Factor ANOVA, the number of subjects per condition is denoted by *n* and the total number of experiment participants is denoted by *N*. For example, if each condition has 10 participants, then *n* = 10. The experiment would have *N* = 40. In other words, 4 conditions with 10 participations (*n* = 10) (4 x 10) = 40 participants in the study.

Hypothesis Testing

We use the same steps for 2- Factor ANOVA that we have used for all other test statistics.

Write the alternative and null hypotheses

- 3 separate set of hypotheses: one set for each F
  - A effect (factor 1)
  - B effect (factor 2)
  - Interaction (A x B or factor 1 x 2)

These are three separate ANOVA tests yielding 3 Fs that are independent and the results are unrelated to the outcome for either of the other two. The hypotheses are set up in the same way as chapter 12. We will see an example for an interaction later in the chapter.

Set criteria for decision making

There are three hypotheses and three F scores so there will be three critical boundaries. The critical boundary of *F* comes from the F distribution table.

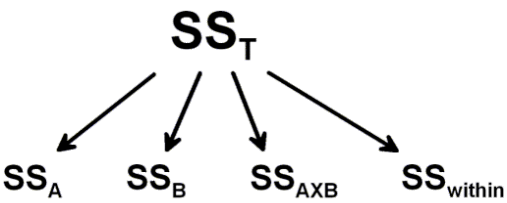
We need to know:

  - Alpha (α)
  - degrees of freedom Factor A = $df_{A =} (k_A -1)$ where k_A is number of levels
  - degrees of freedom Factor B = $df_B = (k_B -1)$ where k_B is number of levels
  - degrees of freedom Interaction (A x B) = $df_{A*B} = k_A * k_B$

- degrees of freedom for within treatment = $df_{total} - (df_A + df_B + df_{A*B})$ [within treatment is also called error]
- degrees of freedom total = $df_{total} = N - 1$ where N is the total number of scores

Note: We would still use the critical value ANOVA table for the critical F-values. The critical values may not be the same for each hypothesis; it will depend on the number of rows and columns used in the study! We will see this in an example later in the chapter.

Sample data are collected and analyzed by performing statistics (calculations for our adjusted step 3)



In the first stage of calculations Sum of Squares (SS) Total is calculated and then separated into the two components SS Between Treatments and SS Within Treatments.
In the second stage the SS Between Treatments is separated into the three factors: Factor A,

Factor B & Factor A X B (interaction factor)

Source	*SS*	df	MS	F
Between Treatment (b/t)	SS_A + SS_B+SS_{A*B}	$(k_A -1)+ (k_B -1)$	$SS_{b/t}/df_{b/t}$	
Factor A	(identify from info. given)	$(k_A -1)$	SS_A / df_A	$MS_A / MS_{w/i}$
Factor B	(identify from info. given)	$(k_B -1)$	SS_B / df_B	$MS_B / MS_{w/i}$
Interaction	(identify from info. given)	$(k_A)(k_B -1)$	SS_{A*B} / df_{A*B}	$MS_{A*B} / MS_{w/i}$
Within Treatment (w/i)	$SS_{total} – SS_{b/t}$	$df_{total} – df_{b/t}$ or $N – df_{b/t}$	$SS_{w/i} / df_{w/i}$	
Total	$SS_{Between}$+SS_{within}	N – 1		

Table 2. ANOVA summary table with calculations

Note: In real life, we would run this through a statistical program with the raw data to calculate the Fs! We are focusing conceptually on calculating the 3 Fs for a two-way factorial ANOVA. Notice that in Table 2, the Sum of Squares Between is adding up the Sum of Squares from each of the factors. You also see that to get to our F-ratios, we need the Mean Squares (just like chapter 14). We have an F for each: Factor A, Factor B and the Interaction Factor. The calculations for Sum of Square for the factors can be found be knowing the df and MS, or knowing the Sum of Squares Between.

You would also be most likely given the means and standard deviations for the 4 study

conditions. Here is an example from Table 1 (made up data). You will see the main value as the mean and the standard deviation in parentheses.

	Gender	
Age	Teen Males M = 3.5 (.3)	Teen Females M = 4.5 (.25)
	Older Males M = 8 (.5)	Older Females M = 12.5 (.8)

Table 2. Means and Standard Deviations example from Table 1 study design.

Make your decision and explain the results (adjusted step 4).

- When making a statistical decision you should begin by looking for patterns in the means from each of the total conditions rather than focusing on the main effects or the interaction. After identifying patterns begin interpreting with the interaction effects first.
- Interaction means that the effect of one factor depends on the level of a second factor – so then there is no consistent main effect. If you get a significant interaction, emphasize that finding over any significant main effects. In other words, if there is an interaction effect, then the main effect cannot be discussed without a qualifier.

Calculate effect size

- Effect size is calculated for each F that is statistically significant.
- Effect size reported is typically eta-square. Remember that from chapter 12, eta-square is the percentage of total variance explained variance by the factor. Again, just as you have a F for factor A, a F for factor B, and an F for the interaction, you would have eta-squares for each.

13.2: Conducting a Two Factor ANOVA is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

13.3: Graphing the Results of Factorial Experiments

Graphing the Results of Factorial Experiments

The results of factorial experiments with two independent variables can be graphed by representing one independent variable on the x-axis and representing the other by using different kinds of bars or lines. The y-axis is always reserved for the dependent variable.

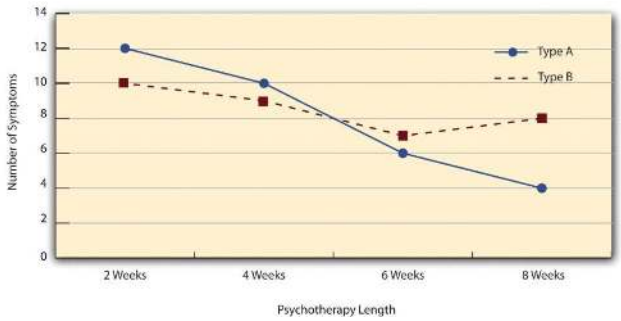


Figure 2. A 4 (Psychotherapy Length) x 2 (Type) ANOVA.

The figure above is a line graph that shows results for a hypothetical 4 x 2 factorial experiment Psychotherapy length, is represented along the x-axis and has four levels (e.g., 2 weeks, 4 weeks, 6 weeks and 8 weeks) and the other variable (psychotherapy type) is represented by differently formatted lines.

Advantages & Disadvantages

Considerations

A 2-Factor ANOVA design is relatively easy to carry out and requires fewer subjects than other types of designs. There is no pre-testing necessary because one group could serve as the control. Although identifying sample sizes and study design for power is an important consideration using a factorial ANOVA.

Disadvantages

A 2-Factor ANOVA using a between-subjects design provides little information about the effect of the independent variable. The statistic provides information about whether the two groups differed (on average) and in which direction but it is not sensitive to individual differences. Other considerations for 2-Factor ANOVAs include using a *repeated measures*

ANOVA. In this case for a 2-factor ANOVA, each person would be in every condition. So if you had a 2×2 an individual would be in all 4 study conditions. Another considerations is having a *mixed design*. For a mixed design, one factor would be between-subjects and the other would be within-subjects (repeated measures). For example, you might wish to conduct a 2×2 study on drug therapy. You can examine gender differences as one factor and type of drug as the other factor. Participants are only in 1 gender category but would receive both types of drug. A mixed design would give you individual differences in how each participant responded to the drug, but also has some of the challenges of using a within-subjects design (see short discussion in chapter 12 on advantages and disadvantages of using a repeated measures design).

Learning Objectives

Having read the chapter, students should be able to:

- Explain the concept of a two-factor research design and recognize a matrix with levels of one factor being rows and levels of the second factor being columns
- Explain main effects and interactions in a two-factor ANOVA including patterns of findings
- Complete a ANOVA table given some information from the study
- Interpret effect size

13.3: Graphing the Results of Factorial Experiments is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

13.E: Exercises

Exercises – Ch 15

1. True or false. The bigger the differences between the sample means, the more likely it is that at least one of the Fs will be significant.
2. True or false. The advantage of combining two factors into a single research study is that the two factor study provides information about the interaction of the two factors and the main effects of each factor.
3. Complete the ANOVA table given this is a 2×3 ANOVA (two-way ANOVA; factor A = 2 levels with n = 5; factor B = 3 levels with n = 5; N = 30)

Source	*SS*	df	MS	F
Between Treatment	60			
Factor A				5
Factor B				
Interaction	30			
Within Treatment			2	
Total	108			

4. What is the df for factor A, B and AxB for the following? What are the corresponding F-critical values?
    1. factor A n=14; factor B n = 18; N = 32

Answers to Exercises – Ch 15

1. true

3.

Source	*SS*	df	MS	F
Between Treatment	60	5		
Factor A	10	1	10	5
Factor B	20	2	10	5
Interaction	30	2	15	7.5
Within Treatment	48	24	2	
Total	108	29		

13.E: Exercises is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

CHAPTER OVERVIEW

14: Correlations

A common theme throughout statistics is the notion that individuals will differ on different characteristics and traits, which we call variance. In inferential statistics and hypothesis testing, our goal is to find systematic reasons for differences and rule out random chance as the cause. By doing this, we are using information on a different variable – which so far has been group membership like in ANOVA – to explain this variance. In correlations, we will instead use a continuous variable to account for the variance.

14.1: Variability and Covariance
14.2: Visualizing Relations
14.3: Three Characteristics
14.4: Pearson's r
14.5: Anxiety and Depression
14.6: Effect Size
14.7: Correlation versus Causation
14.8: Final Considerations
14.E: Correlations (Exercises)

This page titled 14: Correlations is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

14.1: Variability and Covariance

Because we have two continuous variables, we will have two characteristics or score on which people will vary. What we want to know is do people vary on the scores together. That is, as one score changes, does the other score also change in a predictable or consistent way? This notion of variables differing together is called covariance (the prefix "co" meaning "together").

Let's look at our formula for variance on a single variable:

$$s^2 = \frac{\sum (X - \overline{X})^2}{N - 1} \tag{14.1.1}$$

We use X to represent a person's score on the variable at hand, and $\overline{X}$ to represent the mean of that variable. The numerator of this formula is the Sum of Squares, which we have seen several times for various uses. Recall that squaring a value is just multiplying that value by itself. Thus, we can write the same equation as:

$$s^2 = \frac{\sum((X - \overline{X})(X - \overline{X}))}{N - 1} \tag{14.1.2}$$

This is the same formula and works the same way as before, where we multiply the deviation score by itself (we square it) and then sum across squared deviations.

Now, let's look at the formula for covariance. In this formula, we will still use X to represent the score on one variable, and we will now use Y to represent the score on the second variable. We will still use bars to represent averages of the scores. The formula for covariance (cov_{XY} with the subscript XY to indicate covariance across the X and Y variables) is:

$$\text{cov}_{XY} = \frac{\sum((X - \overline{X})(Y - \overline{Y}))}{N - 1} \tag{14.1.3}$$

As we can see, this is the exact same structure as the previous formula. Now, instead of multiplying the deviation score by itself on one variable, we take the deviation scores from a single person on each variable and multiply them together. We do this for each person (exactly the same as we did for variance) and then sum them to get our numerator. The numerator in this is called the Sum of Products.

$$SP = \sum((X - \overline{X})(Y - \overline{Y})) \tag{14.1.4}$$

We will calculate the sum of products using the same table we used to calculate the sum of squares. In fact, the table for sum of products is simply a sum of squares table for X, plus a sum of squares table for Y, with a final column of products, as shown below.

Table 14.1.1: Sum of Products table

X	$(X - \overline{X})$	$(X - \overline{X})^2$	Y	$(Y - \overline{Y})$	$(Y - \overline{Y})^2$	$(X - \overline{X})(Y - \overline{Y})$

This table works the same way that it did before (remember that the column headers tell you exactly what to do in that column). We list our raw data for the X and Y variables in the X and Y columns, respectively, then add them up so we can calculate the mean of each variable. We then take those means and subtract them from the appropriate raw score to get our deviation scores for each person on each variable, and the columns of deviation scores will both add up to zero. We will square our deviation scores for each variable to get the sum of squares for X and Y so that we can compute the variance and standard deviation of each (we will use the standard deviation in our equation below). Finally, we take the deviation score from each variable and multiply them together to get our product score. Summing this column will give us our sum of products. It is very important that you multiply the raw deviation scores from each variable, NOT the squared deviation scores.

Our sum of products will go into the numerator of our formula for covariance, and then we only have to divide by $N-1$ to get our covariance. Unlike the sum of squares, both our sum of products and our covariance can be positive, negative, or zero, and they will always match (e.g. if our sum of products is positive, our covariance will always be positive). A positive sum of products and covariance indicates that the two variables are related and move in the same direction. That is, as one variable goes up, the other will also go up, and vice versa. A negative sum of products and covariance means that the variables are related but move in opposite directions when they change, which is called an inverse relation. In an inverse relation, as one variable goes up, the other variable goes down. If the sum of products and covariance are zero, then that means that the variables are not related. As one variable goes up or down, the other variable does not change in a consistent or predictable way.

The previous paragraph brings us to an important definition about relations between variables. What we are looking for in a relation is a consistent or predictable pattern. That is, the variables change together, either in the same direction or opposite directions, in the same way each time. It doesn't matter if this relation is positive or negative, only that it is not zero. If there is no consistency in how the variables change within a person, then the relation is zero and does not exist. We will revisit this notion of direction vs zero relation later on.

This page titled 14.1: Variability and Covariance is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **12.1: Variability and Covariance** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

14.2: Visualizing Relations

Chapter 2 covered many different forms of data visualization, and visualizing data remains an important first step in understanding and describing out data before we move into inferential statistics. Nowhere is this more important than in correlation. Correlations are visualized by a scatterplot, where our X variable values are plotted on the X-axis, the Y variable values are plotted on the Y-axis, and each point or marker in the plot represents a single person's score on X and Y. Figure 14.2.1 shows a scatterplot for hypothetical scores on job satisfaction (X) and worker well-being (Y). We can see from the axes that each of these variables is measured on a 10-point scale, with 10 being the highest on both variables (high satisfaction and good health and well-being) and 1 being the lowest (dissatisfaction and poor health). When we look at this plot, we can see that the variables do seem to be related. The higher scores on job satisfaction tend to also be the higher scores on well-being, and the same is true of the lower scores.

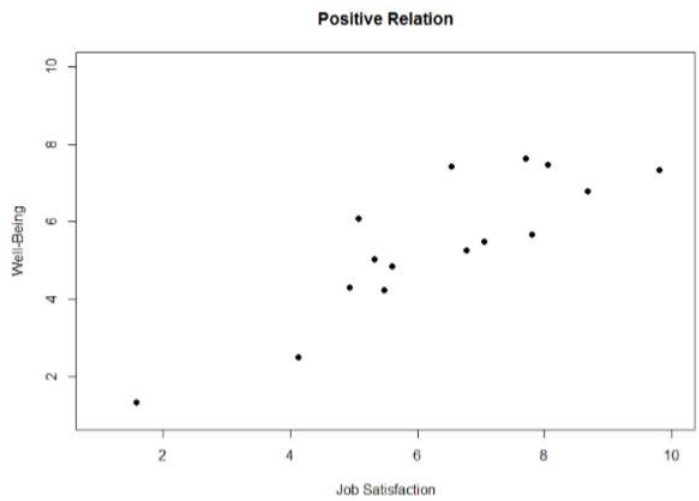


Figure 14.2.1: Plotting satisfaction and well-being scores.

Figure 14.2.1 demonstrates a positive relation. As scores on X increase, scores on Y also tend to increase. Although this is not a perfect relation (if it were, the points would form a single straight line), it is nonetheless very clearly positive. This is one of the key benefits to scatterplots: they make it very easy to see the direction of the relation. As another example, Figure 14.2.2 shows a negative relation between job satisfaction (X) and burnout (Y). As we can see from this plot, higher scores on job satisfaction tend to correspond to lower scores on burnout, which is how stressed, un-energetic, and unhappy someone is at their job. As with Figure 14.2.1, this is not a perfect relation, but it is still a clear one. As these figures show, points in a positive relation moves from the bottom left of the plot to the top right, and points in a negative relation move from the top left to the bottom right.

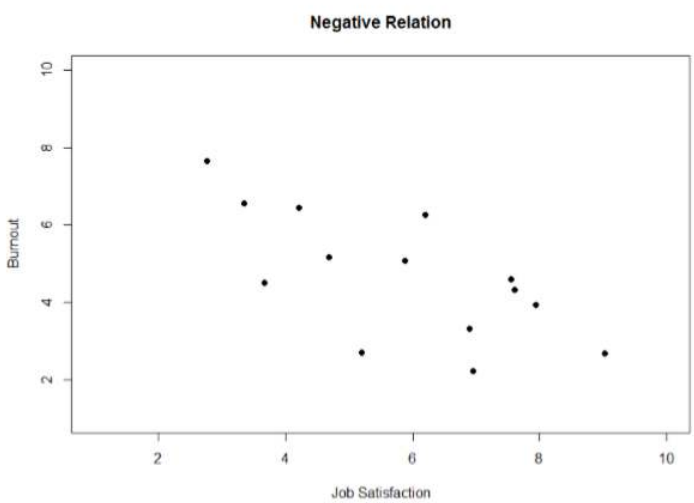


Figure 14.2.2: Plotting satisfaction and burnout scores.

Scatterplots can also indicate that there is no relation between the two variables. In these scatterplots (an example is shown below in Figure 14.2.3 plotting job satisfaction and job performance) there is no interpretable shape or line in the scatterplot. The points appear randomly throughout the plot. If we tried to draw a straight line through these points, it would basically be flat. The low scores on job satisfaction have roughly the same scores on job performance as do the high scores on job satisfaction. Scores in the middle or average range of job satisfaction have some scores on job performance that are about equal to the high and low levels and some scores on job performance that are a little higher, but the overall picture is one of inconsistency.

As we can see, scatterplots are very useful for giving us an approximate idea of whether or not there is a relation between the two variables and, if there is, if that relation is positive or negative. They are also useful for another reason: they are the only way to determine one of the characteristics of correlations that are discussed next: form.

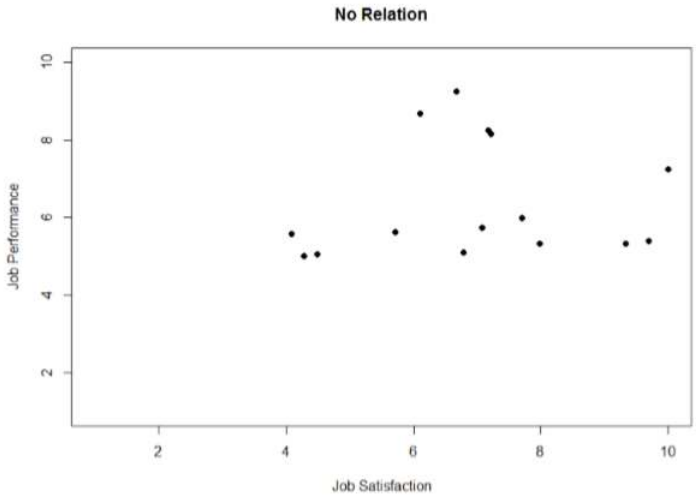


Figure 14.2.3: Plotting no relation between satisfaction and job performance.

This page titled 14.2: Visualizing Relations is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **12.2: Visualizing Relations** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

14.3: Three Characteristics

When we talk about correlations, there are three traits that we need to know in order to truly understand the relation (or lack of relation) between X and Y: form, direction, and magnitude. We will discuss each of them in turn.

Form The first characteristic of relations between variables is their form. The form of a relation is the shape it takes in a scatterplot, and a scatterplot is the only way it is possible to assess the form of a relation. there are two forms we look for: linear or no relation. A linear relation is what we saw in Figures 12.2.1, 12.2.2, and 12.2.3. If we drew a line through the middle points in the any of the scatterplots, we would be best suited with a straight line. The term “linear” comes from the word “line”. A linear relation is what we will always assume when we calculate correlations. All of the correlations presented here are only valid for linear relations. Thus, it is important to plot our data to make sure we meet this assumption.

Sometimes when we create a scatterplot, we end up with no interpretable relation at all. An example of this is shown below in Figure 14.3.4. The points in this plot show no consistency in relation, and a line through the middle would once again be a straight, flat line.

Sometimes when we look at scatterplots, it is tempting to get biased by a few points that fall far away from the rest of the points and seem to imply that there may be some sort of relation. These points are called outliers, and we will discuss them in more detail later in the chapter. These can be common, so it is important to formally test for a relation between our variables, not just rely on visualization. This is the point of hypothesis testing with correlations, and we will go in depth on it soon. First, however, we need to describe the other two characteristics of relations: direction and magnitude.

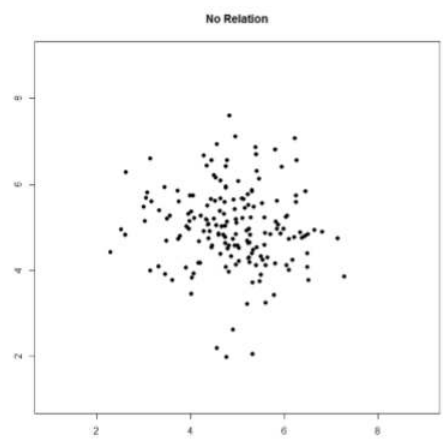


Figure 14.3.4: No relation

Direction

The direction of the relation between two variables tells us whether the variables change in the same way at the same time or in opposite ways at the same time. We saw this concept earlier when first discussing scatterplots, and we used the terms positive and negative. A positive relation is one in which X and Y change in the same direction: as X goes up, Y goes up, and as X goes down, Y also goes down. A negative relation is just the opposite: X and Y change together in opposite directions: as X goes up, Y goes down, and vice versa.

As we will see soon, when we calculate a correlation coefficient, we are quantifying the relation demonstrated in a scatterplot. That is, we are putting a number to it. That number will be either positive, negative, or zero, and we interpret the sign of the number as our direction. If the number is positive, it is a positive relation, and if it is negative, it is a negative relation. If it is zero, then there is no relation. The direction of the relation corresponds directly to the slope of the hypothetical line we draw through scatterplots when assessing the form of the relation. If the line has a positive slope that moves from bottom left to top right, it is positive, and vice versa for negative. If the line it flat, that means it has no slope, and there is no relation, which will in turn yield a zero for our correlation coefficient.

Magnitude

The number we calculate for our correlation coefficient, which we will describe in detail below, corresponds to the magnitude of the relation between the two variables. The magnitude is how strong or how consistent the relation between the variables is. Higher numbers mean greater magnitude, which means a stronger relation.

Our correlation coefficients will take on any value between -1.00 and 1.00, with 0.00 in the middle, which again represents no relation. A correlation of -1.00 is a perfect negative relation; as X goes up by some amount, Y goes down by the same amount, consistently. Likewise, a correlation of 1.00 indicates a perfect positive relation; as X goes up by some amount, Y also goes up by the same amount. Finally, a correlation of 0.00, which indicates no relation, means that as X goes up by some amount, Y may or may not change by any amount, and it does so inconsistently.

The vast majority of correlations do not reach -1.00 or positive 1.00. Instead, they fall in between, and we use rough cut offs for how strong the relation is based on this number. Importantly, the sign of the number (the direction of the relation) has no bearing on how strong the relation is. The only thing that matters is the magnitude, or the absolute value of the correlation coefficient. A correlation of -1 is just as strong as a correlation of 1. We generally use values of 0.10, 0.30, and 0.50 as indicating weak, moderate, and strong relations, respectively.

The strength of a relation, just like the form and direction, can also be inferred from a scatterplot, though this is much more difficult to do. Some examples of weak and strong relations are shown in Figures **14.3.5** and **14.3.6**, respectively. Weak correlations still have an interpretable form and direction, but it is much harder to see. Strong correlations have a very clear pattern, and the points tend to form a line. The examples show two different directions, but remember that the direction does not matter for the strength, only the consistency of the relation and the size of the number, which we will see next.

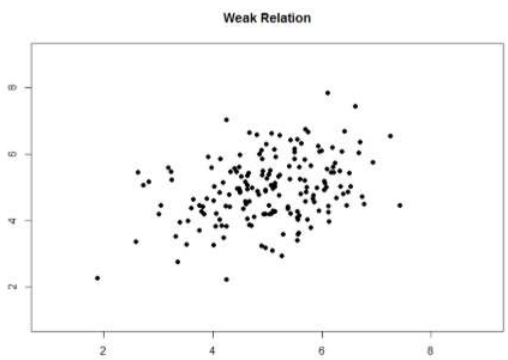


Figure **14.3.5**: Weak positive correlation.

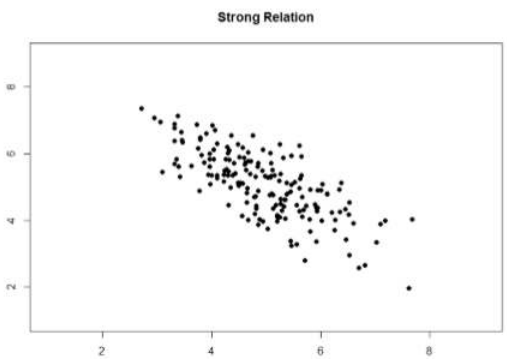


Figure **14.3.6**: Strong negative correlation.

This page titled 14.3: Three Characteristics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **12.3: Three Characteristics** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

14.4: Pearson's r

There are several different types of correlation coefficients, but we will only focus on the most common: Pearson's r. r is a very popular correlation coefficient for assessing linear relations, and it serves as both a descriptive statistic (like $\overline{X}$) and as a test statistic (like t). It is descriptive because it describes what is happening in the scatterplot; r will have both a sign (+/–) for the direction and a number (0 – 1 in absolute value) for the magnitude. As noted above, assumes a linear relation, so nothing about r will suggest what the form is – it will only tell what the direction and magnitude would be if the form is linear (Remember: always make a scatterplot first!). r also works as a test statistic because the magnitude of r will correspond directly to a t value as the specific degrees of freedom, which can then be compared to a critical value. Luckily, we do not need to do this conversion by hand. Instead, we will have a table of r critical values that looks very similar to our t table, and we can compare our r directly to those.

The formula for r is very simple: it is just the covariance (defined above) divided by the standard deviations of X and Y:

$$r = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{SP}{\sqrt{SSX * SSY}} \tag{14.4.1}$$

The first formula gives a direct sense of what a correlation is: a covariance standardized onto the scale of X and Y; the second formula is computationally simpler and faster. Both of these equations will give the same value, and as we saw at the beginning of the chapter, all of these values are easily computed by using the sum of products table. When we do this calculation, we will find that our answer is always between -1.00 and 1.00 (if it's not, check the math again), which gives us a standard, interpretable metric, similar to what z-scores did.

It was stated earlier that r is a descriptive statistic like $\overline{X}$, and just like $\overline{X}$, it corresponds to a population parameter. For correlations, the population parameter is the lowercase Greek letter ρ ("rho"); be careful not to confuse ρ with a p-value – they look quite similar. r is an estimate of ρ just like $\overline{X}$ is an estimate of μ. Thus, we will test our observed value of r that we calculate from the data and compare it to a value of ρ specified by our null hypothesis to see if the relation between our variables is significant, as we will see in our example next.

This page titled 14.4: Pearson's r is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **12.4: Pearson's r** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

14.5: Anxiety and Depression

Anxiety and depression are often reported to be highly linked (or "comorbid"). Our hypothesis testing procedure follows the same four-step process as before, starting with our null and alternative hypotheses. We will look for a positive relation between our variables among a group of 10 people because that is what we would expect based on them being comorbid.

Step 1: State the Hypotheses

Our hypotheses for correlations start with a baseline assumption of no relation, and our alternative will be directional if we expect to find a specific type of relation. For this example, we expect a positive relation:

$$\mathrm{H}_0: \text{ There is no relation between anxiety and depression}$$
$$\mathrm{H}_0: \rho = 0$$

$$\mathrm{H}_\mathrm{A}: \text{There is a positive relation between anxiety and depression}$$
$$\mathrm{H}_0: \rho > 0$$

Remember that ρ ("rho") is our population parameter for the correlation that we estimate with r, just like $\overline{X}$ and μ for means. Remember also that if there is no relation between variables, the magnitude will be 0, which is where we get the null and alternative hypothesis values.

Step 2: Find the Critical Values

The critical values for correlations come from the correlation table, which looks very similar to the t-table (see Figure 14.5.1). Just like our t-table, the column of critical values is based on our significance level (α) and the directionality of our test. The row is

determined by our degrees of freedom. For correlations, we have $N-2$ degrees of freedom, rather than $N-1$ (why this is the case is not important). For our example, we have 10 people, so our degrees of freedom = 10 – 2 = 8.

	Level of Significance for One-Tailed Test				
	.05	.025	.01	.005	
	Level of Significance for Two-Tailed Test				
	.10	.05	.02	.01	
df	1	.988	.997	.9995	.9999
	2	.900	.950	.980	.990
	3	.805	.878	.934	.959
	4	.729	.811	.882	.917
	5	.669	.754	.833	.875
	6	.621	.707	.789	.834
	7	.582	.666	.750	.798
	8	.549	.632	.715	.765
	9	.521	.602	.685	.735
	10	.497	.576	.658	.708
	11	.476	.553	.634	.684
	12	.458	.532	.612	.661
	13	.441	.514	.592	.641
	14	.426	.497	.574	.623
	15	.412	.482	.558	.606
	16	.400	.468	.543	.590
	17	.389	.456	.529	.575
	18	.378	.444	.516	.561
	19	.369	.433	.503	.549
	20	.360	.423	.492	.537
	21	.352	.413	.482	.526
	22	.344	.404	.472	.515
	23	.337	.396	.462	.505
	24	.330	.388	.453	.496
	25	.323	.381	.445	.487
	26	.317	.374	.437	.479
	27	.311	.367	.430	.471
	28	.306	.361	.423	.463

	Level of Significance for One-Tailed Test				
	.05	.025	.01	.005	
	Level of Significance for Two-Tailed Test				
	.10	.05	.02	.01	
	29	.301	.355	.416	.456
	30	.296	.349	.409	.449
	35	.275	.325	.381	.418
	40	.257	.304	.358	.393
	45	.243	.288	.338	.372
	50	.231	.273	.322	.354
	60	.211	.250	.295	.325
	70	.195	.232	.274	.302
	80	.183	.217	.257	.283
	90	.173	.205	.242	.267
	100	.164	.195	.230	.254

Figure 14.5.1: Correlation table

We were not given any information about the level of significance at which we should test our hypothesis, so we will assume $\alpha = 0.05$ as always. From our table, we can see that a 1-tailed test (because we expect only a positive relation) at the $\alpha = 0.05$ level has a critical value of $r^* = 0.549$. Thus, if our observed correlation is greater than 0.549, it will be statistically significant. This is a rather high bar (remember, the guideline for a strong relation is r = 0.50); this is because we have so few people. Larger samples make it easier to find significant relations.

Step 3: Calculate the Test Statistic

We have laid out our hypotheses and the criteria we will use to assess them, so now we can move on to our test statistic. Before we do that, we must first create a scatterplot of the data to make sure that the most likely form of our relation is in fact linear. Figure 14.5.2 below shows our data plotted out, and it looks like they are, in fact, linearly related, so Pearson's r is appropriate.

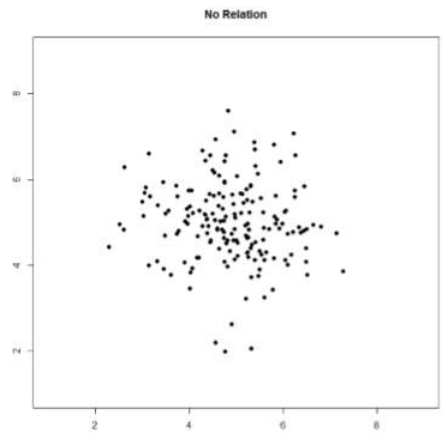


Figure 14.5.2: Scatterplot of anxiety and depression

The data we gather from our participants is as follows:

Table 14.5.1: Data from participants

Dep	2.81	1.96	3.43	3.40	4.71	1.80	4.27	3.68	2.44	3.13
Anx	3.54	3.05	3.81	3.43	4.03	3.59	4.17	3.46	3.19	4.12

We will need to put these values into our Sum of Products table to calculate the standard deviation and covariance of our variables. We will use X for depression and Y for anxiety to keep track of our data, but be aware that this choice is arbitrary and the math will work out the same if we decided to do the opposite. Our table is thus:

Table 14.5.2: Sum of Products table

X	$(X-\overline{X})$	$(X-\overline{X})^2$	Y	$(Y-\overline{Y})$	$(Y-\overline{Y})^2$	$(X-\overline{X})(Y-\overline{Y})$
2.81	-0.35	0.12	3.54	-0.10	0.01	0.04
1.96	-1.20	1.44	3.05	-0.59	0.35	0.71
3.43	0.27	0.07	3.81	0.17	0.03	0.05
3.40	0.24	0.06	3.43	-0.21	0.04	-0.05
4.71	1.55	2.40	4.03	0.39	0.15	0.60
1.80	-1.36	1.85	3.59	-0.05	0.00	0.07
4.27	1.11	1.23	4.17	0.53	0.28	0.59
3.68	0.52	0.27	3.46	-0.18	0.03	-0.09
2.44	-0.72	0.52	3.19	-0.45	0.20	0.32
3.13	-0.03	0.00	4.12	0.48	0.23	-0.01
31.63	0.03	7.97	36.39	-0.01	1.33	2.22

The bottom row is the sum of each column. We can see from this that the sum of the X observations is 31.63, which makes the mean of the X variable $\overline{X} = 3.16$. The deviation scores for X sum to 0.03, which is very close to 0, given rounding error, so everything looks right so far. The next column is the squared deviations for X, so we can see that the sum of squares for X is $SS_X = 7.97$. The same is true of the Y columns, with an average of $\overline{Y} = 3.64$, deviations that sum to zero within rounding error, and a

sum of squares as $SS_Y = 1.33$. The final column is the product of our deviation scores (NOT of our squared deviations), which gives us a sum of products of $SP = 2.22$.

There are now three pieces of information we need to calculate before we compute our correlation coefficient: the covariance of X and Y and the standard deviation of each.

The covariance of two variable, remember, is the sum of products divided by N–1. For our data:

$$\text{cov}_{XY} = \frac{SP}{N-1} = \frac{2.22}{9} = 0.25$$

The formula for standard deviation are the same as before. Using subscripts X and Y to denote depression and anxiety:

$$s_X = \sqrt{\frac{\sum(X-\overline{X})^2}{N-1}} = \sqrt{\frac{7.97}{9}} = 0.94$$

$$s_Y = \sqrt{\frac{\sum(Y-\overline{Y})^2}{N-1}} = \sqrt{\frac{1.33}{9}} = 0.38$$

Now we have all of the information we need to calculate r:

$$r = \frac{\text{cov}_{XY}}{s_X S_Y} = \frac{0.25}{0.94 * 0.38} = 0.70$$

We can verify this using our other formula, which is computationally shorter:

$$r = \frac{SP}{\sqrt{SSX * SSY}} = \frac{2.22}{\sqrt{7.97 * 1.33}} = .70$$

So our observed correlation between anxiety and depression is $r = 0.70$, which, based on sign and magnitude, is a strong, positive correlation. Now we need to compare it to our critical value to see if it is also statistically significant.

Step 4: Make a Decision

Our critical value was $r^* = 0.549$ and our obtained value was $r = 0.70$. Our obtained value was larger than our critical value, so we can reject the null hypothesis.

Reject H_0. Based on our sample of 10 people, there is a statistically significant, strong, positive relation between anxiety and depression, $r(8) = 0.70, p < .05$.

Notice in our interpretation that, because we already know the magnitude and direction of our correlation, we can interpret that. We also report the degrees of freedom, just like with t, and we know that $p < \alpha$ because we rejected the null hypothesis. As we can see, even though we are dealing with a very different type of data, our process of hypothesis testing has remained unchanged.

This page titled 14.5: Anxiety and Depression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **12.5: Anxiety and Depression** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

14.6: Effect Size

Pearson's r is an incredibly flexible and useful statistic. Not only is it both descriptive and inferential, as we saw above, but because it is on a standardized metric (always between -1.00 and 1.00), it can also serve as its own effect size. In general, we use r = 0.10, r = 0.30, and r = 0.50 as our guidelines for small, medium, and large effects. Just like with Cohen's d, these guidelines are not absolutes, but they do serve as useful indicators in most situations. Notice as well that these are the same guidelines we used earlier to interpret the magnitude of the relation based on the correlation coefficient.

In addition to r being its own effect size, there is an additional effect size we can calculate for our results. This effect size is r^2, and it is exactly what it looks like – it is the squared value of our correlation coefficient. Just like η^2 in ANOVA, r^2 is interpreted as the amount of variance explained in the outcome variance, and the cut scores are the same as well: 0.01, 0.09, and 0.25 for small, medium, and large, respectively. Notice here that these are the same cutoffs we used for regular r effect sizes, but squared (0.102 = 0.01, 0.302 = 0.09, 0.502 = 0.25) because, again, the r^2 effect size is just the squared correlation, so its interpretation should be, and is, the same. The reason we use r^2 as an effect size is because our ability to explain variance is often important to us.

This page titled 14.6: Effect Size is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **12.6: Effect Size** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

14.7: Correlation versus Causation

We cover a great deal of material in introductory statistics and, as mentioned chapter 1, many of the principles underlying what we do in statistics can be used in your day to day life to help you interpret information objectively and make better decisions. We now come to what may be the most important lesson in introductory statistics: the difference between correlation and causation.

It is very, very tempting to look at variables that are correlated and assume that this means they are causally related; that is, it gives the impression that X is causing Y.

However, in reality, correlation do not – and cannot – do this. Correlations DO NOT prove causation. No matter how logical or how obvious or how convenient it may seem, no correlational analysis can demonstrate causality. The ONLY way to demonstrate a causal relation is with a properly designed and controlled experiment.

Many times, we have good reason for assessing the correlation between two variables, and often that reason will be that we suspect that one causes the other. Thus, when we run our analyses and find strong, statistically significant results, it is very tempting to say that we found the causal relation that we are looking for. The reason we cannot do this is that, without an experimental design that includes random assignment and control variables, the relation we observe between the two variables may be caused by something else that we failed to measure. These "third variables" are lurking variables or confound variables, and they are impossible to detect and control for without an experiment.

Confound variables, which we will represent with Z, can cause two variables X and Y to appear related when in fact they are not. They do this by being the hidden – or lurking – cause of each variable independently. That is, if Z causes X and Z causes Y, the X and Y will appear to be related . However, if we control for the effect of Z (the method for doing this is beyond the scope of this text), then the relation between X and Y will disappear.

A popular example for this effect is the correlation between ice cream sales and deaths by drowning. These variables are known to correlate very strongly over time. However, this does not prove that one causes the other. The lurking variable in this case is the weather – people enjoy swimming and enjoy eating ice cream more during hot weather as a way to cool off. As another example, consider shoe size and spelling ability in elementary school children. Although there should clearly be no causal relation here, the variables and nonetheless consistently correlated. The confound in this case? Age. Older children spell better than younger children and are also bigger, so they have larger shoes.

When there is the possibility of confounding variables being the hidden cause of our observed correlation, we will often collect data on Z as well and control for it in our analysis. This is good practice and a wise thing for researchers to do. Thus, it would seem that it is easy to demonstrate causation with a correlation that controls for Z. However, the number of variables that could potentially cause a correlation between X and Y is functionally limitless, so it would be impossible to control for everything. That is why we use experimental designs; by randomly assigning people to groups and manipulating variables in those groups, we can balance out individual differences in any variable that may be our cause.

It is not always possible to do an experiment, however, so there are certain situations in which we will have to be satisfied with our observed relation and do the best we can to control for known confounds. However, in these situations, even if we do an excellent job of controlling for many extraneous (a statistical and research term for "outside") variables, we must be very careful not to use causal language. That is because, even after controls, sometimes variables are related just by chance.

Sometimes, variables will end up being related simply due to random chance, and we call these correlation spurious. Spurious just means random, so what we are seeing is random correlations because, given enough time, enough variables, and enough data, sampling error will eventually cause some variables to be related when they should not. Sometimes, this even results in incredibly strong, but completely nonsensical, correlations. This becomes more and more of a problem as our ability to collect massive datasets and dig through them improves, so it is very important to think critically about any relation you encounter.

This page titled 14.7: Correlation versus Causation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **12.7: Correlation versus Causation** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

14.8: Final Considerations

Correlations, although simple to calculate, and be very complex, and there are many additional issues we should consider. We will discuss some other correlations and reporting methods you may encounter.

Outliers

Another issue that can cause the observed size of our correlation to be inappropriately large or small is the presence of outliers. An outlier is a data point that falls far away from the rest of the observations in the dataset. Sometimes outliers are the result of incorrect data entry, poor or intentionally misleading responses, or simple random chance. Other times, however, they represent real people with meaningful values on our variables. The distinction between meaningful and accidental outliers is a difficult one that is based on the expert judgment of the researcher. Sometimes, we will remove the outlier (if we think it is an accident) or we may decide to keep it (if we find the scores to still be meaningful even though they are different).

The plots below in Figure 14.8.3 show the effects that an outlier can have on data. In the first, we have our raw dataset. You can see in the upper right corner that there is an outlier observation that is very far from the rest of our observations on both the X and Y variables. In the middle, we see the correlation computed when we include the outlier, along with a straight line representing the relation; here, it is a positive relation. In the third image, we see the correlation after removing the outlier, along with a line showing the direction once again. Not only did the correlation get stronger, it completely changed direction!

In general, there are three effects that an outlier can have on a correlation: it can change the magnitude (make it stronger or weaker), it can change the significance (make a non-significant correlation significant or vice versa), and/or it can change the direction (make a positive relation negative or vice versa). Outliers are a big issue in small datasets where a single observation can have a strong weight compared to the rest. However, as our samples sizes get very large (into the hundreds), the effects of outliers diminishes because they are outweighed by the rest of the data. Nevertheless, no matter how large a dataset you have, it is always a good idea to screen for outliers, both statistically (using analyses that we do not cover here) and/or visually (using scatterplots).

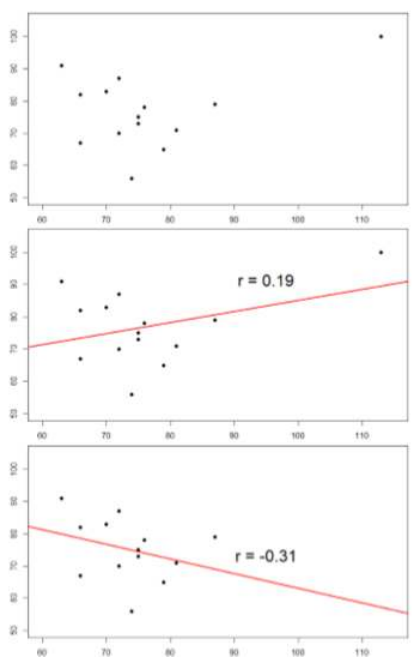


Figure 14.8.3: Three plots showing correlations with and without outliers.

Other Correlation Coefficients

In this chapter we have focused on Pearson's r as our correlation coefficient because it very common and very useful. There are, however, many other correlations out there, each of which is designed for a different type of data. The most common of these is Spearman's rho (ρ), which is designed to be used on ordinal data rather than continuous data. This is a very useful analysis if we

have ranked data or our data do not conform to the normal distribution. There are even more correlations for ordered categories, but they are much less common and beyond the scope of this chapter.

Correlation Matrices

Many research studies look at the relation between more than two continuous variables. In such situations, we could simply list our all of our correlations, but that would take up a lot of space and make it difficult to quickly find the relation we are looking for. Instead, we create correlation matrices so that we can quickly and simply display our results. A matrix is like a grid that contains our values. There is one row and one column for each of our variables, and the intersections of the rows and columns for different variables contain the correlation for those two variables.

At the beginning of the chapter, we saw scatterplots presenting data for correlations between job satisfaction, well-being, burnout, and job performance. We can create a correlation matrix to quickly display the numerical values of each. Such a matrix is shown below.

Table 14.8.1: Correlation matrix to display the numerical values

	Satisfaction	Well-Being	Burnout	Performance
Satisfaction	1.00			
Well-Being	0.41	1.00		
Burnout	-0.54	-0.87	1.00	
Performance	0.08	0.21	-0.33	1.00

Notice that there are values of 1.00 where each row and column of the same variable intersect. This is because a variable correlates perfectly with itself, so the value is always exactly 1.00. Also notice that the upper cells are left blank and only the cells below the diagonal of 1s are filled in. This is because correlation matrices are symmetrical: they have the same values above the diagonal as below it. Filling in both sides would provide redundant information and make it a bit harder to read the matrix, so we leave the upper triangle blank.

Correlation matrices are a very condensed way of presenting many results quickly, so they appear in almost all research studies that use continuous variables. Many matrices also include columns that show the variable means and standard deviations, as well as asterisks showing whether or not each correlation is statistically significant.

This page titled 14.8: Final Considerations is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **12.8: Final Considerations** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

14.E: Correlations (Exercises)

1. What does a correlation assess?

Answer:

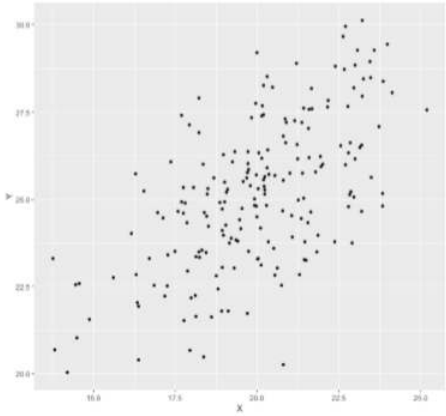
Correlations assess the linear relation between two continuous variables

2. What are the three characteristics of a correlation coefficient?
3. What is the difference between covariance and correlation?

Answer:

Covariance is an unstandardized measure of how related two continuous variables are. Correlations are standardized versions of covariance that fall between negative 1 and positive 1.

4. Why is it important to visualize correlational data in a scatterplot before performing analyses?
5. What sort of relation is displayed in the scatterplot below?



Answer:

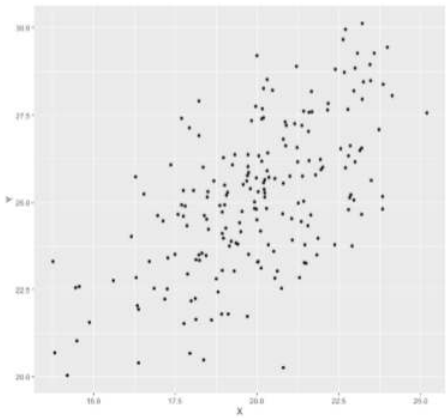
Strong, positive, linear relation

6. What is the direction and magnitude of the following correlation coefficients
   a. -0.81
   b. 0.40
   c. 0.15
   d. -0.08
   e. 0.29
7. Create a scatterplot from the following data:

Hours Studying	Overall Class Performance
0.62	2.02
1.50	4.62
0.34	2.60
0.97	1.59
3.54	4.67
0.69	2.52
1.53	2.28
0.32	1.68
1.94	2.50
1.25	4.04
1.42	2.63
3.07	3.53
3.99	3.90
1.73	2.75
1.9	2.95

Answer:

Your scatterplot should look similar to this:



8. In the following correlation matrix, what is the relation (number, direction, and magnitude) between…
   a. Pay and Satisfaction
   b. Stress and Health

Workplace	Pay	Satisfaction	Stress	Health
Pay	1.00			
Satisfaction	0.68	1.00		
Stress	0.02	-0.23	1.00	
Health	0.05	0.15	-0.48	1.00

9. Using the data from problem 7, test for a statistically significant relation between the variables.

Answer:

Step 1: $H_0 : \rho = 0$, "There is no relation between time spent studying and overall performance in class", $H_A : \rho > 0$, "There is a positive relation between time spent studying and overall performance in class."

Step 2: $df = 15–2 = 13, \alpha = 0.05$, 1-tailed test, $r^* = 0.441$.

Step 3: Using the Sum of Products table, you should find: $\overline{X} = 1.61, SS_X = 17.44, \overline{Y} = 2.95, SS_Y = 13.60, SP = 10.06, r = 0.65$.

Step 4: Obtained statistic is greater than critical value, reject H_0. There is a statistically significant, strong, positive relation between time spent studying and performance in class, $r(13) = 0.65, p < .05$.

10. A researcher collects data from 100 people to assess whether there is any relation between level of education and levels of civic engagement. The researcher finds the following descriptive values: $\overline{X} = 4.02, s_x = 1.15, \overline{Y} = 15.92, s_y = 5.01, SS_X = 130.93, SS_Y = 2484.91, SP = 159.39$. Test for a significant relation using the four step hypothesis testing procedure.

This page titled 14.E: Correlations (Exercises) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **12.E: Correlations (Exercises)** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

15: Chi-square

15.1: Categories and Frequency Tables
15.2: Goodness-of-Fit
15.3: χ^2 Statistic
15.4: Pineapple on Pizza
15.5: Contingency Tables for Two Variables
15.6: Test for Independence
15.7: College Sports
15.E: Chi-square (Exercises)

This page titled 15: Chi-square is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

15.1: Categories and Frequency Tables

Our data for the χ^2 test are categorical, specifically nominal, variables. Recall from unit 1 that nominal variables have no specified order and can only be described by their names and the frequencies with which they occur in the dataset. Thus, unlike our other variables that we have tested, we cannot describe our data for the χ^2 test using means and standard deviations. Instead, we will use frequencies tables.

Table 15.1.1: Pet Preferences

	Cat	Dog	Other	Total
Observed	14	17	5	36
Expected	12	12	12	36

Table 15.1.1 gives an example of a frequency table used for a χ^2 test. The columns represent the different categories within our single variable, which in this example is pet preference. The χ^2 test can assess as few as two categories, and there is no technical upper limit on how many categories can be included in our variable, although, as with ANOVA, having too many categories makes our computations long and our interpretation difficult. The final column in the table is the total number of observations, or N. The χ^2 test assumes that each observation comes from only one person and that each person will provide only one observation, so our total observations will always equal our sample size.

There are two rows in this table. The first row gives the observed frequencies of each category from our dataset; in this example, 14 people reported liking preferring cats as pets, 17 people reported preferring dogs, and 5 people reported a different animal. The second row gives expected values; expected values are what would be found if each category had equal representation. The calculation for an expected value is:

$$E = \frac{N}{C} \tag{15.1.1}$$

Where N is the total number of people in our sample and C is the number of categories in our variable (also the number of columns in our table). The expected values correspond to the null hypothesis for χ^2 tests: equal representation of categories. Our first of two χ^2 tests, the Goodness-of-Fit test, will assess how well our data lines up with, or deviates from, this assumption.

This page titled 15.1: Categories and Frequency Tables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **14.1: Categories and Frequency Tables** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

15.2: Goodness-of-Fit

The first of our two χ^2 tests assesses one categorical variable against a null hypothesis of equally sized frequencies. Equal frequency distributions are what we would expect to get if categorization was completely random. We could, in theory, also test against a specific distribution of category sizes if we have a good reason to (e.g. we have a solid foundation of how the regular population is distributed), but this is less common, so we will not deal with it in this text.

Hypotheses

All χ^2 tests, including the goodness-of-fit test, are non-parametric. This means that there is no population parameter we are estimating or testing against; we are working only with our sample data. Because of this, there are no mathematical statements for χ^2 hypotheses. This should make sense because the mathematical hypothesis statements were always about population parameters (e.g. μ), so if we are non-parametric, we have no parameters and therefore no mathematical statements.

We do, however, still state our hypotheses verbally. For goodness-of-fit χ^2 tests, our null hypothesis is that there is an equal number of observations in each category. That is, there is no difference between the categories in how prevalent they are. Our alternative hypothesis says that the categories do differ in their frequency. We do not have specific directions or one-tailed tests for χ^2, matching our lack of mathematical statement.

Degrees of Freedom and the χ^2 table

Our degrees of freedom for the χ^2 test are based on the number of categories we have in our variable, not on the number of people or observations like it was for our other tests. Luckily, they are still as simple to calculate:

$$df = C - 1 \tag{15.2.1}$$

So for our pet preference example, we have 3 categories, so we have 2 degrees of freedom. Our degrees of freedom, along with our significance level (still defaulted to $\alpha = 0.05$) are used to find our critical values in the χ^2 table, which is shown in figure 1. Because we do not have directional hypotheses for χ^2 tests, we do not need to differentiate between critical values for 1- or 2-tailed tests. In fact, just like our F tests for ANOVA, all χ^2 tests are 1-tailed tests.

	Chi-Square Right-Tail Probability ($\geq \chi^2$)									
DF	**0.995**	**0.99**	**0.975**	**0.95**	**0.9**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

This page titled 15.2: Goodness-of-Fit is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **14.2: Goodness-of-Fit** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

15.3: χ² Statistic

The calculations for our test statistic in χ^2 tests combine our information from our observed frequencies (O) and our expected frequencies (E) for each level of our categorical variable. For each cell (category) we find the difference between the observed and expected values, square them, and divide by the expected values. We then sum this value across cells for our test statistic. This is shown in the formula:

$$\chi^2 = \sum \frac{(\mathrm{O} - \mathrm{E})^2}{\mathrm{E}} \tag{15.3.1}$$

For our pet preference data, we would have:

$$\chi^2 = \frac{(14-12)^2}{12} + \frac{(17-12)^2}{12} + \frac{(5-12)^2}{12} = 0.33 + 2.08 + 4.08 = 6.49$$

Notice that, for each cell's calculation, the expected value in the numerator and the expected value in the denominator are the same value. Let's now take a look at an example from start to finish.

This page titled 15.3: χ² Statistic is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **14.3: χ² Statistic** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

15.4: Pineapple on Pizza

There is a very passionate and on-going debate on whether or not pineapple should go on pizza. Being the objective, rational data analysts that we are, we will collect empirical data to see if we can settle this debate once and for all. We gather data from a group of adults asking for a simple Yes/No answer.

Step 1: State the Hypotheses

We start, as always, with our hypotheses. Our null hypothesis of no difference will state that an equal number of people will say they do or do not like pineapple on pizza, and our alternative will be that one side wins out over the other:

$$\mathrm{H}_0: \text{ An equal number of people do and do not like pineapple on pizza}$$

$$\mathrm{H}_A: A \text{ significant majority of people will agree one way or the other}$$

Step 2: Find the Critical Value

To avoid any potential bias in this crucial analysis, we will leave α at its typical level. We have two options in our data (Yes or No), which will give us two categories. Based on this, we will have 1 degree of freedom. From our χ^2 table, we find a critical value of 3.84.

Step 3: Calculate the Test Statistic

The results of the data collection are presented in Table $15.4.1$. We had data from 45 people in all and 2 categories, so our expected values are $E = 45/2 = 22.50$.

Table $15.4.1$: Results of Data collection

	Yes	No	Total
Observed	26	19	45
Expected	22.50	22.50	45

We can use these to calculate our χ^2 statistic:

$$\chi^2 = \frac{(26-22.50)^2}{22.50} + \frac{(19-22.50)^2}{22.50} = 0.54 + 0.54 = 1.08$$

Step 4: Make the Decision

Our observed test statistic had a value of 1.08 and our critical value was 3.84. Our test statistic was smaller than our critical value, so we fail to reject the null hypothesis, and the debate rages on.

This page titled 15.4: Pineapple on Pizza is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **14.4: Pineapple on Pizza** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

15.5: Contingency Tables for Two Variables

The goodness-of-fit test is a useful tool for assessing a single categorical variable. However, what is more common is wanting to know if two categorical variables are related to one another. This type of analysis is similar to a correlation, the only difference being that we are working with nominal data, which violates the assumptions of traditional correlation coefficients. This is where the χ^2 test for independence comes in handy.

As noted above, our only description for nominal data is frequency, so we will again present our observations in a frequency table. When we have two categorical variables, our frequency table is crossed. That is, each combination of levels from each categorical variable are presented. This type of frequency table is called a contingency table because it shows the frequency of each category in one variable, contingent upon the specific level of the other variable.

An example contingency table is shown in Table 15.5.1, which displays whether or not 168 college students watched college sports growing up (Yes/No) and whether the students' final choice of which college to attend was influenced by the college's sports teams (Yes – Primary, Yes – Somewhat, No):

Table 15.5.1: Contingency table of college sports and decision making

College Sports		Affected Decision			
		Primary	Somewhat	No	Total
Watched	Yes	47	26	14	87
	No	21	23	37	81
	Total	68	49	51	168

In contrast to the frequency table for our goodness-of-fit test, our contingency table does not contain expected values, only observed data. Within our table, wherever our rows and columns cross, we have a cell. A cell contains the frequency of observing it's corresponding specific levels of each variable at the same time. The top left cell in Table 15.5.1 shows us that 47 people in our study watched college sports as a child AND had college sports as their primary deciding factor in which college to attend.

Cells are numbered based on which row they are in (rows are numbered top to bottom) and which column they are in (columns are numbered left to right). We always name the cell using (R,C), with the row first and the column second. A quick and easy way to remember the order is that R/C Cola exists but C/R Cola does not. Based on this convention, the top left cell containing our 47 participants who watched college sports as a child and had sports as a primary criteria is cell (1,1). Next to it, which has 26 people who watched college sports as a child but had sports only somewhat affect their decision, is cell (1,2), and so on. We only number the cells where our categories cross. We do not number our total cells, which have their own special name: marginal values.

Marginal values are the total values for a single category of one variable, added up across levels of the other variable. In table 3, these marginal values have been italicized for ease of explanation, though this is not normally the case. We can see that, in total, 87 of our participants (47+26+14) watched college sports growing up and 81 (21+23+37) did not. The total of these two marginal values is 168, the total number of people in our study. Likewise, 68 people used sports as a primary criteria for deciding which college to attend, 49 considered it somewhat, and 51 did not use it as criteria at all. The total of these marginal values is also 168, our total number of people. The marginal values for rows and columns will always both add up to the total number of participants, N, in the study. If they do not, then a calculation error was made and you must go back and check your work.

Expected Values of Contingency Tables

Our expected values for contingency tables are based on the same logic as they were for frequency tables, but now we must incorporate information about how frequently each row and column was observed (the marginal values) and how many people were in the sample overall (N) to find what random chance would have made the frequencies out to be. Specifically:

$$E_{ij} = \frac{R_i C_j}{N} \tag{15.5.1}$$

The subscripts i and j indicate which row and column, respectively, correspond to the cell we are calculating the expected frequency for, and the R_i and C_j are the row and column marginal values, respectively. N is still the total sample size. Using the data from Table 15.5.1, we can calculate the expected frequency for cell (1,1), the college sport watchers who used sports at their primary criteria, to be:

$$E_{1,1} = \frac{87 * 68}{168} = 35.21$$

We can follow the same math to find all the expected values for this table:

Table 15.5.2: Contingency table of college sports and decision making

College Sports		Affected Decision			Total
		Primary	Somewhat	No	
Watched	Yes	35.21	25.38	26.41	87
	No	32.79	23.62	24.59	81
	Total	68	49	51	

Notice that the marginal values still add up to the same totals as before. This is because the expected frequencies are just row and column averages simultaneously. Our total N will also add up to the same value.

The observed and expected frequencies can be used to calculate the same χ^2 statistic as we did for the goodness-of-fit test. Before we get there, though, we should look at the hypotheses and degrees of freedom used for contingency tables.

This page titled 15.5: Contingency Tables for Two Variables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **14.5: Contingency Tables for Two Variables** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

15.6: Test for Independence

The χ^2 test performed on contingency tables is known as the **test for independence**. In this analysis, we are looking to see if the values of each categorical variable (that is, the frequency of their levels) is related to or independent of the values of the other categorical variable. Because we are still doing a χ^2 test, which is nonparametric, we still do not have mathematical versions of our hypotheses. The actual interpretations of the hypotheses are quite simple: the null hypothesis says that the variables are independent or not related, and alternative says that they are not independent or that they are related. Using this set up and the data provided in Table 14.5.2, let's formally test for whether or not watching college sports as a child is related to using sports as a criteria for selecting a college to attend.

This page titled 15.6: Test for Independence is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **14.6: Test for Independence** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

15.7: College Sports

We will follow the same 4 step procedure as we have since chapter 7.

Step 1: State the Hypotheses

Our null hypothesis of no difference will state that there is no relation between our variables, and our alternative will state that our variables are related:

$$\mathrm{H_0:\ College\ choice\ criteria\ is\ independent\ of\ college\ sports\ viewership\ as\ a\ child}$$

$$\mathrm{H_A:\ College\ choice\ criteria\ is\ related\ to\ college\ sports\ viewership\ as\ a\ child}$$

Step 2: Find the Critical Value

Our critical value will come from the same table that we used for the goodness-of-fit test, but our degrees of freedom will change. Because we now have rows and columns (instead of just columns) our new degrees of freedom use information on both:

$$df = (R-1)(C-1) \tag{15.7.1}$$

In our example:

$$df = (2-1)(3-1) = 1 * 2 = 2$$

Based on our 2 degrees of freedom, our critical value from our table is 5.991.

Step 3: Calculate the Test Statistic

The same formula for χ^2 is used once again:

$$\chi^2 = \sum \frac{(0-\mathrm{E})^2}{\mathrm{E}} \tag{15.7.2}$$

$$\chi^2 = \frac{(47-35.21)^2}{35.21} + \frac{(26-25.38)^2}{25.38} + \frac{(14-26.41)^2}{26.41} + \frac{(21-32.79)^2}{32.79} + \frac{(23-23.62)^2}{23.62} + \frac{(37-24.59)^2}{24.59}$$

Step 4: Make the Decision

The final decision for our test of independence is still based on our observed value (20.31) and our critical value (5.991). Because our observed value is greater than our critical value, we can reject the null hypothesis.

Reject H_0. Based on our data from 168 people, we can say that there is a statistically significant relation between whether or not someone watches college sports growing up and how much a college's sports team factor in to that person's decision on which college to attend, $\chi^2(2) = 20.31, p < 0.05$.

Effect Size for χ^2

Like all other significance tests, χ^2 tests – both goodness-of-fit and tests for independence – have effect sizes that can and should be calculated for statistically significant results. There are many options for which effect size to use, and the ultimate decision is based on the type of data, the structure of your frequency or contingency table, and the types of conclusions you would like to draw. For the purpose of our introductory course, we will focus only on a single effect size that is simple and flexible: Cramer's V.

Cramer's V is a type of correlation coefficient that can be computed on categorical data. Like any other correlation coefficient (e.g. Pearson's r), the cutoffs for small, medium, and large effect sizes of Cramer's V are 0.10, 0.30, and 0.50, respectively. The calculation of Cramer's V is very simple:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{15.7.3}$$

For this calculation, k is the smaller value of either R (the number of rows) or C (the number of columns). The numerator is simply the test statistic we calculate during step 3 of the hypothesis testing procedure. For our example, we had 2 rows and 3 columns, so $k = 2$:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{20.38}{168(2-1)}} = \sqrt{0.12} = 0.35$$

So the statistically significant relation between our variables was moderately strong.

This page titled 15.7: College Sports is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **14.7: College Sports** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

15.E: Chi-square (Exercises)

1. What does a frequency table display? What does a contingency table display?

Answer:

Frequency tables display observed category frequencies and (sometimes) expected category frequencies for a single categorical variable. Contingency tables display the frequency of observing people in crossed category levels for two categorical variables, and (sometimes) the marginal totals of each variable level.

2. What does a goodness-of-fit test assess?
3. How do expected frequencies relate to the null hypothesis?

Answer:

Expected values are what we would observe if the proportion of categories was completely random (i.e. no consistent difference other than chance), which is the same was what the null hypothesis predicts to be true.

4. What does a test-for-independence assess?
5. Compute the expected frequencies for the following contingency table:

	Category A	Category B
Category C	22	38
Category D	16	14

Answer:

Observed	Category A	Category B	Total
Category C	22	38	60
Category D	16	14	30
Total	38	52	90

Expected	Category A	Category B	Total
Category C	$((60*38)/90) = 25.33$	$((60*52)/90) = 34.67$	60
Category D	$((30*38)/90) = 12.67$	$((30*52)/90) = 17.33$	30
Total	38	52	90

6. Test significance and find effect sizes (if significant) for the following tests:
   1. $N = 19, R = 3, C = 2, \chi 2(2) = 7.89, \alpha = .05$
   2. $N = 12, R = 2, C = 2, \chi 2(1) = 3.12, \alpha = .05$
   3. $N = 74, R = 3, C = 3, \chi 2(4) = 28.41, \alpha = .01$
7. You hear a lot of people claim that The Empire Strikes Back is the best movie in the original Star Wars trilogy, and you decide to collect some data to demonstrate this empirically (pun intended). You ask 48 people which of the original movies they liked best; 8 said A New Hope was their favorite, 23 said The Empire Strikes Back was their favorite, and 17 said Return of the Jedi was their favorite. Perform a chi-square test on these data at the .05 level of significance.

Answer:

Step 1: H_0: "There is no difference in preference for one movie", H_A: "There is a difference in how many people prefer one movie over the others."

Step 2: 3 categories (columns) gives $df = 2$, $\chi^2_{\text{crit}} = 5.991$.

Step 3: Based on the given frequencies:

	New Hope	Empire	Jedi	Total
Observed	8	23	17	48
Expected	16	16	16	

$\chi^2 = 7.13$.

Step 4: Our obtained statistic is greater than our critical value, reject H_0. Based on our sample of 48 people, there is a statistically significant difference in the proportion of people who prefer one Star Wars movie over the others, $\chi^2(2) = 7.13$, $p < .05$. Since this is a statistically significant result, we should calculate an effect size: Cramer's $V = \sqrt{\frac{7.13}{48(3-1)}} = 0.27$, which is a moderate effect size.

8. A pizza company wants to know if people order the same number of different toppings. They look at how many pepperoni, sausage, and cheese pizzas were ordered in the last week; fill out the rest of the frequency table and test for a difference.

	Pepperoni	Sausage	Cheese	Total
Observed	320	275	251	
Expected				

9. A university administrator wants to know if there is a difference in proportions of students who go on to grad school across different majors. Use the data below to test whether there is a relation between college major and going to grad school.

		Major		
		Psychology	Business	Math
Graduate School	Yes	32	8	36
	No	15	41	12

Answer:

Step 1: H_0: "There is no relation between college major and going to grad school", H_A: "Going to grad school is related to college major."

Step 2: $df = 2$, $\chi^2_{\text{crit}} = 5.991$.

Step 3: Based on the given frequencies:

Expected Values		**Major**		
		Psychology	Business	Math
Graduate School	Yes	24.81	25.86	25.33
	No	22.19	23.14	22.67

$\chi^2 = 2.09 + 12.34 + 4.49 + 2.33 + 13.79 + 5.02 = 40.05$.

Step 4: Obtained statistic is greater than the critical value, reject H_0. Based on our data, there is a statistically significant relation between college major and going to grad school, $\chi^2(2) = 40.05, \text{p} < .05$, Cramer's $V = 0.53$, which is a large effect

10. A company you work for wants to make sure that they are not discriminating against anyone in their promotion process. You have been asked to look across gender to see if there are differences in promotion rate (i.e. if gender and promotion rate are independent or not). The following data should be assessed at the normal level of significance:

		Promoted in last two years?	
		Yes	No
Gender	Women	8	5
	Men	9	7

This page titled 15.E: Chi-square (Exercises) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Foster et al. (University of Missouri's Affordable and Open Access Educational Resources Initiative) via source content that was edited to the style and standards of the LibreTexts platform.

- **14.E: Chi-square (Exercises)** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

CHAPTER OVERVIEW

16: Appendix A- Statistical Tables

16.1: F-distribution (ANOVA distribution) table
16.2: z-table (aka Standard Normal Distribution Table)
16.3: t Distribution Table

16: Appendix A- Statistical Tables is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

16.1: F-distribution (ANOVA distribution) table

F-Distribution table for alpha=0.05

$df_1 = df_k$ (number of groups), $df_2 = df_N$ (total number of participants - number of groups)

/	df_1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60
df_2=1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	243.9060	245.9499	248.0131	249.0518	250.0951	251.1432	25
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	19.4125	19.4291	19.4458	19.4541	19.4624	19.4707	19
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446	8.7029	8.6602	8.6385	8.6166	8.5944	8.
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117	5.8578	5.8025	5.7744	5.7459	5.7170	5.
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777	4.6188	4.5581	4.5272	4.4957	4.4638	4.
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	3.9999	3.9381	3.8742	3.8415	3.8082	3.7743	3.
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	3.5107	3.4445	3.4105	3.3758	3.3404	3.
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839	3.2184	3.1503	3.1152	3.0794	3.0428	3.
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	3.0061	2.9365	2.9005	2.8637	2.8259	2.
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450	2.7740	2.7372	2.6996	2.6609	2.
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876	2.7186	2.6464	2.6090	2.5705	2.5309	2.
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6169	2.5436	2.5055	2.4663	2.4259	2.
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037	2.5331	2.4589	2.4202	2.3803	2.3392	2.
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	2.5342	2.4630	2.3879	2.3487	2.3082	2.2664	2.
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4034	2.3275	2.2878	2.2468	2.2043	2.
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.4247	2.3522	2.2756	2.2354	2.1938	2.1507	2.
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499	2.3807	2.3077	2.2304	2.1898	2.1477	2.1040	2.
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.3421	2.2686	2.1906	2.1497	2.1071	2.0629	2.
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.3080	2.2341	2.1555	2.1141	2.0712	2.0264	1.
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479	2.2776	2.2033	2.1242	2.0825	2.0391	1.9938	1.
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660	2.3210	2.2504	2.1757	2.0960	2.0540	2.0102	1.9645	1.
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.2258	2.1508	2.0707	2.0283	1.9842	1.9380	1.
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201	2.2747	2.2036	2.1282	2.0476	2.0050	1.9605	1.9139	1.
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547	2.1834	2.1077	2.0267	1.9838	1.9390	1.8920	1.
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649	2.0889	2.0075	1.9643	1.9192	1.8718	1.
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197	2.1479	2.0716	1.9898	1.9464	1.9010	1.8533	1.
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043	2.1323	2.0558	1.9736	1.9299	1.8842	1.8361	1.
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900	2.1179	2.0411	1.9586	1.9147	1.8687	1.8203	1.
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229	2.1768	2.1045	2.0275	1.9446	1.9005	1.8543	1.8055	1.
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921	2.0148	1.9317	1.8874	1.8409	1.7918	1.
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772	2.0035	1.9245	1.8389	1.7929	1.7444	1.6928	1.
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401	1.9926	1.9174	1.8364	1.7480	1.7001	1.6491	1.5943	1.
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588	1.9105	1.8337	1.7505	1.6587	1.6084	1.5543	1.4952	1.
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307	1.7522	1.6664	1.5705	1.5173	1.4591	1.3940	1.

Additional F-tables for Alpha values of 0.10, 0.025, and 0.01 can be found at the source for this table: http://socr.ucla.edu/Applets.dir/F_Table.html

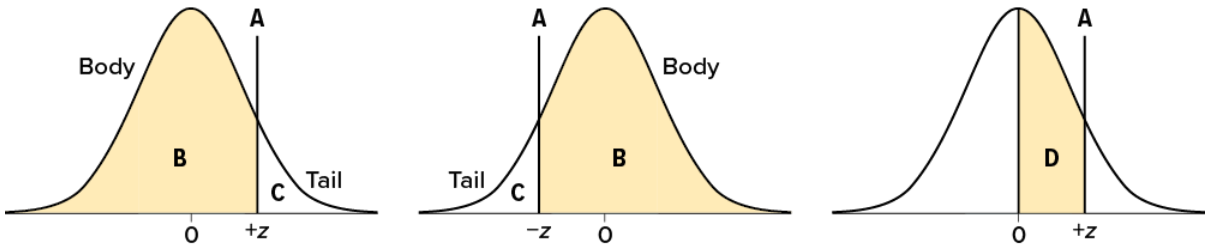
16.1: F-distribution (ANOVA distribution) table is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

16.2: z-table (aka Standard Normal Distribution Table)

- Page ID 49385

Standard Normal Distribution Table (z Table)

(“z Table Curves” by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)



(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
0.00	.5000	.5000	.0000
0.01	.5040	.4960	.0040
0.02	.5080	.4920	.0080
0.03	.5120	.4880	.0120
0.04	.5160	.4840	.0160
0.05	.5199	.4801	.0199
0.06	.5239	.4761	.0239
0.07	.5279	.4721	.0279
0.08	.5319	.4681	.0319
0.09	.5359	.4641	.0359
0.10	.5398	.4602	.0398
0.11	.5438	.4562	.0438
0.12	.5478	.4522	.0478
0.13	.5517	.4483	.0517
0.14	.5557	.4443	.0557
0.15	.5596	.4404	.0596
0.16	.5636	.4364	.0636
0.17	.5675	.4325	.0675
0.18	.5714	.4286	.0714
0.19	.5753	.4247	.0753
0.20	.5793	.4207	.0793
0.21	.5832	.4168	.0832
0.22	.5871	.4129	.0871

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
0.23	.5910	.4090	.0910
0.24	.5948	.4052	.0948
0.25	.5987	.4013	.0987
0.26	.6026	.3974	.1026
0.27	.6064	.3936	.1064
0.28	.6103	.3897	.1103
0.29	.6141	.3859	.1141
0.30	.6179	.3821	.1179
0.31	.6217	.3783	.1217
0.32	.6255	.3745	.1255
0.33	.6293	.3707	.1293
0.34	.6331	.3669	.1331
0.35	.6368	.3632	.1368
0.36	.6406	.3594	.1406
0.37	.6443	.3557	.1443
0.38	.6480	.3520	.1480
0.39	.6517	.3483	.1517
0.40	.6554	.3446	.1554
0.41	.6591	.3409	.1591
0.42	.6628	.3372	.1628
0.43	.6664	.3336	.1664
0.44	.6700	.3300	.1700
0.45	.6736	.3264	.1736
0.46	.6772	.3228	.1772
0.47	.6808	.3192	.1808
0.48	.6844	.3156	.1844
0.49	.6879	.3121	.1879
0.50	.6915	.3085	.1915

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
0.51	.6950	.3050	.1950
0.52	.6985	.3015	.1985
0.53	.7019	.2981	.2019
0.54	.7054	.2946	.2054
0.55	.7088	.2912	.2088
0.56	.7123	.2877	.2123
0.57	.7157	.2843	.2157
0.58	.7190	.2810	.2190
0.59	.7224	.2776	.2224
0.60	.7257	.2743	.2257
0.61	.7291	.2709	.2291
0.62	.7324	.2676	.2324
0.63	.7357	.2643	.2357
0.64	.7389	.2611	.2389
0.65	.7422	.2578	.2422
0.66	.7454	.2546	.2454
0.67	.7486	.2514	.2486
0.68	.7517	.2483	.2517
0.69	.7549	.2451	.2549
0.70	.7580	.2420	.2580
0.71	.7611	.2389	.2611
0.72	.7642	.2358	.2642
0.73	.7673	.2327	.2673
0.74	.7704	.2296	.2704
0.75	.7734	.2266	.2734
0.76	.7764	.2236	.2764
0.77	.7794	.2206	.2794
0.78	.7823	.2177	.2823

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
0.79	.7852	.2148	.2852
0.80	.7881	.2119	.2881
0.81	.7910	.2090	.2910
0.82	.7939	.2061	.2939
0.83	.7967	.2033	.2967
0.84	.7995	.2005	.2995
0.85	.8023	.1977	.3023
0.86	.8051	.1949	.3051
0.87	.8078	.1922	.3078
0.88	.8106	.1894	.3106
0.89	.8133	.1867	.3133
0.90	.8159	.1841	.3159
0.91	.8186	.1814	.3186
0.92	.8212	.1788	.3212
0.93	.8238	.1762	.3238
0.94	.8264	.1736	.3264
0.95	.8289	.1711	.3289
0.96	.8315	.1685	.3315
0.97	.8340	.1660	.3340
0.98	.8365	.1635	.3365
0.99	.8389	.1611	.3389
1.00	.8413	.1587	.3413
1.01	.8438	.1562	.3438
1.02	.8461	.1539	.3461
1.03	.8485	.1515	.3485
1.04	.8508	.1492	.3508
1.05	.8531	.1469	.3531
1.06	.8554	.1446	.3554

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
1.07	.8577	.1423	.3577
1.08	.8599	.1401	.3599
1.09	.8621	.1379	.3621
1.10	.8643	.1357	.3643
1.11	.8665	.1335	.3665
1.12	.8686	.1314	.3686
1.13	.8708	.1292	.3708
1.14	.8729	.1271	.3729
1.15	.8749	.1251	.3749
1.16	.8770	.1230	.3770
1.17	.8790	.1210	.3790
1.18	.8810	.1190	.3810
1.19	.8830	.1170	.3830
1.20	.8849	.1151	.3849
1.21	.8869	.1131	.3869
1.22	.8888	.1112	.3888
1.23	.8907	.1093	.3907
1.24	.8925	.1075	.3925
1.25	.8944	.1056	.3944
1.26	.8962	.1038	.3962
1.27	.8980	.1020	.3980
1.28	.8997	.1003	.3997
1.29	.9015	.0985	.4015
1.30	.9032	.0968	.4032
1.31	.9049	.0951	.4049
1.32	.9066	.0934	.4066
1.33	.9082	.0918	.4082
1.34	.9099	.0901	.4099

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
1.35	.9115	.0885	.4115
1.36	.9131	.0869	.4131
1.37	.9147	.0853	.4147
1.38	.9162	.0838	.4162
1.39	.9177	.0823	.4177
1.40	.9192	.0808	.4192
1.41	.9207	.0793	.4207
1.42	.9222	.0778	.4222
1.43	.9236	.0764	.4236
1.44	.9251	.0749	.4251
1.45	.9265	.0735	.4265
1.46	.9279	.0721	.4279
1.47	.9292	.0708	.4292
1.48	.9306	.0694	.4306
1.49	.9319	.0681	.4319
1.50	.9332	.0668	.4332
1.51	.9345	.0655	.4345
1.52	.9357	.0643	.4357
1.53	.9370	.0630	.4370
1.54	.9382	.0618	.4382
1.55	.9394	.0606	.4394
1.56	.9406	.0594	.4406
1.57	.9418	.0582	.4418
1.58	.9429	.0571	.4429
1.59	.9441	.0559	.4441
1.60	.9452	.0548	.4452
1.61	.9463	.0537	.4463
1.62	.9474	.0526	.4474

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
1.63	.9484	.0516	.4484
1.64	.9495	.0505	.4495
1.65	.9505	.0495	.4505
1.66	.9515	.0485	.4515
1.67	.9525	.0475	.4525
1.68	.9535	.0465	.4535
1.69	.9545	.0455	.4545
1.70	.9554	.0446	.4554
1.71	.9564	.0436	.4564
1.72	.9573	.0427	.4573
1.73	.9582	.0418	.4582
1.74	.9591	.0409	.4591
1.75	.9599	.0401	.4599
1.76	.9608	.0392	.4608
1.77	.9616	.0384	.4616
1.78	.9625	.0375	.4625
1.79	.9633	.0367	.4633
1.80	.9641	.0359	.4641
1.81	.9649	.0351	.4649
1.82	.9656	.0344	.4656
1.83	.9664	.0336	.4664
1.84	.9671	.0329	.4671
1.85	.9678	.0322	.4678
1.86	.9686	.0314	.4686
1.87	.9693	.0307	.4693
1.88	.9699	.0301	.4699
1.89	.9706	.0294	.4706
1.90	.9713	.0287	.4713

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
1.91	.9719	.0281	.4719
1.92	.9726	.0274	.4726
1.93	.9732	.0268	.4732
1.94	.9738	.0262	.4738
1.95	.9744	.0256	.4744
1.96	.9750	.0250	.4750
1.97	.9756	.0244	.4756
1.98	.9761	.0239	.4761
1.99	.9767	.0233	.4767
2.00	.9772	.0228	.4772
2.01	.9778	.0222	.4778
2.02	.9783	.0217	.4783
2.03	.9788	.0212	.4788
2.04	.9793	.0207	.4793
2.05	.9798	.0202	.4798
2.06	.9803	.0197	.4803
2.07	.9808	.0192	.4808
2.08	.9812	.0188	.4812
2.09	.9817	.0183	.4817
2.10	.9821	.0179	.4821
2.11	.9826	.0174	.4826
2.12	.9830	.0170	.4830
2.13	.9834	.0166	.4834
2.14	.9838	.0162	.4838
2.15	.9842	.0158	.4842
2.16	.9846	.0154	.4846
2.17	.9850	.0150	.4850
2.18	.9854	.0146	.4854

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
2.19	.9857	.0143	.4857
2.20	.9861	.0139	.4861
2.21	.9864	.0136	.4864
2.22	.9868	.0132	.4868
2.23	.9871	.0129	.4871
2.24	.9875	.0125	.4875
2.25	.9878	.0122	.4878
2.26	.9881	.0119	.4881
2.27	.9884	.0116	.4884
2.28	.9887	.0113	.4887
2.29	.9890	.0110	.4890
2.30	.9893	.0107	.4893
2.31	.9896	.0104	.4896
2.32	.9898	.0102	.4898
2.33	.9901	.0099	.4901
2.34	.9904	.0096	.4904
2.35	.9906	.0094	.4906
2.36	.9909	.0091	.4909
2.37	.9911	.0089	.4911
2.38	.9913	.0087	.4913
2.39	.9916	.0084	.4916
2.40	.9918	.0082	.4918
2.41	.9920	.0080	.4920
2.42	.9922	.0078	.4922
2.43	.9925	.0075	.4925
2.44	.9927	.0073	.4927
2.45	.9929	.0071	.4929
2.46	.9931	.0069	.4931

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
2.47	.9932	.0068	.4932
2.48	.9934	.0066	.4934
2.49	.9936	.0064	.4936
2.50	.9938	.0062	.4938
2.51	.9940	.0060	.4940
2.52	.9941	.0059	.4941
2.53	.9943	.0057	.4943
2.54	.9945	.0055	.4945
2.55	.9946	.0054	.4946
2.56	.9948	.0052	.4948
2.57	.9949	.0051	.4949
2.58	.9951	.0049	.4951
2.59	.9952	.0048	.4952
2.60	.9953	.0047	.4953
2.61	.9955	.0045	.4955
2.62	.9956	.0044	.4956
2.63	.9957	.0043	.4957
2.64	.9959	.0041	.4959
2.65	.9960	.0040	.4960
2.66	.9961	.0039	.4961
2.67	.9962	.0038	.4962
2.68	.9963	.0037	.4963
2.69	.9964	.0036	.4964
2.70	.9965	.0035	.4965
2.71	.9966	.0034	.4966
2.72	.9967	.0033	.4967
2.73	.9968	.0032	.4968
2.74	.9969	.0031	.4969

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
2.75	.9970	.0030	.4970
2.76	.9971	.0029	.4971
2.77	.9972	.0028	.4972
2.78	.9973	.0027	.4973
2.79	.9974	.0026	.4974
2.80	.9974	.0026	.4974
2.81	.9975	.0025	.4975
2.82	.9976	.0024	.4976
2.83	.9977	.0023	.4977
2.84	.9977	.0023	.4977
2.85	.9978	.0022	.4978
2.86	.9979	.0021	.4979
2.87	.9979	.0021	.4979
2.88	.9980	.0020	.4980
2.89	.9981	.0019	.4981
2.90	.9981	.0019	.4981
2.91	.9982	.0018	.4982
2.92	.9982	.0018	.4982
2.93	.9983	.0017	.4983
2.94	.9984	.0016	.4984
2.95	.9984	.0016	.4984
2.96	.9985	.0015	.4985
2.97	.9985	.0015	.4985
2.98	.9986	.0014	.4986
2.99	.9986	.0014	.4986
3.00	.9987	.0013	.4987
3.01	.9987	.0013	.4987
3.02	.9987	.0013	.4987

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
3.03	.9988	.0012	.4988
3.04	.9988	.0012	.4988
3.05	.9989	.0011	.4989
3.06	.9989	.0011	.4989
3.07	.9989	.0011	.4989
3.08	.9990	.0010	.4990
3.09	.9990	.0010	.4990
3.10	.9990	.0010	.4990
3.11	.9991	.0009	.4991
3.12	.9991	.0009	.4991
3.13	.9991	.0009	.4991
3.14	.9992	.0008	.4992
3.15	.9992	.0008	.4992
3.16	.9992	.0008	.4992
3.17	.9992	.0008	.4992
3.18	.9993	.0007	.4993
3.19	.9993	.0007	.4993
3.20	.9993	.0007	.4993
3.21	.9993	.0007	.4993
3.22	.9994	.0006	.4994
3.23	.9994	.0006	.4994
3.24	.9994	.0006	.4994
3.30	.9995	.0005	.4995
3.40	.9997	.0003	.4997
3.50	.9998	.0002	.4998
3.60	.9998	.0002	.4998
3.70	.9999	.0001	.4999
3.80	.99993	.00007	.49993

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion between Mean and z
3.90	.99995	.00005	.49995
4.00	.99997	.00003	.49997

This table taken from Cote et al, 2021, shared under a CC BY-NC-SA 4.0 license.

16.2: z-table (aka Standard Normal Distribution Table) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

16.3: t Distribution Table

t Distribution Table (t Table)

df	Proportion (a) in One tail								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Proportion (a) in Two tails combined								
	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.578
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.600
3	0.765	1.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	1.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725

df	Proportion (a) in One tail								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Proportion (a) in Two tails combined								
	.50	.40	.30	.20	.10	.05	.02	.01	.001
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.689
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.660
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.290

Adapted from "Tabla t" by Jsmura/Wikimedia Commons, CC BY-SA 4.0.

This table taken from Cote et al, 2021, shared under a CC BY-NC-SA 4.0 license.

16.3: t Distribution Table is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

Detailed Licensing

Overview

Title: PSYC 330: Statistics for the Behavioral Sciences with Dr. DeSouza

Webpages: 124

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 81.5% (101 pages)
- Undeclared: 17.7% (22 pages)
- CC BY-NC 4.0: 0.8% (1 page)

By Page

- PSYC 330: Statistics for the Behavioral Sciences with Dr. DeSouza - *CC BY-NC-SA 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - Foreword - *CC BY-NC-SA 4.0*
  - 1: What are Statistics? - *CC BY-NC-SA 4.0*
    - 1.1: What are statistics? - *CC BY-NC-SA 4.0*
    - 1.2: Why do we study statistics? - *CC BY-NC-SA 4.0*
  - 2: Types of Data and How to Collect Them (by Dr. Alisa Beyer) - *CC BY-NC-SA 4.0*
    - 2.1: Types of Data and How to Collect Them - *CC BY-NC-SA 4.0*
    - 2.2: Who are your participants? Who is your population? - *CC BY-NC-SA 4.0*
    - 2.3: Representative Sample - *CC BY-NC-SA 4.0*
    - 2.4: Type of Research Designs - *Undeclared*
    - 2.5: Working with Data - *Undeclared*
    - 2.6: Levels of Measurement - *Undeclared*
    - 2.7: What level of measurement is used for psychological variables? - *Undeclared*
    - 2.8: Reliability, Validity, and Results - *CC BY-NC 4.0*
    - 2.9: Types of Statistical Analyses - *Undeclared*
    - 2.10: Mathematical Notation - *Undeclared*
    - 2.E: Exercises - *Undeclared*
  - 3: Measures of Central Tendency and Spread - *CC BY-NC-SA 4.0*
    - 3.1: What is Central Tendency? - *CC BY-NC-SA 4.0*
    - 3.2: Measures of Central Tendency - *CC BY-NC-SA 4.0*
    - 3.3: Spread and Variability - *CC BY-NC-SA 4.0*
    - 3.E: Measures of Central Tendency and Spread (Exercises) - *CC BY-NC-SA 4.0*
  - 4: Describing Data using Distributions and Graphs - *CC BY-NC-SA 4.0*
    - 4.1: Graphing Qualitative Variables - *CC BY-NC-SA 4.0*
    - 4.2: Graphing Quantitative Variables - *CC BY-NC-SA 4.0*
    - 4.E: Describing Data using Distributions and Graphs (Exercises) - *CC BY-NC-SA 4.0*
  - 5: Z-scores and the Standard Normal Distribution - *CC BY-NC-SA 4.0*
    - 5.1: Normal Distributions - *CC BY-NC-SA 4.0*
    - 5.2: Z-scores - *CC BY-NC-SA 4.0*
    - 5.3: Z-scores and the Area under the Curve - *CC BY-NC-SA 4.0*
    - 5.E: Z-scores and the Standard Normal Distribution (Exercises) - *CC BY-NC-SA 4.0*
  - 6: Probability - *CC BY-NC-SA 4.0*
    - 6.1: What is Probability - *CC BY-NC-SA 4.0*
    - 6.2: Probability in Graphs and Distributions - *CC BY-NC-SA 4.0*
    - 6.3: The Bigger Picture - *CC BY-NC-SA 4.0*
    - 6.E: Probability (Exercises) - *CC BY-NC-SA 4.0*
  - 7: Sampling Distributions - *CC BY-NC-SA 4.0*
    - 7.1: People, Samples, and Populations - *CC BY-NC-SA 4.0*
    - 7.2: The Sampling Distribution of Sample Means - *CC BY-NC-SA 4.0*
    - 7.3: Sampling Distribution, Probability and Inference - *CC BY-NC-SA 4.0*
    - 7.E: Sampling Distributions (Exercises) - *CC BY-NC-SA 4.0*
  - 8: Introduction to Hypothesis Testing - *CC BY-NC-SA 4.0*

- 8.1: Logic and Purpose of Hypothesis Testing - *CC BY-NC-SA 4.0*
- 8.2: The Probability Value - *CC BY-NC-SA 4.0*
- 8.3: The Null Hypothesis - *CC BY-NC-SA 4.0*
- 8.4: The Alternative Hypothesis - *CC BY-NC-SA 4.0*
- 8.5: Critical values, p-values, and significance level - *CC BY-NC-SA 4.0*
- 8.6: Steps of the Hypothesis Testing Process - *CC BY-NC-SA 4.0*
- 8.7: Movie Popcorn - *CC BY-NC-SA 4.0*
- 8.8: Effect Size - *CC BY-NC-SA 4.0*
- 8.9: Office Temperature - *CC BY-NC-SA 4.0*
- 8.10: Different Significance Level - *CC BY-NC-SA 4.0*
- 8.11: Other Considerations in Hypothesis Testing - *CC BY-NC-SA 4.0*
- 8.E: Introduction to Hypothesis Testing (Exercises) - *CC BY-NC-SA 4.0*

- 9: Introduction to t-tests - *CC BY-NC-SA 4.0*
  - 9.1: The t-statistic - *CC BY-NC-SA 4.0*
  - 9.2: Hypothesis Testing with t - *CC BY-NC-SA 4.0*
  - 9.3: Confidence Intervals - *CC BY-NC-SA 4.0*
  - 9.E: Introduction to t-tests (Exercises) - *CC BY-NC-SA 4.0*
- 10: Repeated Measures - *CC BY-NC-SA 4.0*
  - 10.1: Change and Differences - *CC BY-NC-SA 4.0*
  - 10.2: Hypotheses of Change and Differences - *CC BY-NC-SA 4.0*
  - 10.3: Increasing Satisfaction at Work - *CC BY-NC-SA 4.0*
  - 10.4: Bad Press - *CC BY-NC-SA 4.0*
  - 10.E: Repeated Measures (Exercises) - *CC BY-NC-SA 4.0*
- 11: Independent Samples - *CC BY-NC-SA 4.0*
  - 11.1: Difference of Means - *CC BY-NC-SA 4.0*
  - 11.2: Research Questions about Independent Means - *CC BY-NC-SA 4.0*
  - 11.3: Hypotheses and Decision Criteria - *CC BY-NC-SA 4.0*
  - 11.4: Independent Samples t-statistic - *CC BY-NC-SA 4.0*
  - 11.5: Standard Error and Pooled Variance - *CC BY-NC-SA 4.0*
  - 11.6: Movies and Mood - *CC BY-NC-SA 4.0*
  - 11.7: Effect Sizes and Confidence Intervals - *CC BY-NC-SA 4.0*
  - 11.8: Homogeneity of Variance - *CC BY-NC-SA 4.0*
  - 11.E: Independent Samples (Exercises) - *CC BY-NC-SA 4.0*
- 12: Analysis of Variance - *CC BY-NC-SA 4.0*
  - 12.1: Observing and Interpreting Variability - *CC BY-NC-SA 4.0*
  - 12.2: Sources of Variance - *CC BY-NC-SA 4.0*
  - 12.3: ANOVA Table - *CC BY-NC-SA 4.0*
  - 12.4: ANOVA and Type I Error - *CC BY-NC-SA 4.0*
  - 12.5: Hypotheses in ANOVA - *CC BY-NC-SA 4.0*
  - 12.6: Scores on Job Application Tests - *CC BY-NC-SA 4.0*
  - 12.7: Variance Explained - *CC BY-NC-SA 4.0*
  - 12.8: Post Hoc Tests - *CC BY-NC-SA 4.0*
  - 12.9: Other ANOVA Designs - *CC BY-NC-SA 4.0*
  - 12.10: Analysis of Variance (Exercises) - *CC BY-NC-SA 4.0*
- 13: Two-Factor ANOVAs (by Dr. Alisa Beyer) - *CC BY-NC-SA 4.0*
  - 13.1: Two-Factor ANOVAs - *CC BY-NC-SA 4.0*
  - 13.2: Conducting a Two Factor ANOVA - *Undeclared*
  - 13.3: Graphing the Results of Factorial Experiments - *Undeclared*
  - 13.E: Exercises - *Undeclared*
- 14: Correlations - *CC BY-NC-SA 4.0*
  - 14.1: Variability and Covariance - *CC BY-NC-SA 4.0*
  - 14.2: Visualizing Relations - *CC BY-NC-SA 4.0*
  - 14.3: Three Characteristics - *CC BY-NC-SA 4.0*
  - 14.4: Pearson's r - *CC BY-NC-SA 4.0*
  - 14.5: Anxiety and Depression - *CC BY-NC-SA 4.0*
  - 14.6: Effect Size - *CC BY-NC-SA 4.0*
  - 14.7: Correlation versus Causation - *CC BY-NC-SA 4.0*
  - 14.8: Final Considerations - *CC BY-NC-SA 4.0*
  - 14.E: Correlations (Exercises) - *CC BY-NC-SA 4.0*
- 15: Chi-square - *CC BY-NC-SA 4.0*
  - 15.1: Categories and Frequency Tables - *CC BY-NC-SA 4.0*
  - 15.2: Goodness-of-Fit - *CC BY-NC-SA 4.0*
  - 15.3: χ^2 Statistic - *CC BY-NC-SA 4.0*
  - 15.4: Pineapple on Pizza - *CC BY-NC-SA 4.0*
  - 15.5: Contingency Tables for Two Variables - *CC BY-NC-SA 4.0*
  - 15.6: Test for Independence - *CC BY-NC-SA 4.0*
  - 15.7: College Sports - *CC BY-NC-SA 4.0*
  - 15.E: Chi-square (Exercises) - *CC BY-NC-SA 4.0*
- 16: Appendix A- Statistical Tables - *CC BY-NC-SA 4.0*
  - 16.1: F-distribution (ANOVA distribution) table - *Undeclared*
  - 16.2: z-table (aka Standard Normal Distribution Table) - *CC BY-NC-SA 4.0*

- 16.3: t Distribution Table - *Undeclared*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*

Detailed Licensing

Overview

Title: PSYC 330: Statistics for the Behavioral Sciences with Dr. DeSouza

Webpages: 124

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 81.5% (101 pages)
- Undeclared: 17.7% (22 pages)
- CC BY-NC 4.0: 0.8% (1 page)

By Page

- PSYC 330: Statistics for the Behavioral Sciences with Dr. DeSouza - *CC BY-NC-SA 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - Foreword - *CC BY-NC-SA 4.0*
  - 1: What are Statistics? - *CC BY-NC-SA 4.0*
    - 1.1: What are statistics? - *CC BY-NC-SA 4.0*
    - 1.2: Why do we study statistics? - *CC BY-NC-SA 4.0*
  - 2: Types of Data and How to Collect Them (by Dr. Alisa Beyer) - *CC BY-NC-SA 4.0*
    - 2.1: Types of Data and How to Collect Them - *CC BY-NC-SA 4.0*
    - 2.2: Who are your participants? Who is your population? - *CC BY-NC-SA 4.0*
    - 2.3: Representative Sample - *CC BY-NC-SA 4.0*
    - 2.4: Type of Research Designs - *Undeclared*
    - 2.5: Working with Data - *Undeclared*
    - 2.6: Levels of Measurement - *Undeclared*
    - 2.7: What level of measurement is used for psychological variables? - *Undeclared*
    - 2.8: Reliability, Validity, and Results - *CC BY-NC 4.0*
    - 2.9: Types of Statistical Analyses - *Undeclared*
    - 2.10: Mathematical Notation - *Undeclared*
    - 2.E: Exercises - *Undeclared*
  - 3: Measures of Central Tendency and Spread - *CC BY-NC-SA 4.0*
    - 3.1: What is Central Tendency? - *CC BY-NC-SA 4.0*
    - 3.2: Measures of Central Tendency - *CC BY-NC-SA 4.0*
    - 3.3: Spread and Variability - *CC BY-NC-SA 4.0*
    - 3.E: Measures of Central Tendency and Spread (Exercises) - *CC BY-NC-SA 4.0*
  - 4: Describing Data using Distributions and Graphs - *CC BY-NC-SA 4.0*
    - 4.1: Graphing Qualitative Variables - *CC BY-NC-SA 4.0*
    - 4.2: Graphing Quantitative Variables - *CC BY-NC-SA 4.0*
    - 4.E: Describing Data using Distributions and Graphs (Exercises) - *CC BY-NC-SA 4.0*
  - 5: Z-scores and the Standard Normal Distribution - *CC BY-NC-SA 4.0*
    - 5.1: Normal Distributions - *CC BY-NC-SA 4.0*
    - 5.2: Z-scores - *CC BY-NC-SA 4.0*
    - 5.3: Z-scores and the Area under the Curve - *CC BY-NC-SA 4.0*
    - 5.E: Z-scores and the Standard Normal Distribution (Exercises) - *CC BY-NC-SA 4.0*
  - 6: Probability - *CC BY-NC-SA 4.0*
    - 6.1: What is Probability - *CC BY-NC-SA 4.0*
    - 6.2: Probability in Graphs and Distributions - *CC BY-NC-SA 4.0*
    - 6.3: The Bigger Picture - *CC BY-NC-SA 4.0*
    - 6.E: Probability (Exercises) - *CC BY-NC-SA 4.0*
  - 7: Sampling Distributions - *CC BY-NC-SA 4.0*
    - 7.1: People, Samples, and Populations - *CC BY-NC-SA 4.0*
    - 7.2: The Sampling Distribution of Sample Means - *CC BY-NC-SA 4.0*
    - 7.3: Sampling Distribution, Probability and Inference - *CC BY-NC-SA 4.0*
    - 7.E: Sampling Distributions (Exercises) - *CC BY-NC-SA 4.0*
  - 8: Introduction to Hypothesis Testing - *CC BY-NC-SA 4.0*

- 8.1: Logic and Purpose of Hypothesis Testing - *CC BY-NC-SA 4.0*
- 8.2: The Probability Value - *CC BY-NC-SA 4.0*
- 8.3: The Null Hypothesis - *CC BY-NC-SA 4.0*
- 8.4: The Alternative Hypothesis - *CC BY-NC-SA 4.0*
- 8.5: Critical values, p-values, and significance level - *CC BY-NC-SA 4.0*
- 8.6: Steps of the Hypothesis Testing Process - *CC BY-NC-SA 4.0*
- 8.7: Movie Popcorn - *CC BY-NC-SA 4.0*
- 8.8: Effect Size - *CC BY-NC-SA 4.0*
- 8.9: Office Temperature - *CC BY-NC-SA 4.0*
- 8.10: Different Significance Level - *CC BY-NC-SA 4.0*
- 8.11: Other Considerations in Hypothesis Testing - *CC BY-NC-SA 4.0*
- 8.E: Introduction to Hypothesis Testing (Exercises) - *CC BY-NC-SA 4.0*

- 9: Introduction to t-tests - *CC BY-NC-SA 4.0*
  - 9.1: The t-statistic - *CC BY-NC-SA 4.0*
  - 9.2: Hypothesis Testing with t - *CC BY-NC-SA 4.0*
  - 9.3: Confidence Intervals - *CC BY-NC-SA 4.0*
  - 9.E: Introduction to t-tests (Exercises) - *CC BY-NC-SA 4.0*
- 10: Repeated Measures - *CC BY-NC-SA 4.0*
  - 10.1: Change and Differences - *CC BY-NC-SA 4.0*
  - 10.2: Hypotheses of Change and Differences - *CC BY-NC-SA 4.0*
  - 10.3: Increasing Satisfaction at Work - *CC BY-NC-SA 4.0*
  - 10.4: Bad Press - *CC BY-NC-SA 4.0*
  - 10.E: Repeated Measures (Exercises) - *CC BY-NC-SA 4.0*
- 11: Independent Samples - *CC BY-NC-SA 4.0*
  - 11.1: Difference of Means - *CC BY-NC-SA 4.0*
  - 11.2: Research Questions about Independent Means - *CC BY-NC-SA 4.0*
  - 11.3: Hypotheses and Decision Criteria - *CC BY-NC-SA 4.0*
  - 11.4: Independent Samples t-statistic - *CC BY-NC-SA 4.0*
  - 11.5: Standard Error and Pooled Variance - *CC BY-NC-SA 4.0*
  - 11.6: Movies and Mood - *CC BY-NC-SA 4.0*
  - 11.7: Effect Sizes and Confidence Intervals - *CC BY-NC-SA 4.0*
  - 11.8: Homogeneity of Variance - *CC BY-NC-SA 4.0*
  - 11.E: Independent Samples (Exercises) - *CC BY-NC-SA 4.0*
- 12: Analysis of Variance - *CC BY-NC-SA 4.0*
  - 12.1: Observing and Interpreting Variability - *CC BY-NC-SA 4.0*
  - 12.2: Sources of Variance - *CC BY-NC-SA 4.0*
  - 12.3: ANOVA Table - *CC BY-NC-SA 4.0*
  - 12.4: ANOVA and Type I Error - *CC BY-NC-SA 4.0*
  - 12.5: Hypotheses in ANOVA - *CC BY-NC-SA 4.0*
  - 12.6: Scores on Job Application Tests - *CC BY-NC-SA 4.0*
  - 12.7: Variance Explained - *CC BY-NC-SA 4.0*
  - 12.8: Post Hoc Tests - *CC BY-NC-SA 4.0*
  - 12.9: Other ANOVA Designs - *CC BY-NC-SA 4.0*
  - 12.10: Analysis of Variance (Exercises) - *CC BY-NC-SA 4.0*
- 13: Two-Factor ANOVAs (by Dr. Alisa Beyer) - *CC BY-NC-SA 4.0*
  - 13.1: Two-Factor ANOVAs - *CC BY-NC-SA 4.0*
  - 13.2: Conducting a Two Factor ANOVA - *Undeclared*
  - 13.3: Graphing the Results of Factorial Experiments - *Undeclared*
  - 13.E: Exercises - *Undeclared*
- 14: Correlations - *CC BY-NC-SA 4.0*
  - 14.1: Variability and Covariance - *CC BY-NC-SA 4.0*
  - 14.2: Visualizing Relations - *CC BY-NC-SA 4.0*
  - 14.3: Three Characteristics - *CC BY-NC-SA 4.0*
  - 14.4: Pearson's r - *CC BY-NC-SA 4.0*
  - 14.5: Anxiety and Depression - *CC BY-NC-SA 4.0*
  - 14.6: Effect Size - *CC BY-NC-SA 4.0*
  - 14.7: Correlation versus Causation - *CC BY-NC-SA 4.0*
  - 14.8: Final Considerations - *CC BY-NC-SA 4.0*
  - 14.E: Correlations (Exercises) - *CC BY-NC-SA 4.0*
- 15: Chi-square - *CC BY-NC-SA 4.0*
  - 15.1: Categories and Frequency Tables - *CC BY-NC-SA 4.0*
  - 15.2: Goodness-of-Fit - *CC BY-NC-SA 4.0*
  - 15.3: χ^2 Statistic - *CC BY-NC-SA 4.0*
  - 15.4: Pineapple on Pizza - *CC BY-NC-SA 4.0*
  - 15.5: Contingency Tables for Two Variables - *CC BY-NC-SA 4.0*
  - 15.6: Test for Independence - *CC BY-NC-SA 4.0*
  - 15.7: College Sports - *CC BY-NC-SA 4.0*
  - 15.E: Chi-square (Exercises) - *CC BY-NC-SA 4.0*
- 16: Appendix A- Statistical Tables - *CC BY-NC-SA 4.0*
  - 16.1: F-distribution (ANOVA distribution) table - *Undeclared*
  - 16.2: z-table (aka Standard Normal Distribution Table) - *CC BY-NC-SA 4.0*

- 16.3: t Distribution Table - *Undeclared*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*