INTRO TO STATS IN THE SOCIAL SCIENCES (MPC)



Adrianne Kotecki
Monterey Peninsula College



Intro to Stats in the Social Sciences

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 04/16/2026

TABLE OF CONTENTS

Licensing

Foreword

1: An Introduction to Statistics

- 1.1: What are statistics?
- 1.2: Why do we study statistics?
- 1.3: Statistics in Psychology
- 1.4: Putting It Together- Types of Statistical Studies and Producing Data
- 1.5: Introduction to Psychological Measurement
- 1.6: Types of Data and How to Collect Them
- 1.7: Statistics and Science
- 1.8: How to Succeed in Statistics
- 1.9: Math Refresher

2: Describing Data using Distributions and Graphs

- 2.1: Introduction to Looking at Data (This is what too many numbers looks like)
- 2.2: Frequency Tables
- 2.3: Graphing Qualitative Variables- Bar Graphs
- 2.4: Graphing Qualitative Variables- Pie Charts
- 2.5: Graphing Quantitative Variables
- 2.6: Skew and Kurtosis
- 2.7: Graphing Quantitative Data- Line Graphs
- 2.8: Graphing Quantitative Data- Histograms
- 2.9: Graphing Quantitative Data- Boxplots
- 2.10: Graphing Quantitative Data- Scatterplots
- 2.11: Summary and Some Honesty

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: Z-scores and the Standard Normal Distribution

- 4.1: Normal Distributions
- 4.2: Introduction to z-scores
- 4.3: Calculating z-scores
- 4.4: Introduction to the z table
- 4.5: Z table of critical values
- 4.6: Predicting Amounts
- 4.7: Summary of z Scores
- 4.8: The Write-Up

5: Distributions

- 5.1: Introduction to Distributions
- 5.2: What is Probability
- 5.3: The Law of Large Numbers
- 5.4: Normal Distributions and Probability Distributions
- 5.5: Putting it All Together
- 5.6: Summary- The Bigger Picture

6: Inferential Statistics and an Introduction to Hypothesis Testing

- 6.1: Growth Mindset
- 6.2: People, Samples, and Populations
- 6.3: The Sampling Distribution of Sample Means
- 6.4: Using Standard Error for Probability
- 6.5: Sampling Distribution, Probability and Inference
- 6.6: Logic and Purpose of Hypothesis Testing
- 6.7: The Probability Value
- 6.8: Descriptive versus Inferential Statistics
- 6.9: The Null Hypothesis
- 6.10: The Alternative Hypothesis
- 6.11: Null Hypothesis Significance Testing
- 6.12: Critical Values, p-values, and Significance
  - 6.12.1: Critical Values
  - 6.12.2: Summary of p-values and NHST
- 6.13: Steps of the Hypothesis Testing Process
- 6.14: Movie Popcorn
- 6.15: Effect Size
- 6.16: Office Temperature
- 6.17: The Two Errors in Null Hypothesis Significance Testing
  - 6.17.1: Power and Sample Size
  - 6.17.2: The p-value of a Test

7: Introducing t-tests (One-sample designs)

- 7.1: Introduction to One-Sample t-tests
- 7.2: Hypothesis Testing with t
- 7.3: Confidence Intervals
- 7.4: t table of critical values
- 7.5: Reporting Results
  - 7.5.1: Descriptive and Inferential Calculations and Conclusion Example

8: Two Sample Designs

- 8.1: Repeated Measures
  - 8.1.1: Change and Differences
  - 8.1.2: Hypotheses of Change and Differences
  - 8.1.3: Increasing Satisfaction at Work
  - 8.1.4: Bad Press
- 8.2: Independent Samples
  - 8.2.1: Difference of Means

- 8.2.2: Research Questions about Independent Means
- 8.2.3: Hypotheses and Decision Criteria
- 8.2.4: Independent Samples t-statistic
- 8.2.5: Standard Error and Pooled Variance
- 8.2.6: Movies and Mood
- 8.2.7: Effect Sizes and Confidence Intervals
- 8.2.8: Homogeneity of Variance

9: Analysis of Variance- Independent Samples

- 9.1: Why ANOVA?
  - 9.1.1: Observing and Interpreting Variability
  - 9.1.2: Ratio of Variability
- 9.2: Introduction to ANOVA's Sum of Squares
  - 9.2.1: Summary of ANOVA Summary Table
- 9.3: Hypotheses in ANOVA
- 9.4: Practice with Job Applicants
  - 9.4.1: Table of Critical F-Scores
- 9.5: Introduction to Pairwise Comparisons
  - 9.5.1: Pairwise Comparison Post Hoc Tests for Critical Values of Mean Differences
- 9.6: Practice on Mindset Data
- 9.7: On the Relationship Between ANOVA and the Student t Test
- 9.8: Non-Parametric Analysis Between Multiple Groups

10: Analysis of Variance- Repeated Measures

- 10.1: Introduction to Repeated Measures ANOVA
  - 10.1.1: Things Worth Knowing About RM ANOVAs
- 10.2: ANOVA Summary Table
  - 10.2.1: Repeated Measures ANOVA Sum of Squares Formulas
- 10.3: Practice with RM ANOVA Summary Table
  - 10.3.1: Practice with Mindset
- 10.4: Non-Parametric RM ANOVA

11: Factorial ANOVA (Two-Way)

- 11.1: Introduction to Factorial Designs
  - 11.1.1: Factorial Notations and Square Tables
- 11.2: Introduction to Main Effects and Interactions
  - 11.2.1: Example with Main Effects and Interactions
  - 11.2.2: Graphing Main Effects and Interactions
  - 11.2.3: Interpreting Main Effects and Interactions in Graphs
  - 11.2.4: Interpreting Interactions- Do Main Effects Matter?
  - 11.2.5: Interpreting Beyond 2x2 in Graphs
- 11.3: Two-Way ANOVA Summary Table
  - 11.3.1: Calculating Sum of Squares for the Factorial ANOVA Summary Table
- 11.4: When Should You Conduct Post-Hoc Pairwise Comparisons?
- 11.5: Practice with a 2x2 Factorial Design- Attention

- 11.5.1: Practice 2x3 Factorial ANOVA on Mindset
- 11.6: Choosing the Correct Analysis- Mean Comparison Edition

12: Correlation (and a little about Regression)

- 12.1: If something caused something else to change, what would that look like?
- 12.2: Pearson's r
- 12.3: Turning the numbers into a measure of co-variance
- 12.4: Examples with Data
- 12.5: Regression — A mini intro
- 12.6: Interpreting Correlations
- 12.7: Summary
- 12.8: References

13: Chi-Square

- 13.1: Introduction to Chi-Square
  - 13.1.1: Assumptions of the Test(s)
- 13.2: Introduction to Goodness-of-Fit Chi-Square
  - 13.2.1: Critical Values of Chi-Square Table
  - 13.2.2: Interpretation of the Chi-Square Goodness-of-Fit Test
- 13.3: Goodness of Fit χ^2 Formula
- 13.4: Practice Goodness of Fit- Pineapple on Pizza
- 13.5: Introduction to Test of Independence
- 13.6: Practice Chi-Square Test of Independence- College Sports
  - 13.6.1: Practice- Fast Food Meals
- 13.7: RM Chi-Square- The McNemar Test
- 13.8: Choosing the Correct Test- Chi-Square Edition

14: Wrapping it Up

- 14.1: Introduction to Wrapping Up
- 14.2: Choosing the Test
- 14.3: Why did you take this class?

Index

Glossary

Detailed Licensing

Detailed Licensing

Licensing

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

Foreword

Below you'll find the original Foreword from Dr. Garrett Foster, one of the original authors to a book that I drew from heavily in creating this text. LibreText and open source textbooks, have allowed me to customize Dr. Foster's text, bringing in other content from other authors, primarily Dr. Michelle Oja, to create a text specifically for our class.

~Prof. Kotecki

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

Oja, Michelle. (2022). *Behavioral Statistics*. Libretexts. https://stats.libretexts.org/Courses...Sciences_(Oja)

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: An Introduction to Statistics

1.1: What are statistics?
1.2: Why do we study statistics?
1.3: Statistics in Psychology
1.4: Putting It Together- Types of Statistical Studies and Producing Data
1.5: Introduction to Psychological Measurement
1.6: Types of Data and How to Collect Them
1.7: Statistics and Science
1.8: How to Succeed in Statistics
1.9: Math Refresher

1: An Introduction to Statistics is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

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.

1.3: Statistics in Psychology

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

does psychology have so much statistics?

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

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

Can't someone else do the statistics?

To some extent, but not completely. It's true that you don't need to become a fully trained statistician just to do psychology, but you do need to reach a certain level of statistical competence. In my view, there's three reasons that every psychological researcher ought to be able to do basic statistics:

- Firstly, there's the fundamental reason: statistics is deeply intertwined with research design. If you want to be good at designing psychological studies, you need to at least understand the basics of stats.
- Secondly, if you want to be good at the psychological side of the research, then you need to be able to understand the psychological literature, right? But almost every paper in the psychological literature reports the results of statistical analyses. So if you really want to understand the psychology, you need to be able to understand what other people did with their data. And that means understanding a certain amount of statistics.
- Thirdly, there's a big practical problem with being dependent on other people to do all your statistics: statistical analysis is *expensive*. If you ever get bored and want to look up how much the Australian government charges for university fees, you'll notice something interesting: statistics is designated as a "national priority" category, and so the fees are much, much lower than for any other area of study. This is because there's a massive shortage of statisticians out there. So, from your perspective as a psychological researcher, the laws of supply and demand aren't exactly on your side here! As a result, in almost any real life situation where you want to do psychological research, the cruel facts will be that you don't have enough money to afford a statistician. So the economics of the situation mean that you have to be pretty self-sufficient.

Note that a lot of these reasons generalise beyond researchers. If you want to be a practicing psychologist and stay on top of the field, it helps to be able to read the scientific literature, which relies pretty heavily on statistics.

don't care about jobs, research, or clinical work. Do I need statistics?

Okay, now you're just messing with me. Still, I think it should matter to you too. Statistics should matter to you in the same way that statistics should matter to *everyone*: we live in the 21st century, and data are *everywhere*. Frankly, given the world in which we live these days, a basic knowledge of statistics is pretty damn close to a survival tool! Which is the topic of the next section…

This page titled 1.3: Statistics in Psychology is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Danielle Navarro via source content that was edited to the style and standards of the LibreTexts platform.

- **1.3: Statistics in Psychology** by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.

1.4: Putting It Together- Types of Statistical Studies and Producing Data

Let's Summarize

- There are four steps in a statistical investigation:
  - Ask a question that can be answered by collecting data.
  - Decide what to measure, and then collect data.
  - Summarize and analyze.
  - Draw a conclusion, and communicate the results.
- There are two types of statistical studies:
  - Observational studies: An *observational study* observes individuals and measures variables of interest. We conduct observational studies to investigate questions about a population or about an association between two variables. An observational study alone does not provide convincing evidence of a cause-and-effect relationship.
  - Experiments: An *experiment* intentionally manipulates one variable in an attempt to cause an effect on another variable. The primary goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables.
- In statistics, a *variable* is information we gather about individuals or objects.

Observational Studies

- In an observational study, we draw a conclusion about the population on the basis of a sample. To draw a valid conclusion, the sample must be representative of the population. A representative sample is a subset of the population. It also reflects the characteristics of the population.
- A sample is biased if it systematically favors a certain outcome. Voluntary response samples (such as Internet polls) and convenience samples (such as surveys at a mall) are biased.
- Random selection eliminates bias. In a simple random sample, everyone in the population has an equal chance of being chosen. In this way, random selection helps ensure that the sample is representative of the population.
- Larger samples tend to be more accurate than smaller samples if the samples are chosen randomly. The size of the population does not affect the accuracy of a random sample as long as the population is large.
- If an attempt is made to include every individual from a population in a sample, then the investigation is called a *census.*

Experiments

The goal of the experiment is to provide evidence for a cause-and-effect relationship between two variables. When we investigate a relationship between two variables, we identify an explanatory variable and a response variable. To establish a cause-and-effect relationship, we want to make sure the explanatory variable is the only factor that impacts the response variable. But other factors, called *confounding variables*, may also influence the response.

- A well-designed experiment takes steps to eliminate the effects of confounding variables. These steps include direct control, random assignment of people to treatment groups, use of a control or placebo, and blind conditions. Incorporating such precautions, a well-designed experiment provides convincing evidence of cause-and-effect.
- Random assignment uses random chance to assign participants to treatments, which creates similar treatment groups. With random assignment, we can be fairly confident that any differences we observe in the response of treatment groups is due to the explanatory variable. In this way, we have evidence for a cause-and-effect relationship.
- A well-designed experiment provides evidence for a cause-and-effect relationship. But even in a well-designed experiment, differences in the response might be due to chance. We learn to describe chance behavior when we study probability later in the course.

CC licensed content, Shared previously

- Concepts in Statistics. **Provided by**: Open Learning Initiative. **Located at**: http://oli.cmu.edu. **License**: *CC BY: Attribution*

This page titled 1.4: Putting It Together- Types of Statistical Studies and Producing Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Bill Pelz via source content that was edited to the style and standards of the LibreTexts platform.

1.5: Introduction to Psychological Measurement

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

Some thoughts about psychological measurement

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

- My **age** is *33 years*.
- I *do not* **like anchovies**.
- My **chromosomal gender** is *male*.
- My **self-identified gender** is *male*.[7]

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

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

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

Looking at this a bit more closely, you might also realise that the concept of "age" isn't actually all that precise. In general, when we say "age" we implicitly mean "the length of time since birth". But that's not always the right way to do it. Suppose you're interested in how newborn babies control their eye movements. If you're interested in kids that young, you might also start to worry that "birth" is not the only meaningful point in time to care about. If Baby Alice is born 3 weeks premature and Baby Bianca is born 1 week late, would it really make sense to say that they are the "same age" if we encountered them "2 hours after birth"? In one sense, yes: by social convention, we use birth as our reference point for talking about age in everyday life, since it defines the amount of time the person has been operating as an independent entity in the world, but from a scientific perspective that's not the only thing we care about. When we think about the biology of human beings, it's often useful to think of ourselves as organisms that have been growing and maturing since conception, and from that perspective Alice and Bianca aren't the same age at all. So you might want to define the concept of "age" in two different ways: the length of time since conception, and the length of time since birth. When dealing with adults, it won't make much difference, but when dealing with newborns it might.

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

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

- You could look up official records, like birth certificates. This is time consuming and annoying, but it has its uses (e.g., if the person is now dead).

Operationalisation: defining your measurement

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

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

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

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

- ***A theoretical construct***. This is the thing that you're trying to take a measurement of, like "age", "gender" or an "opinion". A theoretical construct can't be directly observed, and often they're actually a bit vague.
- ***A measure***. The measure refers to the method or the tool that you use to make your observations. A question in a survey, a behavioural observation or a brain scan could all count as a measure.
- ***An operationalisation***. The term "operationalisation" refers to the logical connection between the measure and the theoretical construct, or to the process by which we try to derive a measure from a theoretical construct.
- ***A variable***. Finally, a new term. A variable is what we end up with when we apply our measure to something in the world. That is, variables are the actual "data" that we end up with in our data sets.

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

This page titled 1.5: Introduction to Psychological Measurement is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Danielle Navarro via source content that was edited to the style and standards of the LibreTexts platform.

- **2.1: Introduction to Psychological Measurement** by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.

1.6: 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.

1.6: 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.

1.7: Statistics and Science

Statistics are important to many fields, including social, behavioral, and health sciences such as psychology, sociology, nursing, and education to name just a few. They are also used in many areas of the physical sciences. This is because of their important position in the scientific method.

What distinguishes sciences from other fields is their empirical nature and use of the scientific method. To be empirical is to be evidence-based or evidence-focused. Statistics is an empirical field focused on the use of data which is often used as an important step in the scientific method. The scientific method is a method of inquiry in which questions, assumptions, and/or hypotheses are tested through the collection and analysis of data before conclusions are drawn. In keeping, the use of statistics is an important part of research in a diversity of fields.

Statistics is so central to the field of psychology, for example, that the standard format for research articles in psychology includes a section detailing from whom and how data were collected, followed by a section summarizing how the data were analyzed and what the results of the analyses were, and, finally, a section discussing the implications of the findings. These sections are titled as Method, Results, and Discussion, respectively, when following the recommended format known as American Psychological Association (APA) format. The focus of this book is the analysis step of research and how it is applied. Therefore, a review of the proper way to report results in APA-format will be included in several chapters throughout this book. This will provide us some of the necessary skills to read the results from, and participate in the process of, science.

Reading Review 1.3

1. When inferential statistics are used, the goal is to _____ results from samples to the populations from which they were drawn.
2. _____ statistics include calculations such as the mean, median, and mode to summarize data from samples univariately.

This page titled 1.7: Statistics and Science is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by .

- **1.5: Statistics and Science** is licensed CC BY-NC-SA 4.0.

1.8: How to Succeed in Statistics

Statistics can feel challenging at times, especially for those who are new to it. However, there are some things you should keep in mind that can help you.

First, focus on organization. There will be lots of numbers, qualitative responses, and symbols throughout your work in statistics which can get confusing and, at times, make it hard for you to identify the exact piece of information you are looking for. However, if you carefully organize the information and steps, it will make finding the information more efficient and less stressful. Imagine that there are 1,000 books about statistics all mixed up in a pile and you have to find one called “Why I Love Statistics” by Dr. Peter to do a book report. Looking for that book in the mess can be overwhelming and you might accidentally grab “Why I Loathe Statistics” by Dr. Pester instead to use for your book report. This is like having disorganized data and notes. Now imagine that instead of a pile, there are 1,000 books about statistics all nicely organized on shelves by title. You can quickly get to the book by Dr. Peter you want and even see it next to the similar book you don’t want by Dr. Pester. You can see clearly which one is the better match to what you are looking for and confidently get right to work on your book report. Let’s take this analogy a step further. Imagine you read the book and put it back in its appropriate place on the organized shelf. Then, three months later, you find out you are going to have a quiz on some topics from that book so you want to review it again. You can quickly locate it again and start refreshing your memory. Having organized and clear data and notes will benefit you as you progress through statistics. It will make it easier to follow your work and find what you need in the moment and be able to quickly return to the right information later when it is time to review.

Second, remember that the order of operations is an important process that is here to make our work easier. We will use order of operations in many places in this book. Just as organization of data and notes helps us, order of operations is a formal way of specifying the order in which the steps of formulas must be completed. This ensures that we always know what to do next and don’t get stymied in the middle of the steps to a complex formula. For a review of the order of operations, see Appendix B).

Third, go back to your foundational math skills when you get confused, frustrated, or overwhelmed. The level of statistics covered in this book uses combinations of only six basic mathematical operations: adding, subtracting, multiplying, dividing, squaring, and square rooting. That’s it. Though we will encounter new symbols for these operations and some formulas which can look quite complex, they are all just versions of the same six basic mathematical operations. For a review basic mathematical operations, see Appendix A). Fourth, consider what statistics can offer you. Statistics is an applied field focused on understanding aspects of the real world. It is used in many disciplines and careers so nearly everyone can find applicability to their academic and occupational pursuits. There are several fields and careers where the use of statistics is obvious such as mathematics, economics, biostatistics, actuarial science, and research more broadly. In addition, health statistics are important to nurses and doctors, test theory is important to educators, developmental modeling is important to child and adolescent development professionals, experimental analytics are important to the sciences, business, health, and education, and knowing how to read and explain statistics is important for journalists, meteorologists, and financial planners. In addition, several roles in the technology and engineering sectors such as user experience (UX) researcher, data scientist, data engineer, machine learning specialist, and engineering statistician require knowledge in statistics. Further, those who pursue careers as content creators are now provided with a variety of analytics that they can learn to use to potentially enhance their success. Thus, statistical skills have broad (and potentially highly valuable) application.

Fifth, statistics can help us to circumvent our biases and make better decisions. Statistics are all around us. They are used to make decisions about us but can also be harnessed to improve decisions made by us. Each of us has information available to us that is at least partially limited to or by our experiences. This in itself is not a problem; it is simply a description of an aspect of human reality. Unfortunately, however, our limited experiences can lead us to biased conclusions and poorer decision-making. Data which capture experiences beyond our own can allow us to see a bigger picture and statistics can allow us to see patterns that aren’t readily apparent through our day-to-day experiences. Thus, when used properly and thoughtfully, statistics can help us circumvent our own biases, become more knowledgeable about our worlds, and become better-informed decision makers.

Sixth, and related, take clear notes including examples to help you as you study statistics. One of the beauties of working in statistics is that we can use real world examples to help us value and think through problems. Thinking about real-world topics or questions can help you understand what the numbers represent and why the formulas include the steps that they do. This can keep you connected to the “why” of statistics at all levels: Why measure the data this way? Why do this specific mathematical

operation? Why analyze the data in this way? Why are these patterns appearing in the data and what does this tell us about the world? Why bother with statistics? We will begin to answer these questions in this book.

This page titled 1.8: How to Succeed in Statistics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by .

- **1.6: How to Succeed in Statistics** is licensed CC BY-NC-SA 4.0.

1.9: Math Refresher

Statistics is not math. It does, however, use math as a tool. This section will review some of the tools that you'll use in statistical equations, with an ending section on calculating percentages. You learned much of this in junior high or early high school, but that was a long time ago for most of us!

Order of Operations

Example 1.9.1

What does “Please Excuse My Dear Aunt Sally” or “PEMDAS” mean?

Solution

It's the order in which you should do each mathematical operation.

“Please Excuse My Dear Aunt Sally” or PEMDAS tells you what kind of calculations must be completed before other calculations. However, this can still be overwhelming. Although you must follow the PEMDAS order, you can start at the top left, and go right and down, just like how you read. In fact, PEMDAS is sorta like a sentence structure but with equations.

Note

- P- You must complete the calculations within parentheses before you can use those numbers in the rest of the equation.
- E- You must complete the exponents (which will always just be squaring a number in statistics) before you can use that number in the rest of the equation.
- M- You must complete the multiplication before you can add or subtract it with other numbers in the equation. I actually think that you're supposed to divide before you multiply?
- D- You must complete the division before you can add or subtract it with other numbers in the equation. I actually think that you're supposed to divide before you multiply?
- A- Once all of the other mathematical calculations are completed, you can now add or subtract. I actually think that you're supposed to subtract before you add?
- S- Once all of the other mathematical calculations are completed, you can now add or subtract. I actually think that you're supposed to subtract before you add?

What might throw you off in actual equations that we'll use is square rooting. Square rooting is actually the last operation because you have to get down to one number.

Notation & Symbols

Table **1.9.1** will cover some of the major mathematical symbols used in statistical equations. The column on the right is where you can write explanations or examples in your own words.

Table **1.9.1**- Common Mathematical Symbols

Symbol	Read As	Example or Description
>	"greater than"	
<	"less than"	
* or XY	"multiply" or "times"	
$()^2$	"square" or "times itself"	
√	"square root" or "find the number that, multiplied by itself, equals the number in the square root sign"	
|#|	“absolute value” or "make it a positive number" (the lines, not the #-sign)	

Summing (Sigma or Σ)

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 four grapes show in Table **1.9.2**:

Table **1.9.2**- Grape Weights in Grams

Grape	X
A	4.6
B	5.1
C	4.9
D	4.4

The following formula means to sum up the weights of the four grapes:

$$\sum X$$

The Greek letter **Σ** indicates summation. The X is a placeholder for all of the scores. Therefore,

$$\sum X = X_A + X_B + X_C + X_D = 4.6 + 5.1 + 4.9 + 4.4 = 19$$

Sums of Squares

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

$$\begin{aligned}(\sum X)^2 &= 4.6^2 + 5.1^2 + 4.9^2 + 4.4^2 \\ &= 21.16 + 26.01 + 24.01 + 19.36 = 90.54\end{aligned}$$

Notice that:

$$\left(\sum \mathrm{X}\right)^2 \neq \sum\left(\mathrm{X}^2\right)$$

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).

Multiplication (Products)

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

Table **1.9.2**- Table of Scores and Products

Participant	X	Y	XY
A	1	3	3
B	2	2	4
C	3	7	21
$\sum$	$\sum = 1+2+3 = 6$	$\sum = 3+2+7 = 12$	$\sum = 3+4+21 = 28$

Percentages

Providing the percentages is often more useful than merely displaying frequency information, especially when comparing distributions that have different sample sizes (different total amount of units/numbers).

Example 1.9.2

1. Having 1 red-head in a group of 10= _____%
2. Having 1 red-head in a group of 100= _____%

Solution

1. Having 1 red-head in a group of 10 $= 10\%$
2. Having 1 red-head in a group of 100 $= 1\%$

In Example 1.9.1, showing the frequency of one doesn't tell us as much as the percentage does because although 10% isn't a large proportion, it's way bigger than 1%. The only time when providing frequencies (counts) is better than percentages is when you have a very small population. For example, having 1 out of 3 people like your cologne is 33%, but it that doesn't really tell you anything meaningful about how you smell.

Once you get to a decimal point, you then multiply by 100 to get the percentage.

- $0.17 = 17\%$
- $0.94 = 94\%$

You probably all do this in your head, but try to write it out for now so that you don't skip a step.

Example 1.9.3

Imagine that there are 35 students in your class and 5 of them had earned full credit on all assignments. What's the percentage of students in your class earning an A+?

Solution

1. Division: $\frac{5}{35} = 0.1429$ (We would normally round decimal points to 2 numbers after the decimal, but for percentages we should include 4 numbers because we'll need to multiply them.
2. Multiplication: $0.1429 \times 100 = 14.29\%$
3. Conclusion: The percentage of students in your class earning an A+ was 14.29%

As we move on to the actual calculations used to describe distributions of data, you can check out this Crash Course video on data visualization which has a bit on percentages.

This page titled 1.9: Math Refresher is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **1.7: Mathematical Notation** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.
- **3.2: Math Refresher** by Michelle Oja is licensed CC BY-NC-SA 4.0.

CHAPTER OVERVIEW

2: Describing Data using Distributions and Graphs

2.1: Introduction to Looking at Data (This is what too many numbers looks like)
2.2: Frequency Tables
2.3: Graphing Qualitative Variables- Bar Graphs
2.4: Graphing Qualitative Variables- Pie Charts
2.5: Graphing Quantitative Variables
2.6: Skew and Kurtosis
2.7: Graphing Quantitative Data- Line Graphs
2.8: Graphing Quantitative Data- Histograms
2.9: Graphing Quantitative Data- Boxplots
2.10: Graphing Quantitative Data- Scatterplots
2.11: Summary and Some Honesty

This page titled 2: Describing Data using Distributions and Graphs is shared under a not declared license and was authored, remixed, and/or curated by Michelle Oja.

2.1: Introduction to Looking at Data (This is what too many numbers looks like)

I have scores on a 100-point final exam from one class provided by OpenIntro.org Can you look at them in Table **2.1.1** and quickly tell how the class did?

Table **2.1.1**- Final Exam Scores

79	83	66	81
83	72	89	88
57	74	78	69
82	73	81	77
94	71	78	79

Even if you could get a general idea of how these 20 students did on the final exam, this is not the best way to understand data sets. And just imagine if the class had 40 or 100 students in it!

When you deal with data, you may have so many numbers to you that you will be overwhelmed by them. That is why we need ways to describe the data in a more manageable fashion.

We'll start that in this chapter on charts, then describe whole sets of data in only a few important numbers in the following chapter.

This page titled 2.1: Introduction to Looking at Data (This is what too many numbers looks like) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **2.1: This is what too many numbers looks like** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

2.2: Frequency Tables

A table will have columns and rows. Columns go up and down (think of pillars or Greek columns), and rows go left to right on the horizon line.

A graph (also called a chart) is like a drawing in which the data is plotted on two axes. These will be discussed next.

Table **2.2.1** shows the same data as Table 2.1.1 (scores on a 100-point final exam from one class from an unknown college provided by OpenIntro.org). The only difference is that they are all in one column in Table **2.2.1**, which is typically how lists of raw data is provided.

Table **2.2.1**- Final Exam Scores

Final Exam Scores
79
83
57
82
94
83
72
74
73
71
66
89
78
81
78
81
88
69
77
79

Let's practice with what we learned from the prior chapter on this new set of data.

Example 2.2.1

1. Who was the sample?
2. Who could this sample represent (population)?
3. What was measured?
4. How many students scored above 90 points on the exam?

Solution

1. 20 students from an unknown college

2. Anything reasonable; maybe students in college?
3. Scores on a 100-point final exam
4. One

It's difficult to answer that last question with the raw data shown in Table **2.2.1**. It would be easier if the data was organized from highest score to lowest score, but an even easier way to answer these kinds of "how many" questions is to create a frequency table.

Definition: Frequency Table

Table showing each score in the "x" column, and how many people earned that score in the "f" column. The "x" stands in for whatever the score is, and the "f" stands for frequency.

Table **2.2.2** shows the same data as Table **2.2.1**, but in a Frequency Table generated by software (Software Package for the Social Sciences, or SPSS). As you can see in Table **2.2.2**, only one person scored each amount listed, except for the scores of 78, 79, 81, and 83 points (in which two students earned those scores).

Table **2.2.2**- Frequency Table of Final Exam Scores

x	f
57.00	1
66.00	1
69.00	1
71.00	1
72.00	1
73.00	1
74.00	1
77.00	1
78.00	2
79.00	2
81.00	2
82.00	1
83.00	2
88.00	1
89.00	1
94.00	1
Total	20

You might notice that the Table **2.2.2** does not include final exam scores that no student earned (for example, no scores are in the "x" column between 57 to 66 points). You can include all of the scores with a frequency ("f" column) of zero, it does make the table much complicated to interpret. What can be confusing is frequency tables must include scores of zeros ("x" is the score, not the frequency or number of people who earned that score). For example, if a student skipped the final exam, then their zero points earned would need to be included in the "x" column, with a frequency ("f" column) of one student.

The following practice Exercise should be much easier with Table **2.2.2**!

Exercise 2.2.1

1. How many students scored above 90 points on the exam?
2. How many students scored below 70 points?

Answer

1. One
2. Three

Did you notice that the x-variable, the final exam scores, was a quantitative variable (ratio scale of measurement)? Frequency Tables can also be used with qualitative variables (nominal scale of measurement). Table **2.2.3** shows the frequency ("f" column) of Associate of Arts for Transfer degrees (which are slightly different from Associate of Arts degrees in general) earned by California community college students in academic year 2019-2020 ("x" column). This data can be found on the California Community Colleges Chancellor's Office DataMart.

Table **2.2.3**- AA-T Majors

x	f
Biological Sciences	19
Education	8
Family and Consumer Sciences	616
Fine and Applied Arts	2,778
Foreign Language	717
Health	1,982
Humanities (Letters)	6,667
Interdisciplinary Studies	1,034
Law	36
Media and Communications	444
Psychology	10,843
Public and Protective Services	33
Social Sciences	13,164

The majors in Table **2.2.3** are listed alphabetically, which makes understanding the numbers a little more confusing. Because qualitative variables have no natural order, you can list them however you like. To find your own major, an alphabetical order works best. But if you'd like emphasize which major has the most graduates, then order from the highest to the lowest number of graduates would be best. As we will learn more with charts, the best way to display data depends on what message you are trying to tell.

Most of the graphs shown in the rest of the chapter are derived from frequency tables, so I hope that you find some simple data and practice making a few before we starting using them more!

Contributors and Attributions

- Foster et al. (University of Missouri-St. Louis, Rice University, & University of Houston, Downtown Campus)
- Dr. MO (Taft College)

This page titled 2.2: Frequency Tables is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

2.3: Graphing Qualitative Variables- Bar Graphs

Graphing Qualitative Variables

The first types of charts that will be covered are bar charts and pie charts. Both types of charts should be used on qualitative variables only. There's another whole mess of chart types to use on quantitative variables! Review the section on quantitative and qualitative variables in the first chapter to refresh yourself on qualitative variables and the nominal scale of measurement.

Bar Charts

Bar charts are meant to display the frequencies of qualitative variables. Figure **2.3.1** shows the same data as the frequency table in Table 2.2.3, but with bars.

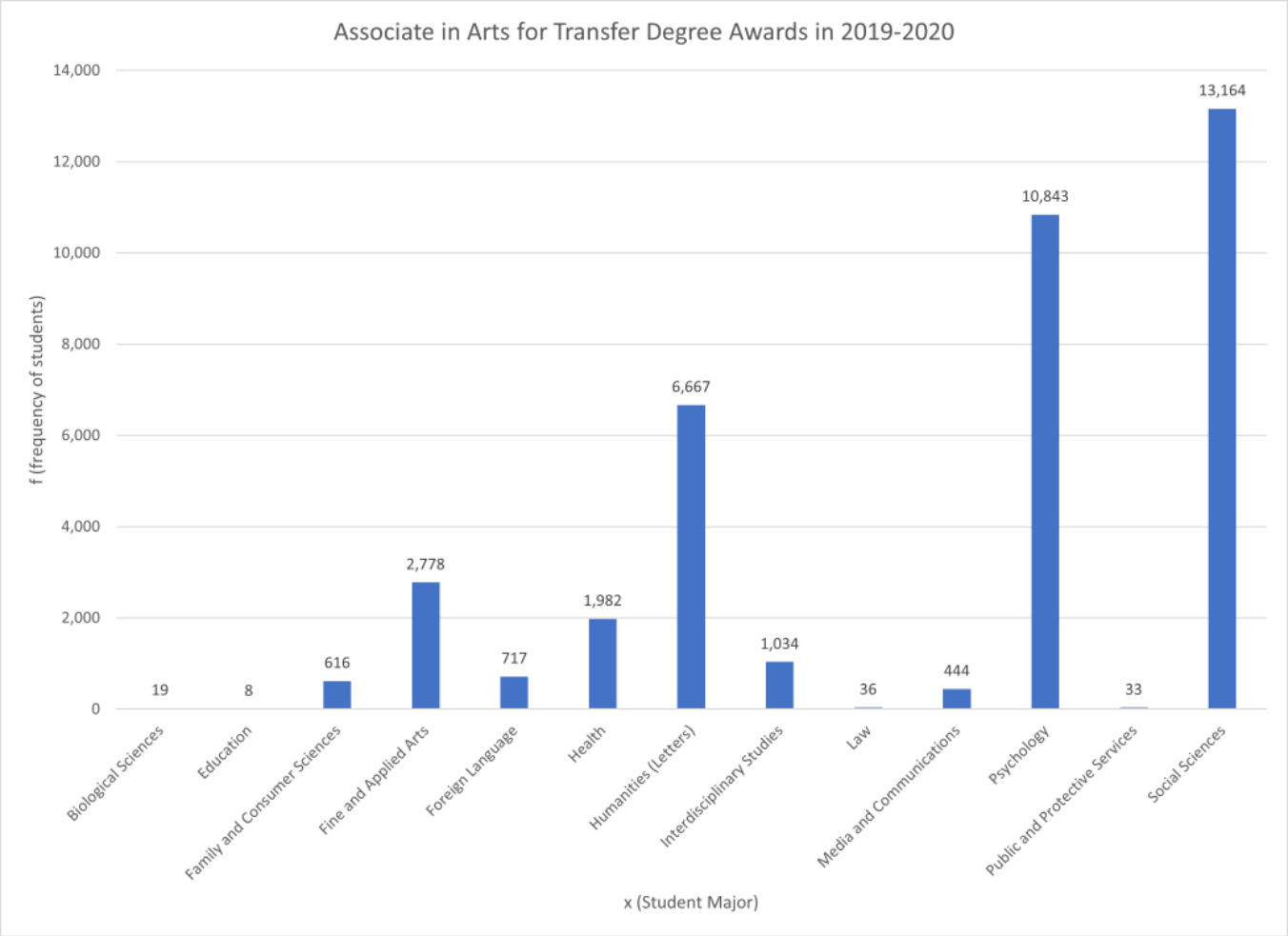


Figure **2.3.1**: AA-T Graduates by Major (Copyright CC0; chart created by Michelle Oja via California Community Colleges Chancellor's Office DataMart)

Let's unpack Figure **2.3.1** a bit.

First, the x-axis has the same information as the "x" column in Table 2.2.3, the major that the graduating student earned the degree in. The x-axis is always the one on the bottom going from left to right. Because we are talking about a variable that has categories (nominal scale of measurement; qualitative variable), the information on the x-axis is the category names. They can be in any order on the chart because qualitative variables don't have a natural order. As you can see, they are ordered alphabetically. This makes no sense to me, so I usually order from lowest frequency to highest frequency but wanted to let you see what is looks like when not ordered numerically. You can choose any order that you'd like since qualitative variables have no natural order.

Second, the y-axis in Figure **2.3.1** shows the same information as the "f" column in Table 2.2.3, the number of students who earned that degree. The y-axis is always going up and down. Stand with your hands touching above your head to show that this is the axis that is going up and down, then move your hands out to a "Y" (like in the song YMCA) to remember that the y-axis is the one that goes up and down.

Third, notice that the bars do not touch; this is what makes it a bar chart and not a histogram (which will be covered later in this chapter). This chart was created with common commercial spreadsheet software (Excel), but note that this type of chart is called a "column bar chart" in the software. There are fancier types of bar charts, but often the fancier a chart looks, the more difficult it is to easily and quickly understand what the chart is showing.

Interpreting Bar Charts

The following questions will be asked for each type of graph in this chapter. In most cases, the questions will also be answered so that you can get an idea of what you should be looking for in graphs. Some of these questions won't make sense for specific types of graphs, but I want you to get in the habit of answering them to yourself when reviewing graphs.

1. What kind of graph is Figure **2.3.1**?
    1. This is a bar graph (notice that the bars are not touching) because the variable is a qualitative category (nominal scale of measurement).
2. What does the x-axis measure in Figure **2.3.1**?
    1. The x-axis is the one on the bottom, and it was named "Student Major". This is the major that the student earned an Associate of Arts for Transfer degree from a California community college student during the 2019-2020 academic year.
3. What does the y-axis measure in Figure **2.3.1**?
    1. The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. In this figure, that means that the y-axis shows how many students earned the Associate of Arts for Transfer degree from each major.
4. What do you notice from Figure **2.3.1**? What pops out to you?
    1. What I noticed is that most students are earning the general "Social Science" degree, rather than something more specific like Sociology or Psychology. I also noticed that there were very few students earned an Associate of Arts for Transfer degree with an Education major.
5. What does Figure **2.3.1** make you wonder about?
    1. I wonder what kind of degree future teachers are earning, because it does not look like they are earning an Associate of Arts for Transfer! Maybe a general Associate of Arts? Or is there a different major that they are in?
6. What is a catchy headline for Figure **2.3.1**?
    1. Be Like the Ten Thousand Psychology Graduates- Go to a California Community College
7. How could you summarize the info in Figure **2.3.1** into one sentence?
    1. The most common Associate of Arts for Transfer degrees were in the Social Sciences, Psychology, Humanities, and Fine and Applied Arts.
8. Who might want to know the information in Figure **2.3.1**?
    1. California community college students? Potential California community college students? Faculty or administrators of California community colleges?

? Exercise 2.3.1

Remember those videos about the best way to learn from Dr. Chew, a cognitive psychologist, first discussed in the first chapter? Well, Dr. Chew's second video has a bar chart (Part 2: What Students Should Know about How People Learn! I suggest that you watch the whole video to understand the experiment, but the bar chart shows up around 3:10 minutes. Then, answer the same questions as above.

Answer

1. What kind of graph is in the video?
    1. This is a bar graph (notice that the bars are not touching).
2. What does the x-axis measure in the video?
    1. The x-axis is the one on the bottom, named “Level of Processing” and it has three levels (shallow, deep, and the control group). There are different colored bars, and they represent another variable (memory goal?) that has two levels, intentional and incidental.
3. What does the y-axis measure in the video?
    1. The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. However, in this video, the y-axis is showing the percentage of words that were successfully recalled (remembered).
4. What do you notice from the graph in the video? What pops out to you?
    1. What I noticed is that shallow processing (looking for the letter “e”) leads to worse recall than not having any processing goal (the control group). I also find it interesting that the intent to learn didn’t affect memory really at all. But you might have noticed other things!
5. What does the graph in the video make you wonder about?
    1. I wonder why deep processing (elaborating) isn’t more well known to help students?! But what does the graph make you wonder?
6. What is a catchy headline for the graph in the video?
    1. "Learning is Deep" seems catchy to me, but I bet you came up with something better!
7. How could you summarize the info in the graph in the video into one sentence?
    1. You could say something like, "Focusing on the meaning of words doesn’t seem to help you remember much more than just trying to remember information, but it is way better than shallow processing."
8. Who might want to know the information from the graph in the video?
    1. I imagine that both students and teachers would find this information very useful. Tutors and other learning professionals should also know about it, too.

Summary

Bar charts effectively portraying qualitative data. Bar charts are a good option when there are more than just a few categories, or for comparing two or more distributions.

This page titled 2.3: Graphing Qualitative Variables- Bar Graphs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

2.4: Graphing Qualitative Variables- Pie Charts

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. Instead of showing frequencies, a pie chart shows *proportions*. Figure **2.4.1** shows the same information as the frequency table in Table 2.2.3 and the bar chart in Figure 2.4.1, but as a pie chart.

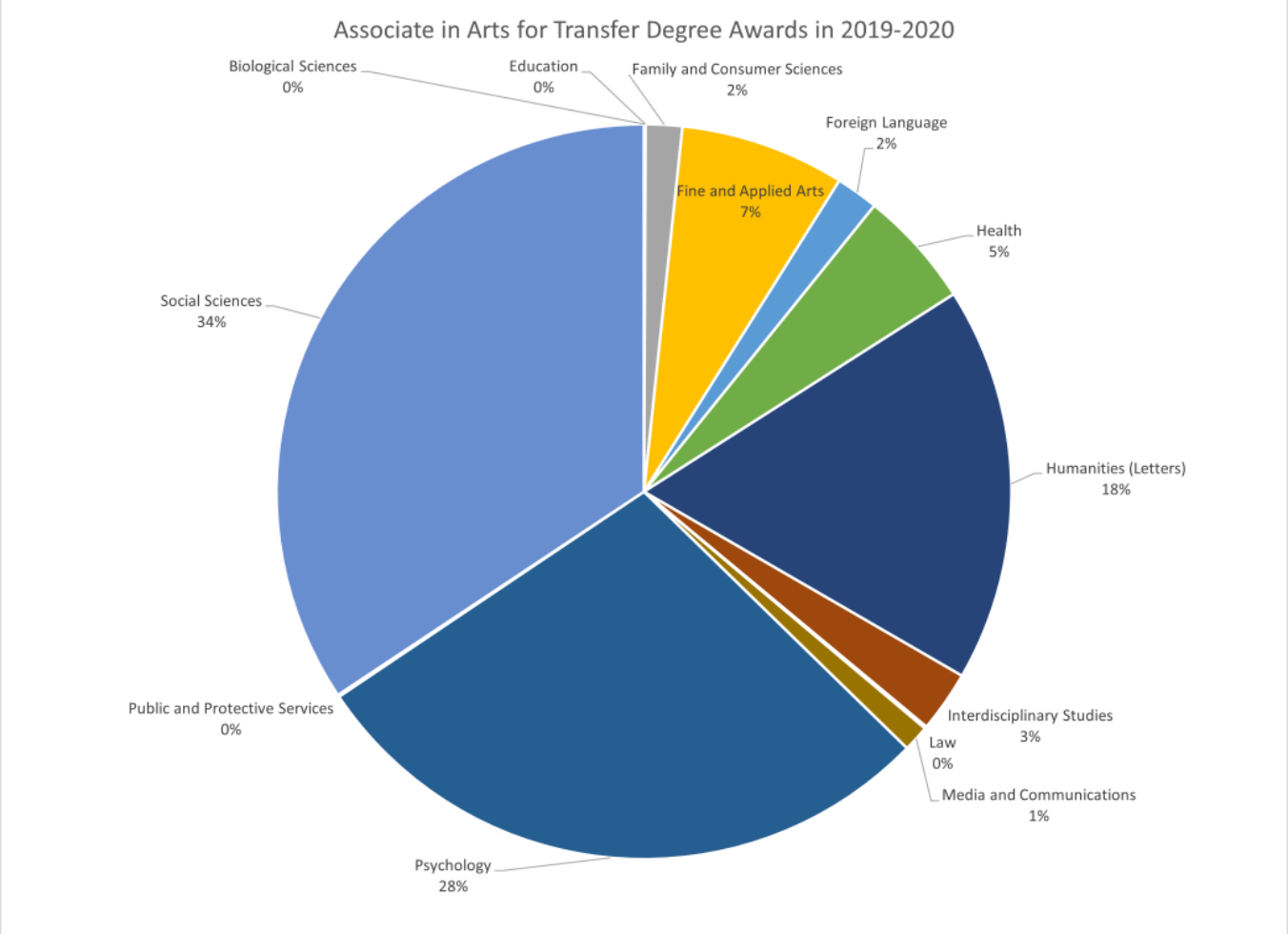


Figure **2.4.1**- Pie Chart of Graduate's Majors (Copyright CC0; chart created by Michelle Oja via California Community Colleges Chancellor's Office DataMart)

Both bar charts and pie charts show information about how many people are in each group, but bar charts can show the exam numbers while pie charts show the proportion, or percentage. You can learn different things from each chart, so it's up to you to decide which chart is best for you needs. Remembering that both of these kinds of charts show how many people are in each qualitative variable's group!

Pie Chart Interpretation

The following will have similar answers as the bar chart interpretation because it's the same data. But the answers aren't exactly the same because some things pop out in pie charts but not in bar charts, and vice versa.

1. What kind of graph is Figure **2.4.1**?
    1. This is a pie chart. Pie charts are used with variables that have qualitative categories (nominal scale of measurement) when you are want to compare proportions (percentages).
2. What does the x-axis measure in Figure **2.4.1**?

1. Trick questions! There's no axes in pie charts! What is normally on the x-axis is the category or each score, and that is what the colors and labels identify.

3. What does the y-axis measure in Figure **2.4.1**?
    1. Another trick question! Again, there are no axes in pie charts! What is normally on the y-axis is the frequency; instead of having the numbers on an axis, the size of the pie slices indicates the size or portion of each category. In this example, I also included the percentage next to the name of the category.
4. What do you notice fromFigure **2.4.1**? What pops out to you?
    1. The pie chart shows which groups have the most graduates with Associate of Arts for Transfer, and that's Social Sciences, Psychology, Humanities, and Fine and Applied Arts. There are also a bunch of majors that show 0% of the pie. They really don't have 0%, or they wouldn't be in the chart. Instead, those are the ones with really small frequencies, and their percentage is a decimal point (like 0.05% for Biological Sciences).
5. What does Figure **2.4.1** make you wonder about?
    1. Because the general "Social Sciences" is such a large proportion (slice of pie), I wonder what kind of careers they go into?
6. What is a catchy headline for Figure **2.4.1**?
    1. Your Rainbow of Choices- Associate of Arts for Transfer from California Community Colleges
7. How could you summarize the info in Figure **2.4.1** into one sentence?
    1. I would still say that the most common Associate of Arts for Transfer degrees were in the Social Sciences, Psychology, Humanities, and Fine and Applied Arts, but the pie chart format shows how small of a proprotion many of other majors earn.
8. Who might want to know the information in Figure **2.4.1**?
    1. Same as from the bar chart! California community college students? Potential California community college students? Faculty or administrators of California community colleges?

Pie charts can be misleading if they are based on a small number of observations. For example, if just five people had graduated with an Associate of Arts for Transfer, and three earned their transfer degree in Psychology, it would be misleading to display a pie chart with the Psychology major slice showing 60%. Problems like that can happen with small samples. In this case, the slices should be labeled with the actual frequencies observed (e.g., 3) instead of with percentages.

Summary

Pie charts and bar charts can both be effective methods of portraying qualitative data. Let's move on to graphing quantitative data!

Contributors and Attributions

This page was extensively adapted by Michelle Oja (Taft College) from work by Foster et al. (sourced from University of Missouri's Affordable and Open Access Educational Resources Initiative).

This page titled 2.4: Graphing Qualitative Variables- Pie Charts is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

2.5: Graphing Quantitative Variables

Review the first chapter to refresh yourself on what quantitative variables and the ratio and interval scale of measurement describes.

Briefly, 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 qualitative (sometimes called nominal or categorical) 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:

1. Line graphs (sometimes called frequency polygons)
2. Histograms
3. Box plots
4. Scatter plots

Each type of chart highlights different kinds of information. It is up to you to decide which fits your purposes best!

The Shape of Distribution

The next section will go more in-depth, and the following chapter will go into even more detail, but here is an introduction to describing the shape of graphs of quantitative variables.

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 two 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. 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 **2.5.1**. 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.

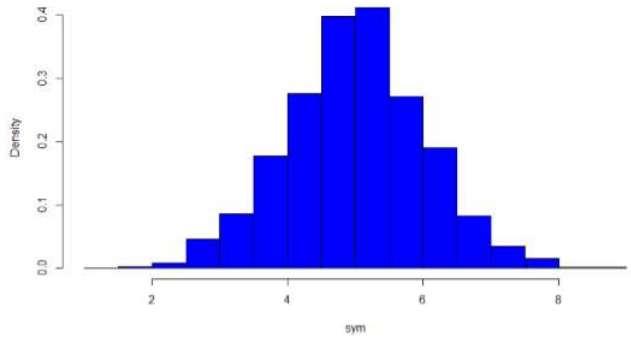


Figure **2.5.1**: A symmetrical distribution in a histogram (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Symmetrical distributions can also have multiple peaks. Figure **2.5.2** shows a bimodal distribution (bimodal means two modes), named for the two peaks that lie roughly symmetrically on either side of the center point. This is relatively difficult characteristic to detect numerically. Thus, it is important to visualize your data before moving ahead with any statistical analyses.

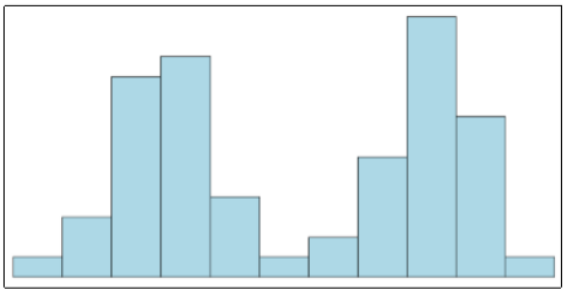


Figure 2.5.2: A bimodal distribution in a histogram (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 2.5.3 and 2.5.4 show positive (right) and negative (left) skew, respectively.

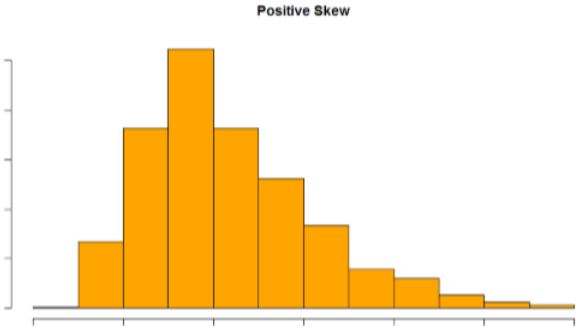


Figure 2.5.3: A positively skewed distribution in a histogram (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

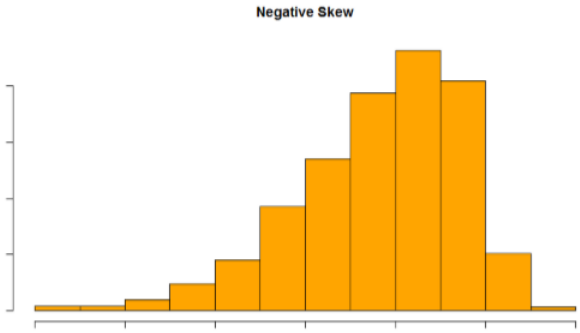


Figure 2.5.4: A negatively skewed distribution in a histogram (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

The next section will go into more detail about skew, and introduce you to a new term to describe different shapes of symmetrical distributions.

Contributors and Attributions

- Foster et al. (University of Missouri-St. Louis, Rice University, & University of Houston, Downtown Campus)
- Dr. MO (Taft College)

This page titled 2.5: 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.

2.6: Skew and Kurtosis

Statisticians and researchers look at a lot of frequency charts (usually line graphs and histograms), so they know what to look for. Khan Academy has a great video previewing what will be discussed in this section.

Skew

Since it's the more interesting of the two, let's start by talking about the skew.

The shape of a frequency chart (line graph or histogram) can tell you a lot of about that data set. The simplest things to look for are any extreme scores (called *outliers*) that seem to be much higher or much lower than most of the other scores. These outliers may affect the shape of the distribution, making it skewed. A skewed distribution is one in which many scores are bunched up to one side, and there are only a few scores on the other side. A distribution can be positively skewed (meaning that the scores are bunched to the left, and the thin tail is pointing to the right) or negatively skewed (meaning that the scores are bunched to the right, and the thin tail is pointing to the left). Figure **2.6.1** shows examples of a positively skewed line graph (on the right) and a negatively skewed line graph on the left.

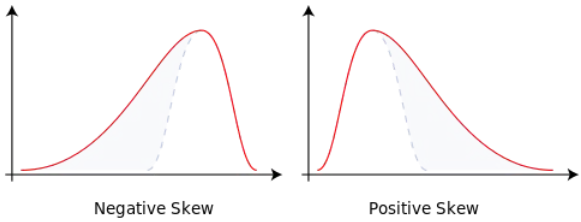


Figure **2.6.1**- Diagram of Skew. (CC-BY-SA Rodolfo Hermans (Godot), CC BY-SA 3.0, via Wikimedia Commons)

Skewness is basically a measure of asymmetry, and the easiest way to explain it is by drawing some pictures. As Figure **2.6.2** illustrates in histograms, if the data tend to have a lot of extreme small values (i.e., the lower tail is "longer" than the upper tail) and not so many extremely large values (left panel), then we say that the data are *negatively skewed*. On the other hand, if there are more extremely large values than extremely small ones (right panel) we say that the data are *positively skewed*. That's the qualitative idea behind skewness.

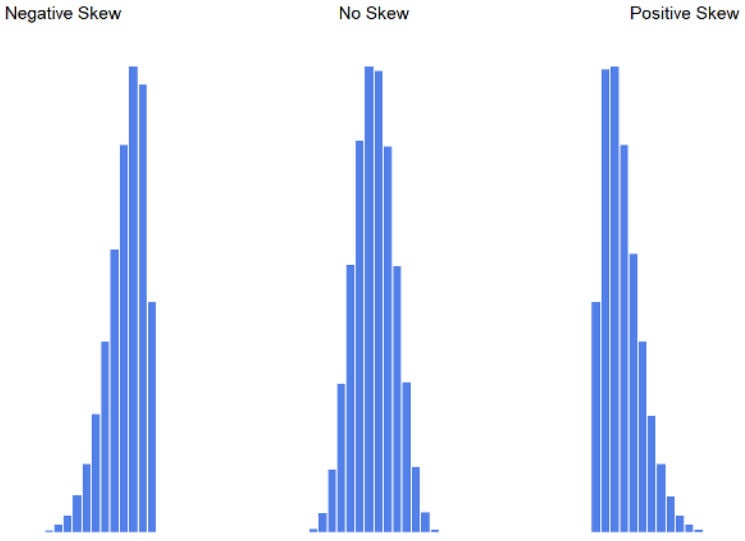


Figure **2.6.2**- Diagram of Skew and No Skew (CC-BY-SA Danielle Navarro from Learning Statistics with R)

Kurtosis

For distributions that are not skewed, you can look at how short and fat the distribution is, or if it is a medium bell-shaped curve, or if it is tall and narrow. This way of describing the shape of symmetrical (not-skewed) distributions, how broad the distribution is, is called kurtosis. Put simply, kurtosis is a measure of the "tailedness" of the data. As you can see in the line graph in Figure **2.6.3**,

there are three main types: wide, medium, or tall. Wide and flat graphs are called platykurtic. Medium, bell-shaped graphs are called mesokurtic or a normal distribution. Tall and narrow graphs are called leptokurtic.

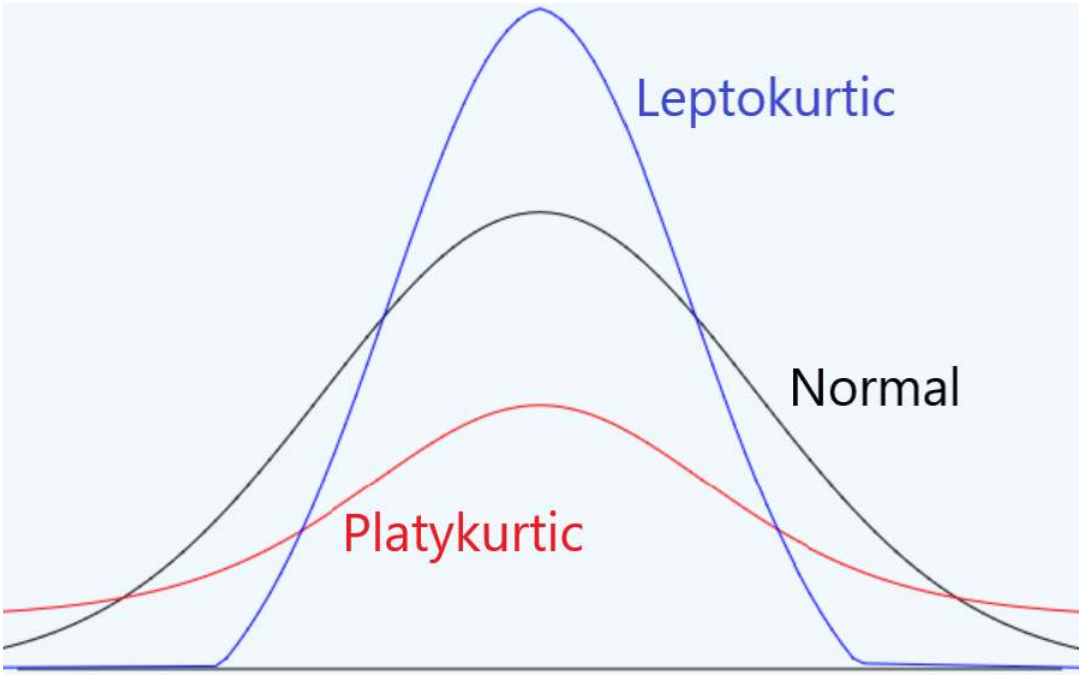


Figure **2.6.3**- Diagram of Three Kurtosis Types in Line Graphs (CC-BY Larry Green)

Figure **2.6.4** shows the three types of kurtosis in histograms. The histogram on the left shows platykurtic data, the middle histogram shows mesokurtic data, and on the right, we have a leptokurtic data set. By mathematical calculations, the "normal curve" (black lines) has zero kurtosis.

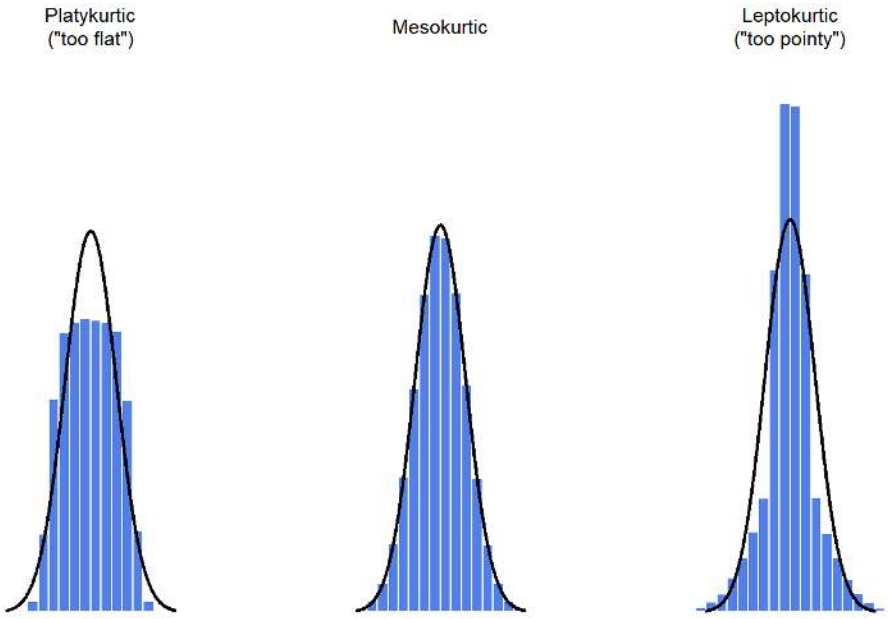


Figure **2.6.4**: Diagram of Three Kurtosis Types in Histograms (CC-BY-SA Danielle Navarro from Learning Statistics with R)

The types of kurtosis are summarized in Table **2.6.1**.

Table **2.6.1**- Informal Description of Types of Kurtosis

Technical Name	Informal Description
mesokurtic	just pointy enough
leptokurtic	too pointy
platykurtic	too flat

Summary

In practice, neither skew nor kurtosis is used anywhere near as frequently as the measures of central tendency and variability that will be discussed in next chapter. Skew is pretty important, though, so you do see it mentioned a fair bit; but I've actually never seen kurtosis reported in a scientific article to date

Contributors

- Danielle Navarro (University of New South Wales)
- Peter H. Westfall (Paul Whitfield Horn Professor and James and Marguerite Niver Professor, Texas Tech University)
- Dr. MO (Taft College)

- Figure **2.6.1** Image Attribution: Rodolfo Hermans (Godot), CC BY-SA 3.0, via Wikimedia Commons
- Figure **2.6.3** Image Attribution: CC-BY Larry Green from Lake Tahoe Community College

This page titled 2.6: Skew and Kurtosis is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Danielle Navarro.

2.7: Graphing Quantitative Data- Line Graphs

Although line graphs are great at showing how data changes through time, we will mostly use line graphs to show frequency distributions; line graphs of frequency distributions are sometimes called *frequency polygons*, but that just reminds me of high school geometry. For example, the frequency table in Table 2.2.2 showed the frequency of final exam scores, a quantitative variable, while Figure **2.7.1** shows the same information but as a graph:

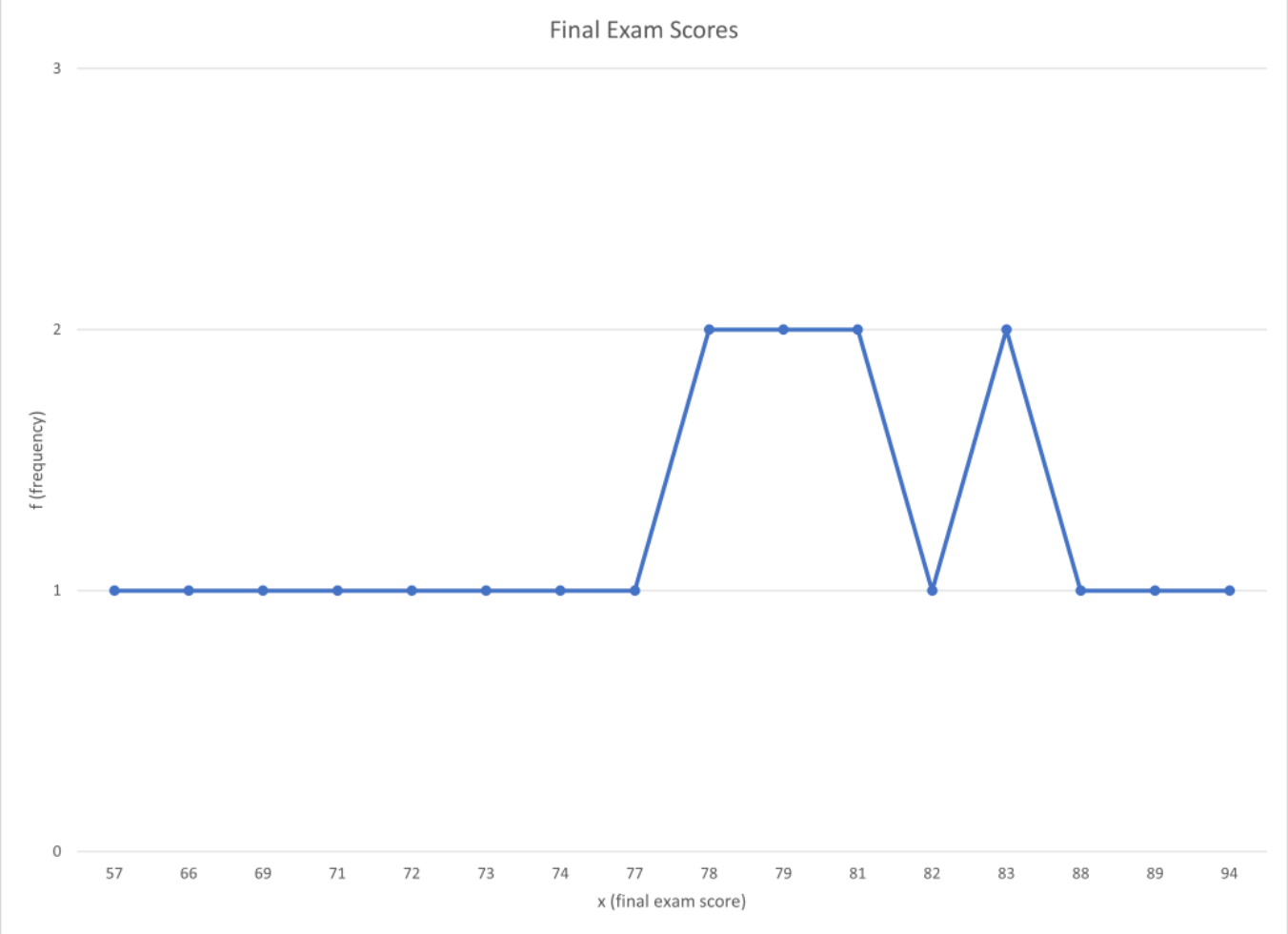


Figure **2.7.1**: Line Graph of Frequency of Final Exam Scores. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

Let's unpack Figure **2.7.1**, as well. First, the x-axis has the same information as the "x" column in Figure 2.1, each score that was earned on the final exam. The x-axis is always the one on the bottom going from left to right. Because we are talking about a numerical variable (ratio scale of measurement; quantitative variable), the information on the x-axis is *each score*. In a line graph, you did not create your own categories; for example, you would not create categories based on letter grades. The scores should be from lowest to highest. Second, the y-axis shows the same information as the "f" column in Figure 2.2.2, the number of students who earned that score. The y-axis is always going up and down. This chart was also created with a common commercial spreadsheet software (Excel), and there are again fancier types of line graph. And again, the fancier a chart looks, the more difficult it is to easily and quickly understand what the chart is showing.

Interpreting Line Graphs

I added a question to your interpretations.

Note

Can you see which question was added? Why do you think it was added for line charts?

1. What kind of graph is Figure **2.7.1**?
    1. This is a line graph. Line graphs are used to display quantitative data (or changes through time)
2. What does the x-axis measure in Figure **2.7.1**?
    1. The x-axis on line graphs show each score (or each time period). The x-axis on line graphs shows each time period. In a line graph showing changes through time, the x-axis will always be some marker of time.
3. What does the y-axis measure in Figure **2.7.1**?
    1. The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. However, the y-axis in Figure **2.7.2** shows the average percentages earned out of all points possible in the class each semester.
4. [New question!] Is Figure **2.7.1** skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
    1. Because the tail is so long on th left and there seems to be a pile of scores towards the right, I would say that this line graph is negatively skewed. This seems backwards to me, so I looked back at Figure 2.xxx (the picture of the skewed charts). Because it is skewed, it can't have kurtosis.
5. What do you notice from Figure **2.7.1**? What pops out to you?
    1. Two things that pop out to me. First, there is a bump of more frequent scores between 78 to 83 points. This suggests that most students in this unknown class earned a C+ or a B- on the Final Exam. The second thing I notice is that the left tail is pretty long. This means there are also quite a few students who earned low scores on the Final Exam.
6. What does Figure **2.7.1** make you wonder about?
    1. This chart makes me wonder if the professor ended up "curving" the exam since the highest score was 94 points, and that score was earned by only one student. The next highest score was 89 points, again with only one student earning that score on the Final Exam. If I haven't said this before, THIS IS NOT MY CLASS.
7. What is a catchy headline for Figure **2.7.1**?
    1. 85% of the Class Passed! [17 out of 20 students earned 70 points or higher]
8. How could you summarize the info in Figure **2.7.1** into one sentence?
    1. Almost everyone in the class passed the Final Exam, but only one student earned more than 90 points.
9. Who might want to know the information in Figure **2.7.1**?
    1. I would guess that students who are going to take this class from this professor might be interested. College administrators might also want to know so that they can see that most students are passing their classes.

Line Graphs Through Time

But for thoroughness, let's also look at a line graph that shows data through time. Figure **2.7.2** shows the percentage of points earn out of all points in Dr. MO's behavioral statistics courses in seven consecutive semesters (starting in Fall 2013, and ending in Fall 2016 when she started teaching a modified version of the course).

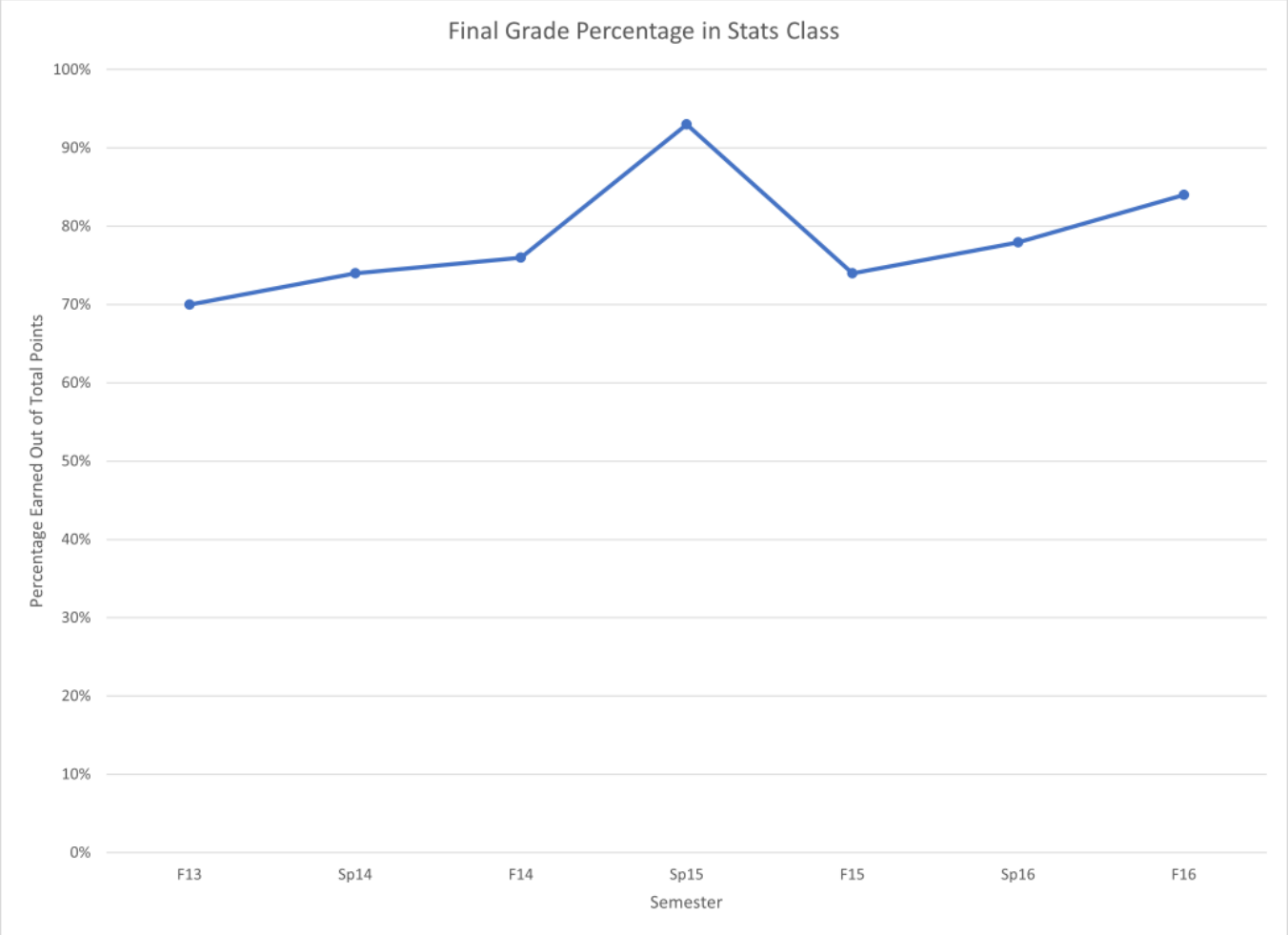


Figure 2.7.2- Line Graph of Percentages Through Time (Copyright CC-BY-SA; Michelle Oja)

Unpacking Figure 2.7.2, what do you see? First, let's compare Figure 2.7.2 to Figure 2.7.1. In Figure 2.7.1, the x-axis was each score, while in Figure 2.7.2 the x-axis is the time period. Similarly, in Figure 2.7.1, the y-axis in the frequency, the number of students who earned that score. In Figure 2.7.2, the y-axis is the percentage earned. Frequency distributions are really important for statistical analyses, but they aren't the only type of chart that you'll come across.

Interpreting Line Graphs

- What kind of graph is Figure 2.7.2?
  - This is a line graph. Line graphs are used to display quantitative data (or changes through time)
- What does the x-axis measure in Figure 2.7.2?
  - The x-axis on line graphs shows each time period. In a line graph showing changes through time, the x-axis will always be some marker of time.
- What does the y-axis measure in Figure 2.7.2?
  - The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. However, the y-axis in Figure 2.7.2 shows the average percentages earned out of all points possible in the class each semester.
- Is Figure 2.7.2 skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
  - This question doesn't really make sense when the x-axis isn't scores (quantitative data points). In line graphs through time, we are looking for trends not kurtosis or skew.
- What do you notice from Figure 2.7.2? What pops out to you?

- Spring 2015 was a great year for students!
- What does Figure **2.7.2** make you wonder about?
  - What happened in Spring 2015?! And why didn't it keep happening?
- What is a catchy headline for Figure **2.7.2**?
  - Average is Passing in Dr. MO's Behavioral Statistics Class
- How could you summarize the info in Figure **2.7.2** into one sentence?
  - The average percentages were all passing, with Spring 2015 being a particularly high average score.
- Who might want to know the information in Figure **2.7.2**?
  - I would guess that students who are going to take this class from this professor might be interested. College administrators might also want to know so that they can see that most students are passing this professor's class.

Another kind of graph that is used to show quantitative variables is a histogram, discussed next!

Contributors and Attributions

This page was extensively adapted by Michelle Oja (Taft College) from work by Matthew J. C. Crump (Brooklyn College of CUNY)

This page titled 2.7: Graphing Quantitative Data- Line Graphs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

2.8: Graphing Quantitative Data- Histograms

Another common type of graph for frequency distributions for quantitative variables (interval or ratio scales of measurement) is a histogram.

? Exercise 2.8.1

What other type of chart shows a frequency distribution for quantitative variables? Hint: We've already talked about it.

Answer

Line graph (sometimes called a frequency polygon)

? Exercise 2.8.1

Histograms look more like bar graphs, but are actually more like line graphs. Why?

Answer

Your answer probably talks about how the bars in bar charts don't touch, but the bars (or bins) in histograms do touch. Your answer should say something about the type of variable: quantitative variables (ratio or interval scale of measurements) should be graphed with histograms, and qualitative variables (nominal scale of measurement) should be graphed with bar charts.

Figure **2.8.1** shows a histogram (created by SPSS) of the frequency distribution of the final exam score data from Table 2.2.2 (and also shown in the line graph in Figure 2.8.1.

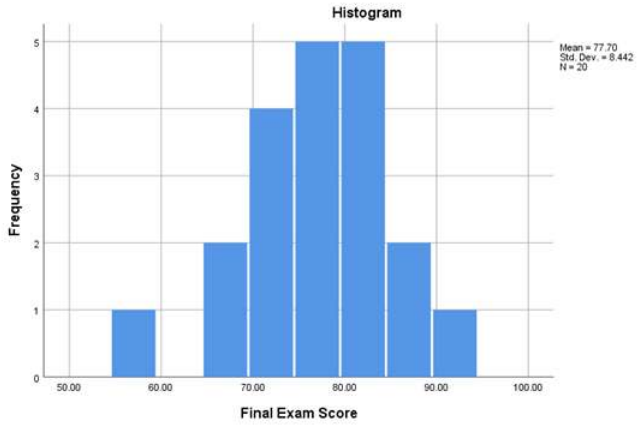


Figure **2.8.1**: Histogram of Frequency of Final Exam Scores. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

As with all frequency distributions, the y-axis of a histogram shows the frequency of each score (how many people had each score) and the x-axis has each score. Figure **2.8.1** shows the number of students who in each score category on the y-axis (Frequency), and labels the score categories on the x-axis (Final Exam Score, with each 10 points labeled). To make the histogram, we just count up the number of data points falling inside each bin, then plot those frequency counts as a function of the bins. Voila, a histogram.The 10-point categories are called bins. Bin #1 goes from about 55 to abou 59 points, bin #2 goes about 65 to 69 points, and so on until the last bin. The difference in the x-axis between a line graph and a histogram is that the scores are combined into ranges, or categories, in a histogram. This makes it easier to quickly understand what a graph is showing if there are a lot of different scores (like everyone's GPA in your class), but it does lose some accuracy.

Each bar (bin), and can be used to answer questions about the frequency of scores within these bin categories. For example, how many people scored between 90 to 100 points on the Final Exam? The seventh bar (or first on the right), the one between 90 and

100 on the x-axis, tells you how many. Look how tall that bar is. How tall is it? The height is shown on the y-axis, which provides the frequency. One person scored between 90 points and 100 points on the Final Exam.

Just like a line graph, a histogram can show the shape (kurtosis) or skew, and the range of our data set. The shape of the histogram refers to how it goes up and down. The shape tells us where the data is. For example, when the bars are low we know there isn't much data there. When the bars are high, we know there is more data there. So, where is most of the data? It looks like it's mostly in the middle two bins, between 75-85 points. The range of the data tells us the minimum (lowest score) and the maximum (highest score) of the data. In Figure **2.8.1**, most of the scores are between 65 points and 95 points, but the minimum scores is between 55-59 points.

When you make a histogram you get to choose how wide each bar will be. For example, below are four different histograms of the same data (a happiness survey). What changes is the width of the bins.

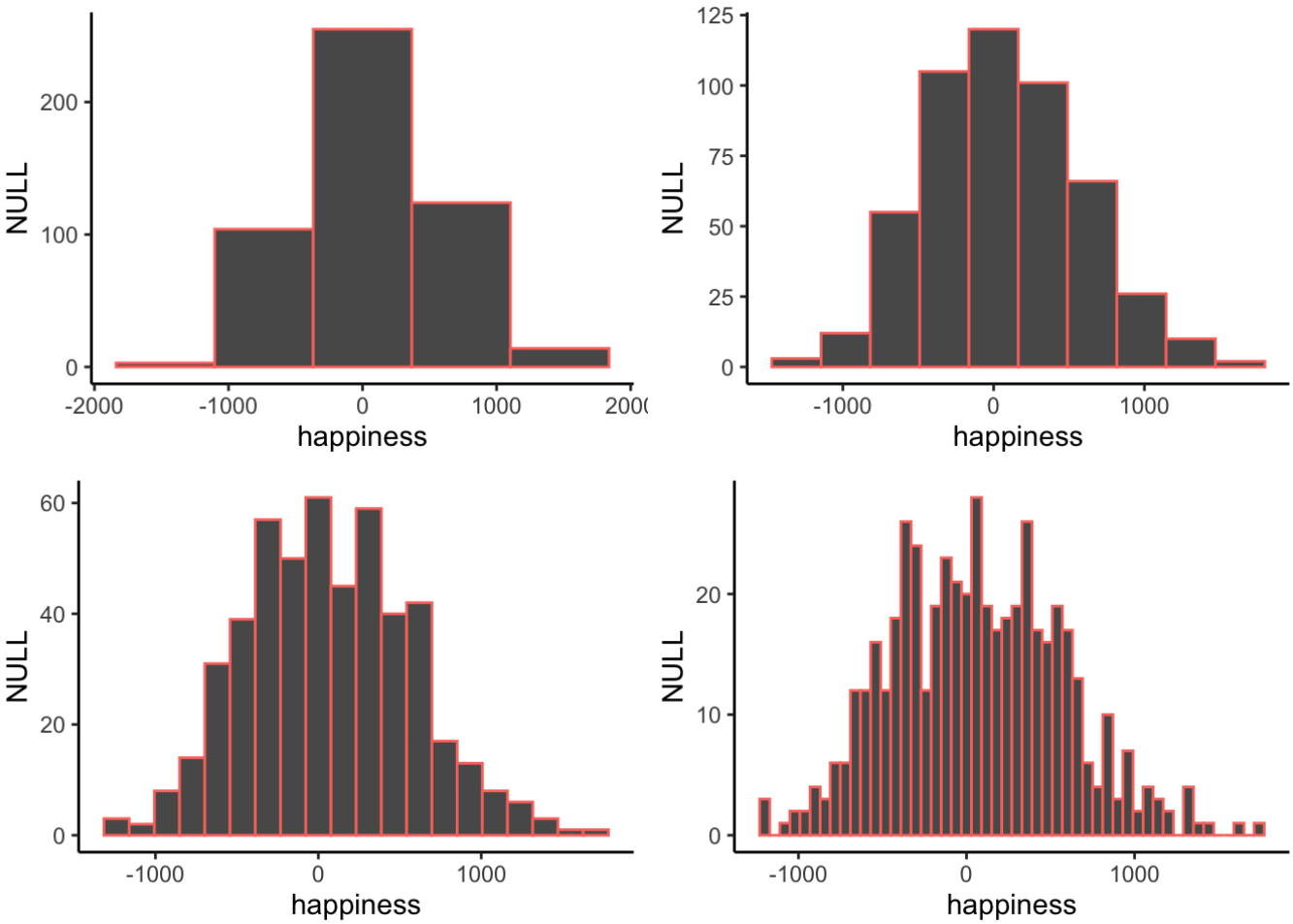


Figure **2.8.2**: Four histograms of the same data using different bin widths. (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

All of the histograms have roughly the same overall shape: From left to right, the bars start off small, then go up, then get small again. In other words, as the numbers get closer to zero, they start to occur more frequently. We see this general trend across all the histograms. But, some aspects of the trend fall apart when the bars get really narrow. For example, although the bars generally get taller when moving from -1000 to 0, there are some exceptions and the bars seem to fluctuate a little bit. When the bars are wider, there are less exceptions to the general trend. How wide or narrow should your histogram be? It's a Goldilocks question. Make it just right for your data.

Students often get confused between a bar graph and a histogram because they both have bars. There are two differences, however. First, in a histogram, the bars must touch; there should **not** be space between the categories that were created from the quantitative

data. This leads to the second difference between histograms and bar charts: Histograms show quantitative variables, while bar charts show qualitative variables (nominal scale of measurement).

Histogram Interpretation

The following are the same questions asked about the Figure 2.8.1, the line graph showing the frequency distribution of this same data. You might find it interesting how the answers change a little depending on how the data is presented.

1. What kind of graph is Figure **2.8.1**?
    1. This is a histogram. Histograms are used to display quantitative data, but it combines each score into groups. In Figure 2.xxx, the groups are 10-point ranges.
2. What does the x-axis measure in Figure **2.8.1**?
    1. The x-axis on histograms show categories based on each score. In Figure **2.8.1**, the software created 10-point categories for each Final Exam Score that a student earned.
3. What does the y-axis measure in Figure **2.8.1**?
    1. The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. In Figure **2.8.1**, that means that the y-axis shows how many students earned scores on the Final Exam in each 10-point range.
4. Is Figure **2.8.1** skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
    1. Because there is a gap between the lowest score category and the next one, I would say that this histrogram is negatively skewed. However, it looks less skewed than the same data in the line graph in Figure 2.8.1. In fact, just by looking at Figure **2.8.1** and not knowing the exact scores, I might have said that it is not skewed and that is has approximately normal kurtosis (mesokurtic).
5. What do you notice from Figure **2.8.1**? What pops out to you?
    1. Two things that pop out to me. First, there seemed to be the most scores around the middle of the distribution. The second thing I notice is that one column that's to the left. This shows that one student (frequency = 1) score below 60 points on the Final Exam.
6. What does Figure **2.8.1** make you wonder about?
    1. This histogram still makes me wonder if the professor ended up "curving" the exam so that the highest earning score was now like earning 100% on the Final Exam. If I haven't said this before, THIS IS NOT MY CLASS.
7. What is a catchy headline for Figure **2.8.1**?
    1. Final Exam Scores are Normal [This is catchy if you know what a Normal Distribution is; more in Ch. 4!]
8. How could you summarize the info in Figure **2.8.1** into one sentence?
    1. The class did really well on the Final Exam, although one student scored pretty low.
9. Who might want to know the information in Figure **2.8.1**?
    1. I still am guessing that students who are going to take this class from this professor might be interested. College administrators might also want to know so that they can see that most students are passing their classes.

Summary

Making a histogram is our first act of officially *summarizing* the data. We are no longer look at the individual bits of data, instead we will see how the numbers group together. Let's look at a histogram of the happiness data, and then explain it.

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 2.8: Graphing Quantitative Data- Histograms is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **2.2: Look at the data** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

2.9: Graphing Quantitative Data- Boxplots

You might think that you've never seen a box plot, but you probably have seen something similar.

An alternative to line graphs and histograms is a *boxplot*, sometimes called a “box and whiskers” plot. Like line graphs and histograms, they’re best suited to quantitative data (interval or ratio scale of measurement). The idea behind a boxplot is to provide a simple visual depiction of the score in the exact middle (median), the where about each fourth (quartile) of the scores are, and the range of the data. And because boxplots do so in a fairly compact and intuitive way, they have become a very popular statistical graphic. When you look at this plot, this is how you should interpret it: the thick line in the middle of the box is the median; the box itself spans the range from the 25th percentile to the 75th percentile; and the “whiskers” cover the full range from the minimum value to the maximum value. This is summarized in the annotated plot in Figure **2.9.1**.

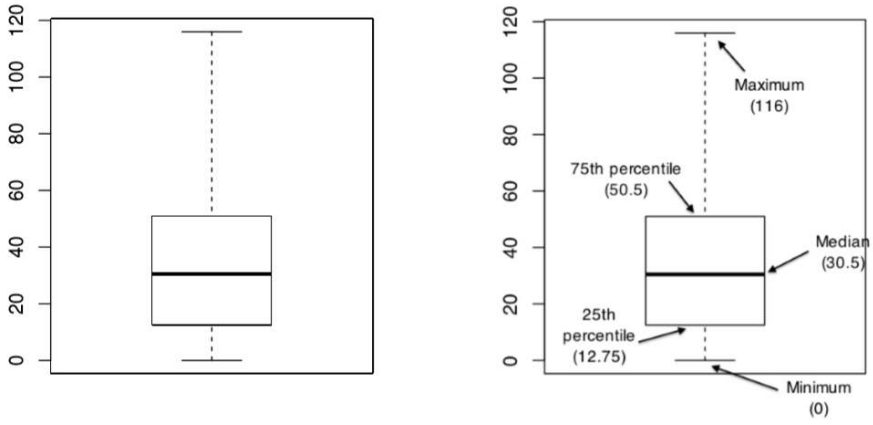


Figure **2.9.1**- A basic boxplot (panel a), plus the same plot with annotations added to explain what aspect of the data set each part of the boxplot corresponds to (panel b). (CC-BY-SA- Danielle Navarro from Learning Statistics with R)

In most applications, the “whiskers” don’t cover the full range from minimum to maximum. Instead, they actually go out to the most extreme data point that doesn’t exceed a certain bound. By default, this value is 1.5 times the interquartile range (you don't have to know what all of that means). Any observation whose value falls outside this range is plotted as a circle instead of being covered by the whiskers, and is commonly referred to as an *outlier*. Because the boxplot automatically separates out those observations that lie outside a certain range, people often use them as an informal method for detecting outliers: observations that are “suspiciously” distant from the rest of the data.

Our Final Exam Score data provides just such an example, as shown in Figure **2.9.2**, a boxplot created with SPSS software (standard spreadsheet software generally does not include boxplot options). If you look closely, there's a little circle under the bottom whisker with a "3" next to it. This is showing that three scores were below the range of the whiskers. These three scores can be considered outliers. If you go back to Figure 2.2.2, you can easily see what those three scores are (57, 66, 69).

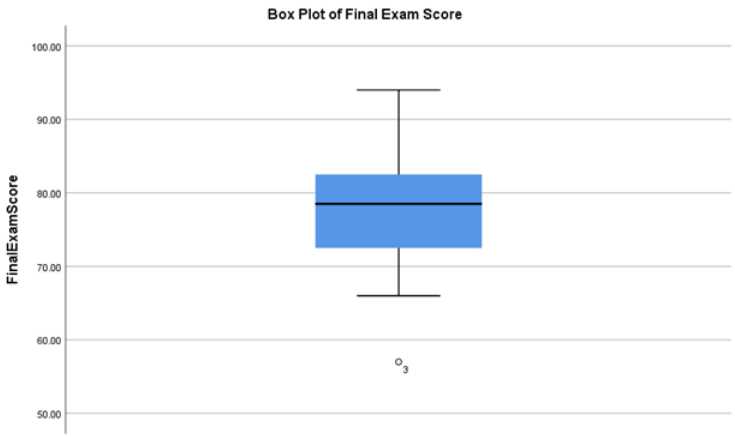


Figure 2.9.2- Boxplot of Frequency of Final Exam Scores. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

Box Plots Interpretation

Let's answer the same questions that we've been answering, over this same data set, but based on the boxplot graphic in Figure 2.9.2.

1. What kind of graph is Figure Figure 2.9.2?
    1. This box plot.
2. What does the x-axis measure in Figure 2.9.2?
    1. There is no x-axis because there is only one group. If there were more groups, each would have its own box and the name of the group would be labeled on the x-axis.
3. What does the y-axis measure in Figure 2.9.2?
    1. The y-axis is the axis that goes up and down. In Figure 2.9.2,the y-axis shows how many students earned different scores on the Final Exam by the size and location of the square in the middle. The “whiskers” or bars show information about the expected variation. You can also see a circle under the 60.00 line, with the number 3 next to it; this shows that 3 students scored that extreme, which is outside of what is expected.
4. Is Figure Figure 2.9.2? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
    1. This is sort of a trick question. Skew is not shown in a box plot in the same way that it is shown in a line graph or histogram. However, you can see that the box in the middle is not centered evenly between the whiskers; the top whisker is longer than the bottom whisker. This shows that the bulk of scores fall within the box, but that there are some that go higher and lower, and the circle shows that there are three scores that are really low.
5. What do you notice from Figure 2.9.2? What pops out to you?
    1. Those three scores at the bottom jump out at me! Plus, the fact that the top whisker doesn’t go all the way to 100 suggests that no one scored that high.
6. What does Figure 2.9.2 make you wonder about?
    1. This box plot makes me wonder about those three students outside of the whiskers more than the other charts did.
7. What is a catchy headline for Figure 2.9.2?
    1. Chances Are You’ll Be One of the 85% [Because 17 out of 20 students earned a passing grade on the Final Exam; 17/20 = .85 x 100 = 85%]
8. How could you summarize the info in Figure 2.9.2 into one sentence?
    1. The class did really well on the Final Exam, although three students did not pass.
9. Who might want to know the information in Figure 2.9.2?
    1. I still am guessing that students who are going to take this class from this professor might be interested. College administrators might also want to know so that they can see that most students are passing their classes.

That's it on boxplots! Let's move on to the last kind of graph that will be discussed, scatterplots!

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)

This page titled 2.9: Graphing Quantitative Data- Boxplots is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

2.10: Graphing Quantitative Data- Scatterplots

So far, we've mostly covered charts that show how many people were in each category or earned each score. Scatterplots are a different kind of chart that can show you about how many people scored each variable, but the real purpose is show a person's score on *two* quantitative variables (ratio or interval scale of measurement) at once. In scatterplots, the both the x-axis and the y-axis are scores on a quantitative variable, and there is no direct measurement of frequency. Figure **2.10.1** shows an example with GPA and the hours that the student studied (the GPA_Study_Hours file from OpenIntro.org). Each dot represents a student, and shows both their GPA and how many hours they studied. Find the dot that's most right; that student who spent 70 hours studying, and did not quite earn a 4.0. In fact, it looks like no one who studied over 35 hours earned a 4.0; what do you think that can tell you about time spent studying? On the other hand, only a few people who spent less than 20 hours studying earned a 4.0; what can *that* tell you about how the time you spend studying leads to your GPA?

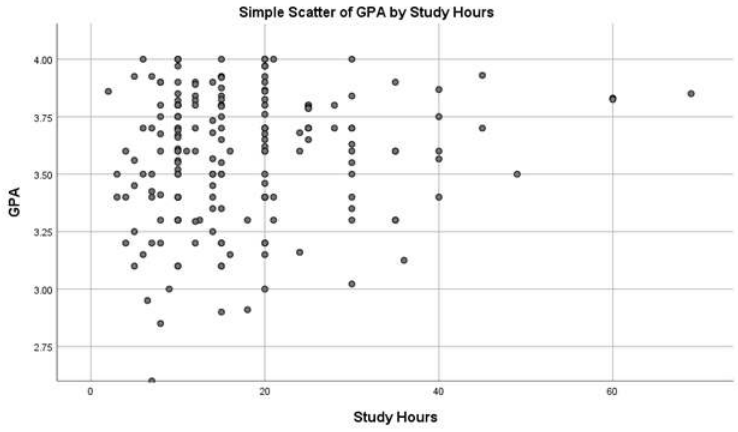


Figure **2.10.1**: Scatterplot of Study Hours and GPA. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

Remember IVs and DVs from Ch. 1? In scatterplots, the cause (IV) is usually plotted on the x-axis, and the effect (DV) is plotted on the y-axis. What we are looking for is: what happens to the DV when the IV increases? In Figure **2.10.1**, as Study Hours increase, GPA tends to increase. However, this isn't a strong relationship because there are a couple students who spend a lot of time studying and don't get the GPA payoff, while there are even more students who spend 5-10 hours studying but still earn good grades. In general, however, it looks like spending more time studying is related to higher GPAs.

Note

Use this Statistics How To to practice making a scatterplot. Once you can make one, with the IV on the bottom and the DV on the side, you'll better understand what they can show.

Scatterplot Interpretation

1. What kind of graph is Figure **2.10.1**?
    1. This is a scatterplot. Scatterplots are used to plot two quantitative variables for each person. In Figure **2.10.1**, the two variables are Study Hours and GPA.
2. What does the x-axis measure in Figure **2.10.1**?
    1. The x-axis in Figure **2.10.1** shows the hours that each person studied. Lower hours of studying are on the left, and higher hours of studying are on the right.
3. What does the y-axis measure in Figure **2.10.1**?
    1. The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. However, the y-axis of the scatterplot in Figure **2.10.1** shows another quantitative variable, GPA of each person. Lower GPAs are towards the bottom of the axis, and higher GPAs are towards the top.
4. Is Figure **2.10.1** skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?

1. Trick question! Because scatterplots do not show frequency distributions, they aren't described as skewed and they don't have kurtosis. Instead, we look to see if there is a general trend; as one variable goes up, does the other go up, go down, or are they unrelated? It does look like more studying is related to higher GPAs, but this isn't quite as strong of a relationship as we'd like to see.

5. What do you notice from Figure **2.10.1**? What pops out to you?
    1. What I first noticed is that a lot of students study 20 hours or less, and still have high GPAs. I also noticed that there are lots of scores in 5-hour increments (e.g., 5 hours, 10 hours, 15 hours, etc.). This suggests that students were estimated the number of hours that they studied, and rounding to the nearest 5-hour mark.
6. What does Figure **2.10.1** make you wonder about?
    1. I wonder if the results would change if the hours studied were measured more accurately, rather than having students estimate at a later time? I also wonder what was going with those two students who studied more than 60 hours (and still didn't earn a 4.0). What were they doing that they considered "studying"? Or is this a data-entry mistake? Like, did someone say that they studied 6 hours but it was entered as 60 hours?
7. What is a catchy headline for Figure **2.10.1**?
    1. Study Less and Still Get a 4.0!
8. How could you summarize the info in Figure **2.10.1** into one sentence?
    1. A lot of students studied about 5 to 20 hours and earned GPAs above 3.0.
9. Who might want to know the information in Figure **2.10.1**?
    1. I still am guessing that students who might be interested. Maybe tutors and folks working in a Learning Center would find this useful?

Remember those videos about the best way to learn from Dr. Chew, a cognitive psychologist? First discussed in 1.3: The Scientific Method and all 5 linked in 1.8 (What can you do to learn material for this class?). Well, Dr. Chew's first video has a scatter plot (Part 1: Beliefs that Make You Fail)! I suggest that you watch the whole video, but the experiment and scatterplot show up around 4:30 minutes.

? Exercise 2.10.1

Watch Dr. Chew's first video (Part 1: Beliefs that Make You Fail), then answer the same set of questions. I suggest that you watch the whole video, but the experiment and scatterplot show up around 4:30 minutes.

Answer

1. What kind of graph is in Dr. Chew's Beliefs that Make you Fail video?
    1. This is a scatterplot. Scatterplots are used to plot two quantitative variables for each person. In the scatterplot in the video, the two variables are Estimated Percentage Correct and Actual Percentage Correct on an exam.
2. What does the x-axis measure in the scatterplot in Dr. Chew's Beliefs that Make you Fail video?
    1. The x-axis in the scatterplot in Dr. Chew's Beliefs that Make you Fail video shows the percentage that students thought that they got correct on the exam. Lower scores are on the left, and higher percentages are on the right.
3. What does the y-axis measure iin the scatterplot in Dr. Chew's Beliefs that Make you Fail video?
    1. The y-axis is the axis that goes up and down. For most charts discussed in this textbook, the y-axis will be frequencies. However, the y-axis of the scatterplot in the scatterplot in Dr. Chew's Beliefs that Make you Fail video shows another quantitative variable, the actual percentage correct on the exam. Lower scores are towards the bottom of the axis, and higher percentages are towards the top.
4. Is the scatterplot in Dr. Chew's Beliefs that Make you Fail video skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
    1. Trick question! Because scatterplots do not show frequency distributions, they aren't described as skewed and they don't have kurtosis. Instead, we look to see if there is a general trend; as one variable goes up, does the other go up, go down, or are they unrelated? It does look like most students estimated their percentage correct pretty close to

their actual percentage correct, but were a few students who estimated that they did way better than they actually did. Dr. Chew says that these students have low meta-cognition, meaning that they are unclear on what material they actually know.

5. What do you notice from the scatterplot in Dr. Chew's Beliefs that Make you Fail video? What pops out to you?
    1. What I first noticed was that most students' estimates were pretty close to their actual percentages earned. Yay, most students had good meta-cognition!
6. What does the scatterplot in Dr. Chew's Beliefs that Make you Fail video make you wonder about?
    1. I wonder if students get better at estimating what they know throughout the semester?
7. What is a catchy headline for the scatterplot in Dr. Chew's Beliefs that Make you Fail video?
    1. Meta-Cognition: Predicting Your Future!
8. How could you summarize the info in the scatterplot in Dr. Chew's Beliefs that Make you Fail video into one sentence?
    1. Most students' estimates of their exam grades were relatively close to their actual exam grades, but there were a few students who badly overestimated their scores.
9. Who might want to know the information in the scatterplot in Dr. Chew's Beliefs that Make you Fail video?
    1. I still am guessing that students might be interested. Maybe tutors and folks working in a Learning Center would find this useful?

Positive or Negative?

Scatterplots can show when there is a linear relationship., meaning that the two variables vary together and when plotted look like a straight line. Scatterplots can show positive linear relationships, negative linear relationships, or show that there is no relationship between the two variables.

- Positive linear relationship: When one variable goes up, the other goes up.
    - Positive *doesn't* mean good!! It means that the two variables change in the **same** direction.
- Negative linear relationship: When one variable goes up, the other variable goes down.
    - Negative *doesn't* mean bad!!! It means that the two variables change in **different** directions.

Figure 2.10.2 shows a scatterplot for hypothetical scores on Job Satisfaction (x-axis) and Well-Being of the worker (on the y-axis). 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). This scale suggests an interval scale of measurement for both variables, which means that there are two quantitative variables.

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.

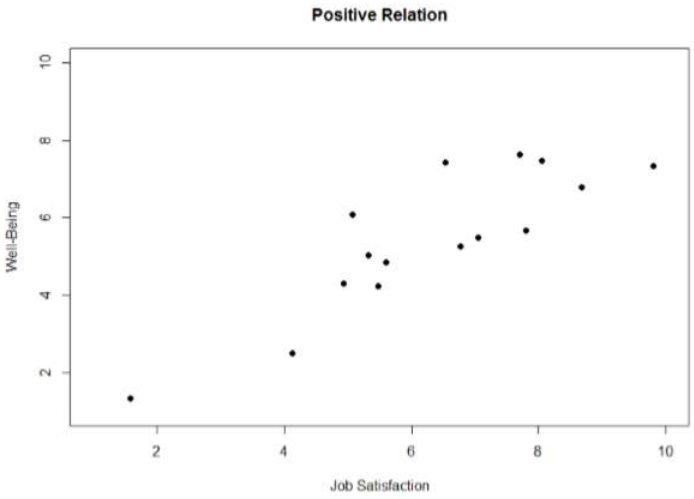


Figure 2.10.2- Scatterplot of Job Satifaction & Well-Being (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Figure 2.10.2 demonstrates a positive relation. As Job Satisfacation scores increase, Well-Being scores 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 2.10.3 shows a negative relation between Job Satisfacation on the x-axis and Burnout on the y-axis. As we can see from Figure 2.10.3, higher scores on Job Satisfacation tend to correspond to lower scores on Burnout, which is how stressed, un-energetic, and unhappy someone is at their job. As with Figure 2.10.2, 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.

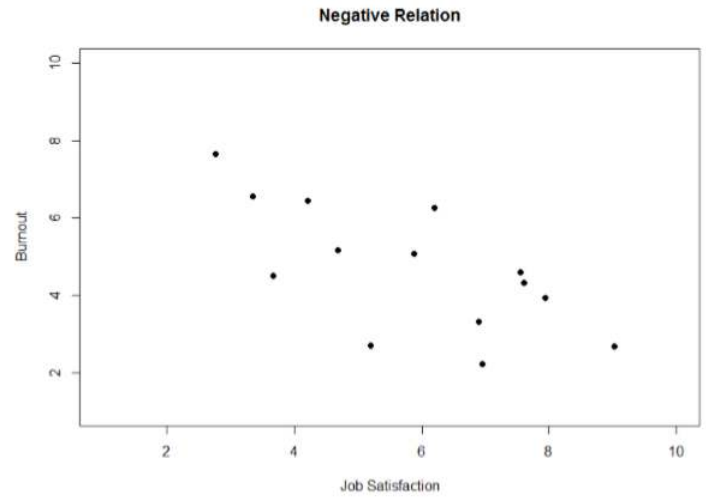


Figure 2.10.3- Scatterplot of Job Satisfaction and Burnout (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Or None?

Scatterplots can also indicate that there is no relation between the two variables. In these scatterplots (an example is shown below in Figure 2.10.4 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.

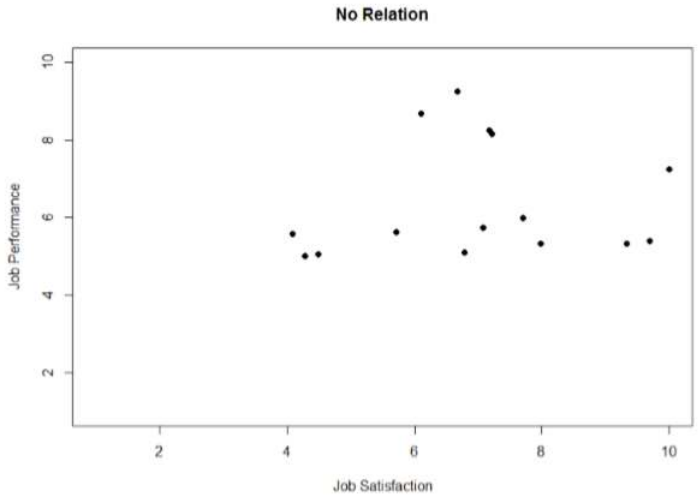


Figure 2.10.4- Scatterplot of Job Satisfaction and Job Performance (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Curvilinear Relationships

Sometimes the fastest way isn't a straight line...

A linear relation is what we saw in Figure 2.10.2 and Figure 2.10.3. 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”. The relation between two variables can also be curvilinear. As the name suggests, a curvilinear relation is one in which a line through the middle of the points will be curved rather than straight. Two examples are presented in Figures 2.10.5 and 2.10.6.

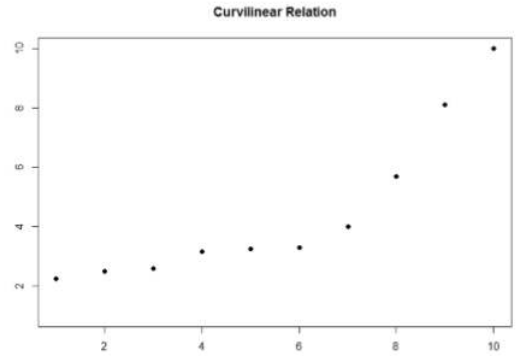


Figure 2.10.5: Curvilinear Scatterplot with an Elbow (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

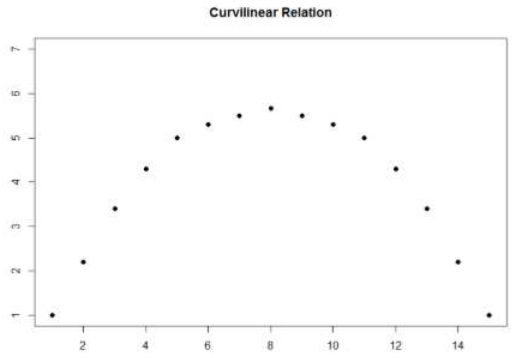


Figure 2.10.6: Curvilinear Scatterplot with an Inverted-U (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics).

Curvilinear relations can take many shapes, and the two examples above are only a small sample of the possibilities. What they have in common is that they both have a very clear pattern but that pattern is not a straight 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 outliers, which we discussed before.

We'll return to scatterplots in Chapters 14 and 15 to discuss ways to statistically analyze this relationship to see if what the scatterplot looks like it shows is really what is happening.

Contributors and Attributions

- Foster et al. (University of Missouri-St. Louis, Rice University, & University of Houston, Downtown Campus)
- Dr. MO (Taft College)

This page titled 2.10: Graphing Quantitative Data- Scatterplots is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

2.11: Summary and Some Honesty

Intentional Lies or Accidental Misleading?

It's easy to mislead with graphs. Some of the fancy graphics make it more difficult to see the exact numbers, but there are many ways to be misled with graphs. To avoid being fooled by a fancy chart you should also check out the axes. In a frequency distribution, the y-axis should generally be one or zero. This article "Lying with Graphs" (website address: http://chmullig.com/2011/04/lying-with-graphs/) by Chris Mulligan in 2011 shows an example of both a confusing type of chart, and mis-leading axes (plus, he didn't think that the original article calculated the actual numbers correctly, either); always be vigilant!

Example of Misleading

We will use data from interviews of computer users to show how graphs in general, but especially bar graphs, can be misleading. Here's the research scenario: 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? Was it purchased by newcomers to the computer market> Were previous Windows users 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.

Figure **2.11.1** shows that three-dimensional bars are usually not as effective as two-dimensional bars because it's unclear where the top of the bar is.

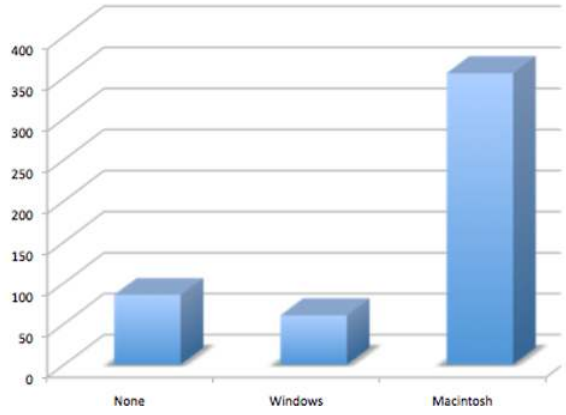


Figure **2.11.1**: A three-dimensional bar chart (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure **2.11.2** presents iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure **2.11.2** is misleading because the viewer's attention will be captured by the area (width) of the computers, not just the height. 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 **2.11.2** 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 **2.11.2**! 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.

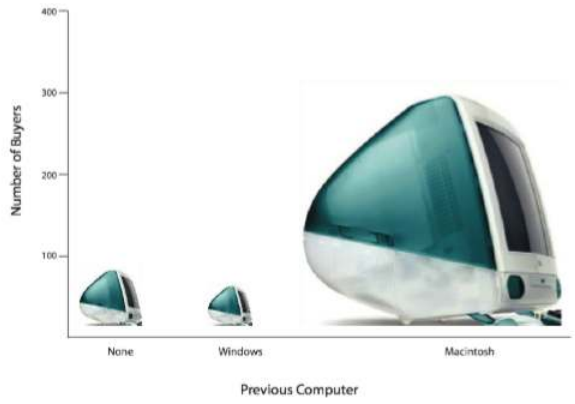


Figure 2.11.2: Example with a lie factor greater than 8. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Another distortion in bar charts results from setting the bottom of the y-axis to a something other than zero. The bottom of the Y-axis should represent the least number of cases that could have occurred in a category. Normally, this number should be zero. Figure 2.11.3 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%.

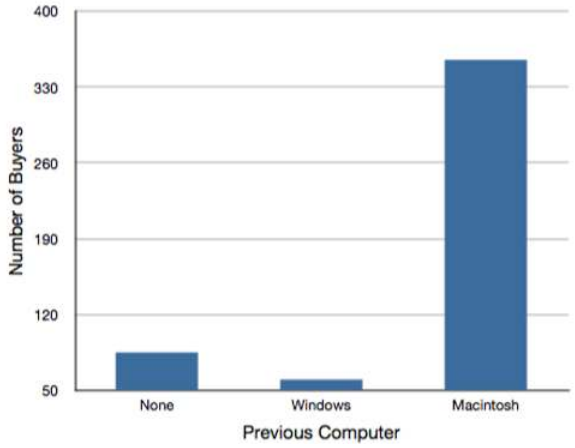


Figure 2.11.3: Example with a baseline of 50. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Note

Why don't you use this iMac data to play around and create your own bar chart by hand or with computer software, without making it misleading?

Examples of Misleading, or Lies?

Example 2.11.1

Go through the examples from this 2014 BuzzFeed report by Katie Notopoulos (https://www.buzzfeednews.com/article/katienotopoulos/graphs-that-lied-to-us), and decide one thing that is wrong with each chart.

1. The time an upside down y-axis made "Stand Your Ground" seem much more reasonable.
2. The time 7 million was 5x more than 6 million.
3. The Govenor race where one guy's 37% was WAY more than just 37%
4. This bar graph that shows the devastating drop in this pitcher's speed after one year.
5. The time when Scotland really gave 110%.

6. This poll which gives about the same infuriating response as when you ask someone what they want for dinner.

7. The pollster who really doesn't want people to think A levels are getting harder.

8. This graph that measure..... units of "innovation"?

9. The number skipping y-axis on this totally legit graph.

10. This graph showing the astonishing growth rate to .06 hawks?

11. Whatever happened in this school paper.

12. The graph where last year, last week, and today are equally far apart.

13. This chart showing the giant gulf between 35% and 39.6%.

Solution

1. Check out the y-axis. Instead of zero on the bottom, with the higher numbers being higher, the y-axis is “upside down” from what is typical.

2. Check out the y-axis. Not only is it not labeled, but if the y-axis had started with zero, the 1 million difference would have looked like much.

3. The two 37%’s probably look difference chart doesn’t show decimal points, but if the y-axis had started with zero, the two 37%’s would be indistinguishable.

4. It’s the missing y-axis problem again, and not starting the y-axis at zero.

5. There’s nothing technically wrong with these numbers, but you should always have someone else check your math.

6. There’s nothing technically wrong with these numbers, but it is a reminder that designing survey questions is harder than it seems!

7. This is an appropriate way to use a pie chart (showing percentages that add up to close to 100%), but the pie slices need to represent the percentage, so the blue should only be 49% of the pie. It looks like 60% or about two-thirds…

8. It’s important to label your y-axis with units that make sense!

9. The y-axis should have equal intervals, meaning that the difference between the first number and the second number (2 points in this example, from 1 to 3) is the same as the difference between the highest number and the second highest number (which is 1803-516 = 1287 in this example).

10. There is nothing technically wrong with this chart, but even going from zero to 1.0 would have seemed more accurate when dealing with such small numbers on the y-axis.

11. The colors in the pie charts should take up the portion of the pie that the numbers show. Also, pie charts show 100% of the sample, but 26% + 26% + 26% +26% = 104%, so someone needed a friend to check their math here, as well.

12. Here is an example of a line graph showing changes through time, except that the x-axis doesn’t have equal intervals. It should probably have been a bar chart showing the three time categories, rather than trying to show a change through time.

13. Another example of a missing y-axis that is misleading because the missing y-axis doesn’t start at zero.

You can also read this great 2014 blog by Ravi Parikh of Heap showing some of the same examples of graphs that seem to be trying to intentionally mislead readers, and an explanation about what is inappropriate or missing: https://heap.io/blog/data-stories/how-to-lie-with-data-visualization

Be careful to avoid misleading your audience, or being misled!

Graphs Summary

You've learned about many types of graphs, and why they should be used. You also should have practiced interpreting what they can tell you.

Example 2.11.1

Play around with this interactive website (website address: https://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm), then answer what kind of graphs they are.

Solution

Although both have lines to show a Normal Distribution, the actual graphs of the data are histograms because they are bars that touch, and the bars show the frequency of different quantitative variables.

Now that you've learned so much about charts and graphs, go through these charts from the New York Times (website address: https://www.nytimes.com/column/whats-going-on-in-this-graph) and determine if the x-axis is showing qualitative or quantitative variables and if you can describe the shape of any frequency distributions.

Contributors and Attributions

- Foster et al. (University of Missouri-St. Louis, Rice University, & University of Houston, Downtown Campus)
- Dr. MO (Taft College)

This page titled 2.11: Summary and Some Honesty 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.

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)

3: Measures of Central Tendency and Spread is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

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.

	8	05
	7	156
	6	233
	5	168
330	4	06
9420	3	
622	2	

Figure 3.1.1: Back-to-back stem and leaf display. 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.

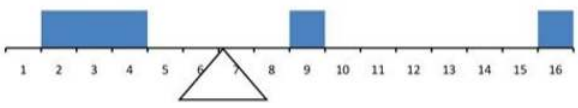


Figure 3.1.2: A balance scale.

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

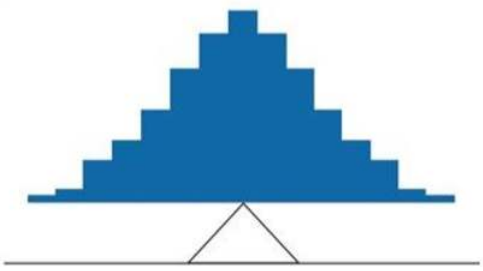


Figure 3.1.3: A distribution balanced on the tip of a triangle.

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

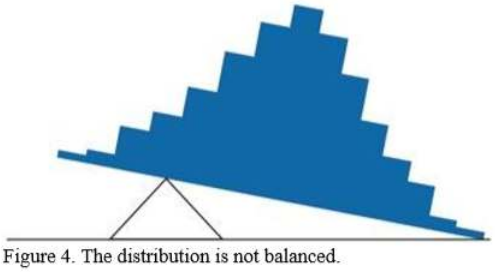


Figure 3.1.4: The distribution is not balanced.

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.

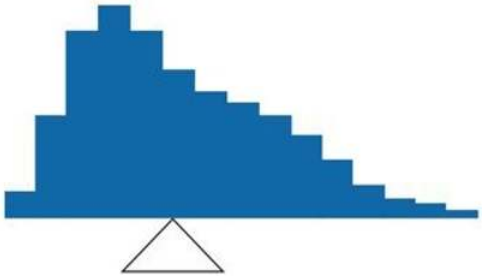


Figure **3.1.5** An asymmetric distribution balanced on the tip of a triangle.

Smallest Absolute Deviation Another way to define the center of a distribution is based on the concept of the sum of the absolute deviations (differences). Consider the distribution made up of the five numbers 2, 3, 4, 9, 16. Let's see how far the distribution is from 10 (picking a number arbitrarily). Table **3.1.2** shows the sum of the absolute deviations of these numbers from the number 10.

Table **3.1.2**: An example of the sum of absolute deviations

Values	Absolute Deviations from 10
2	8
3	7
4	6
9	1
16	6
Sum	28

The first row of the table shows that the absolute value of the difference between 2 and 10 is 8; the second row shows that the absolute difference between 3 and 10 is 7, and similarly for the other rows. When we add up the five absolute deviations, we get 28. So, the sum of the absolute deviations from 10 is 28. Likewise, the sum of the absolute deviations from 5 equals 3 + 2 + 1 + 4 + 11 = 21. So, the sum of the absolute deviations from 5 is smaller than the sum of the absolute deviations from 10. In this sense, 5 is closer, overall, to the other numbers than is 10.

We are now in a position to define a second measure of central tendency, this time in terms of absolute deviations. Specifically, according to our second definition, the center of a distribution is the number for which the sum of the absolute deviations is smallest. As we just saw, the sum of the absolute deviations from 10 is 28 and the sum of the absolute deviations from 5 is 21. Is there a value for which the sum of the absolute deviations is even smaller than 21? Yes. For these data, there is a value for which the sum of absolute deviations is only 20. See if you can find it.

Smallest Squared Deviation

We shall discuss one more way to define the center of a distribution. It is based on the concept of the sum of squared deviations (differences). Again, consider the distribution of the five numbers 2, 3, 4, 9, 16. Table **3.1.3** shows the sum of the squared deviations of these numbers from the number 10.

Table **3.1.3**: An example of the sum of squared deviations

Values	Squared Deviations from 10
2	64
3	49
4	36
9	1
16	36
Sum	186

The first row in the table shows that the squared value of the difference between 2 and 10 is 64; the second row shows that the squared difference between 3 and 10 is 49, and so forth. When we add up all these squared deviations, we get 186. Changing the target from 10 to 5, we calculate the sum of the squared deviations from 5 as 9 + 4 + 1 + 16 + 121 = 151. So, the sum of the squared deviations from 5 is smaller than the sum of the squared deviations from 10. Is there a value for which the sum of the squared deviations is even smaller than 151? Yes, it is possible to reach 134.8. Can you find the target number for which the sum of squared deviations is 134.8?

The target that minimizes the sum of squared deviations provides another useful definition of central tendency (the last one to be discussed in this section). It can be challenging to find the value that minimizes this sum.

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

In the section “What is central tendency,” we saw that the center of a distribution could be defined three ways:

1. the point on which a distribution would balance
2. the value whose average absolute deviation from all the other values is minimized
3. the value whose squared difference from all the other values is minimized.

The mean is the point on which a distribution would balance, the median is the value that minimizes the sum of absolute deviations, and the mean is the value that minimizes the sum of the squared deviations.

Table **3.2.2** shows the absolute and squared deviations of the numbers 2, 3, 4, 9, and 16 from their median of 4 and their mean of 6.8. You can see that the sum of absolute deviations from the median (20) is smaller than the sum of absolute deviations from the mean (22.8). On the other hand, the sum of squared deviations from the median (174) is larger than the sum of squared deviations from the mean (134.8).

Table **3.2.2**: Absolute & squared deviations from the median of 4 and the mean of 6.8.

Value	Absolute Deviation from Median	Absolute Deviation from Mean	Squared Deviation from Median	Squared Deviation from Mean
2	2	4.8	4	23.04
3	1	3.8	1	14.44
4	0	2.8	0	7.84
9	5	2.2	25	4.84
16	12	9.2	144	84.64
Total	20	22.8	174	134.8

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.

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.

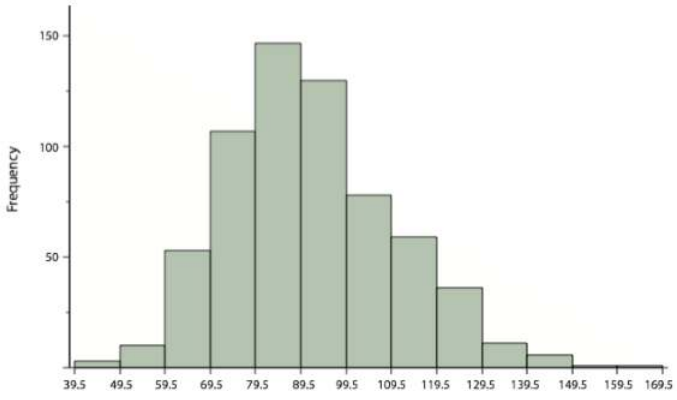


Figure **3.2.3**: A distribution with a positive skew.

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**.

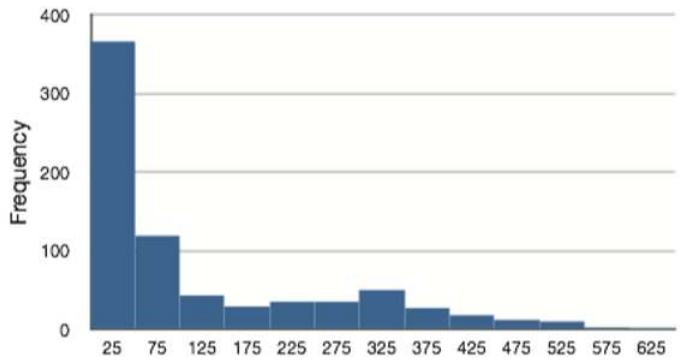


Figure **3.2.3**: A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars).

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.

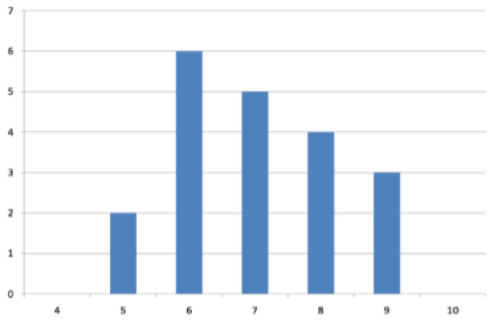


Figure **3.3.1**: Bar chart of quiz one.

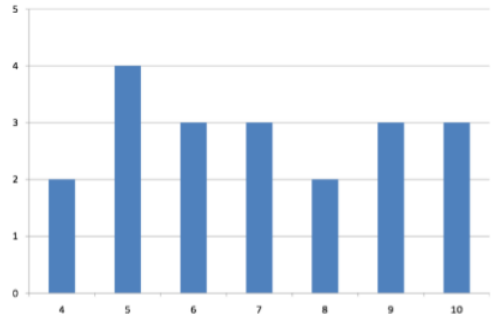


Figure **3.3.2**: Bar chart of quiz two.

The terms variability, spread, and dispersion are synonyms, and refer to how spread out a distribution is. Just as in the section on central tendency where we discussed measures of the center of a distribution of scores, in this chapter we will discuss measures of the variability of a distribution. There are three frequently used measures of variability: range, variance, and standard deviation. In the next few paragraphs, we will look at each of these measures of variability in more detail.

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.

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. Recall that in the discussion of box plots, the 75th percentile was called the upper hinge and the 25th percentile was called the lower hinge. Using this terminology, the interquartile range is referred to as the H-spread.

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.

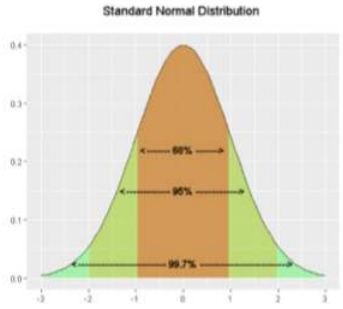


Figure **3.3.3**: Percentages of the normal distribution

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.

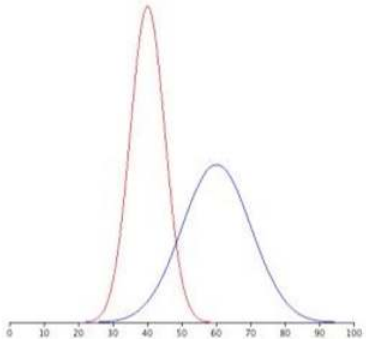


Figure **3.3.4**: Normal distributions with standard deviations of 5 and 10.

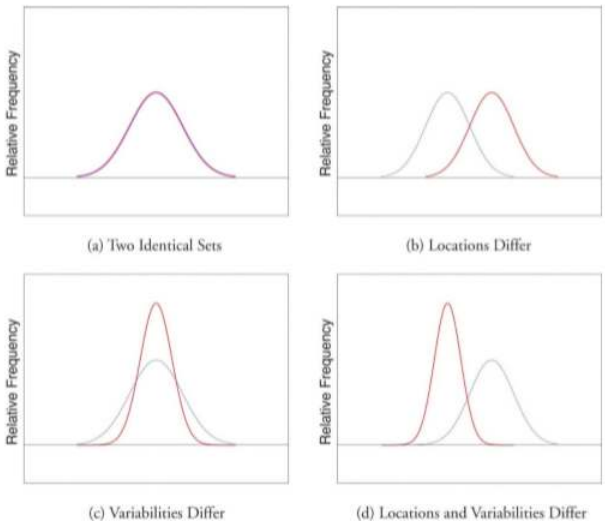


Figure **3.3.5**: Differences between two datasets.

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: Z-scores and the Standard Normal Distribution

4.1: Normal Distributions
4.2: Introduction to z-scores
4.3: Calculating z-scores
4.4: Introduction to the z table
4.5: Z table of critical values
4.6: Predicting Amounts
4.7: Summary of z Scores
4.8: The Write-Up

4: Z-scores and the Standard Normal Distribution is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

4.1: 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 **4.1.1**).

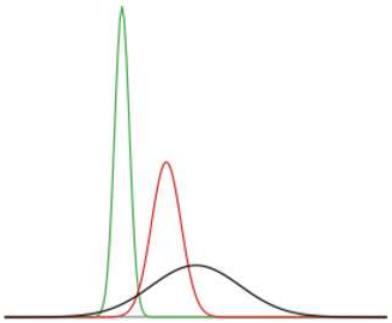


Figure **4.1.1**: Normal distributions differing in mean and standard deviation.

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 4.1: 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.

4.2: Introduction to z-scores

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 **4.2.1** 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.

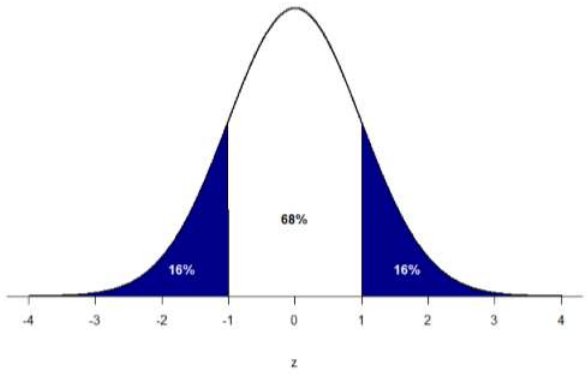


Figure **4.2.1**: Shaded areas represent the area under the curve in the tails

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 4.2: Introduction to 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.

- **4.3: 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.

4.3: Calculating 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 from a sample into a z-score is:

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

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.

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. 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).

Practice

Raw Score to z-score

We can 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.

$$z = \frac{x - \overline{X}}{s} = \frac{68 - 54}{8} = 1.75$$

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!

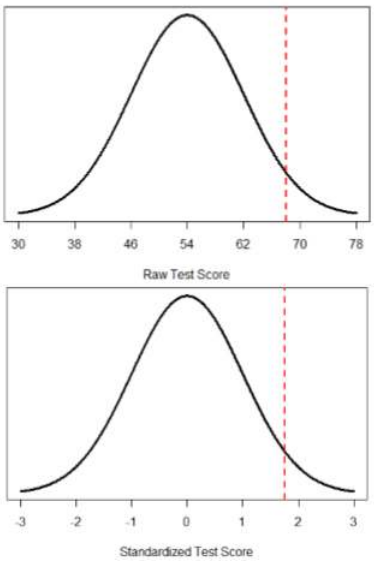


Figure 4.3.1: Raw and standardized versions of a single score (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Figure 4.3.1 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 4.3.1:

Table 4.3.1: 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 4.3.2.

Table 4.3.2: 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

z-score to Raw Score

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 formula for transforming z to x for a sample is:

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

Notice that this is just a rearrangement of the original formulas for calculating z from raw scores. Dr. MO prefers to just use the original z-score formula and do algebra to figure out the raw score when she has a z-score because she doesn't want to have to remember more formulas, but others prefer to do less algebra and have more formulas; it's your choice!

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:

$$z = \frac{52 - 40}{7} = 1.71$$
$$z = \frac{43 - 40}{7} = 0.43$$
$$z = \frac{34 - 40}{7} = -0.80$$

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.

Contributors and Attributions

- Foster et al. (University of Missouri-St. Louis, Rice University, & University of Houston, Downtown Campus)
- Dr. MO (Taft College)

This page titled 4.3: Calculating z-scores is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

4.4: Introduction to the z table

To introduce the table of critical z-scores, we'll first refresh and add to what you learned last chapter about distributions

Probability Distributions and Normal Distributions

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 standard deviations from the mean. These lines marked specific z-scores. These sections between the marked lines have specific probabilities of scores falling in these areas under the normal curve.

First, let's look back at the area between z = -1.00 and z = 1.00 presented in Figure **4.4.1**. 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 $z = 1.00$ (within one standard deviation below and one standard deviation above the mean) because those are the z-scores that satisfy our criteria.

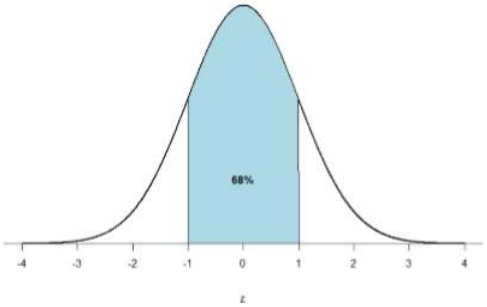


Figure **4.4.1**: There is a 68% chance of selection a z-score from the blue-shaded region (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Take a look at the normal distribution in Figure **4.4.2** 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.

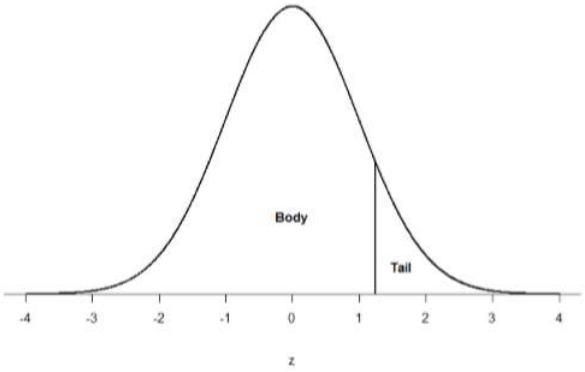


Figure **4.4.2**: Body and tail of the normal distribution (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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, but we don't need to know how to do all that! The exact proportions for 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.

Let's look at an example: 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). Dr. Foster didn't just pick this z-score out of nowhere, but we'll get to that later. Let's find the area corresponding to the region illustrated in Figure **4.4.3**, which corresponds to the area more extreme than z = -1.96 and z = 1.96.

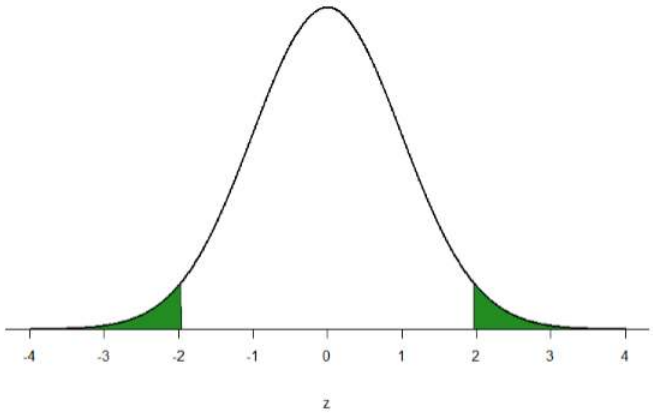


Figure 4.4.3: Area in the tails beyond z = -1.96 and z = 1.96 (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

If we go to the z-table shown in the Critical Values of z Table page (which can also be found from the Common Critical Value Tables at the end of this book in the Back Matter with the glossary and index), we will see one column header that has a z, bidirectional arrows, and then p. This means that, for the entire table (all 14ish columns), there are really two columns (or sub-columns. The numbers on the left (starting with -3.00 and ending with 3.00) are z-scores. The numbers on the right (starting with .00135 and ending with .99865) are probabilities (p-values). So, if you multiply the p-values by 100, you get a percentage.

Let's start with the tail for z = 1.96. What p-value corresponds to 1.96 from the z-table in Table 4.4.1?

Example 4.4.1

What p-value corresponds to 1.96 from the z-table in Table 4.4.1?

Solution

For z = 1.96, p = .97500

If we multiply that by 100, that means that 97.50% of the scores in this distribution will be *below* this score. Look at Figure 4.4.3 again. This is saying that 97.5 % of scores are outside of the shaded area on the right. That means that 2.5% of scores in a normal distribution will be higher than this score (100% - 97.50% = 2.50%). In other words, the probability of a raw score being higher than a z-score is p=.025.

If do the same thing with |(z = -1.96|), we find that the p-value for $z = -1.96$ is .025. That means that 2.5% of raw scores should be below a z-score of -1.96; according to Figure 4.4.3, that is the shaded area on the left side. What did we just learn? That the shaded areas for the same z-score (negative or positive) are the same p-value, the same probability. We can also find the total probabilities of a score being in the two shaded regions by simply adding 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 later). And, because we know that z-scores are really just standard deviations, this means that it is very unlikely (probability of 5%) to get a score that is almost two standard deviations away from the mean (-1.96 below the mean or 1.96 above the mean).

Attributions & Contributors

- Foster et al. (University of Missouri-St. Louis, Rice University, & University of Houston, Downtown Campus)
- Dr. MO (Taft College)

This page titled 4.4: Introduction to the z table is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

4.5: Z table of critical values

4.5: Z table of critical values is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

4.6: Predicting Amounts

You might see how z-scores are useful for determining probabilities of scores, but you can use the z-score, probability, and the characteristics of a standard normal distribution so predict how many people could earn lower (or higher) than scores. To show how, we'll use Final Exam scores (scores on a 100-point final exam from 20 students from an unknown college provided by OpenIntro.org). The average Final Exam score 77.7 points, with a standard deviation of 8.44 points.

Let's start with some context about who and what we are looking at.

Example 4.6.1

1. For the Final Exam scores, who is the sample? *Your answer should include the number of people in the sample.*

2. Based on this sample, who could be the population?

3. What was measured?

4. What information was provided?

a. N: _____

b. Mean: _____ points

c. Standard deviation: _____

Solution

Add text here.

1. 20 students from an unknown college

2. Anything reasonable (probably college students).

3. Final Exam scores (on a 100-point exam)

4. What information was provided?

a. N: 20

b. Mean: 77.70 points

c. Standard deviation: 8.44

The Research Question will be: ***How many students passed the Final Exam (earned a score of 70 points or higher)?*** And we'll use z-scores and the z-table to figure that out! Let's walk through each step.

Example 4.6.2

1. What's the formula for computing a z-score?

2. Compute the z-score for X (70 points, from the Research Question).

3. Find the p-value from the z Table for that z-score.

4. To answer the Research Question, do we need to subtract this from 1 (or 18.141% from 100%)? Why or why not? If so, do it.

5. You're not done yet! The Research Question asks *how many students*, not what proportion (percentage).

6. Write a concluding sentence for the Research Question. Round to the nearest whole person.

Solution

1. What's the formula for computing a z-score?

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

2. Compute the z-score for X (70 points, from the Research Question).

$$z = \frac{x - \overline{X}}{s} = \frac{70 - 77.70}{8.44} = \frac{-7.7}{8.44} = -0.91$$

3. z = -0.91 so p = 0.18141

4. Yes, we do need to subtract because the p-value is telling us that 18.141% of the Final Exam scores are *lower* than our 70 points (z=-0.91), but the Research Question asked how many students scored *higher* then 70 points.

$$1 – 0.18141 = 0.81859 (or 100\% - 18.14\% = 81.86\%)$$

5. We have 20 students (N = 20). To find the amount of students, we multiply the N by the proportion (before it is a percentage):

$$Amount = 20 \times 0.81859 = 16.37$$

6. Based on the mean, standard deviation, and size of this sample, 16 of the 20 students should pass the Final Exam (earned a score of 70 points or higher).

Your turn! We'll use the same sample of Final Exam scores (scores on a 100-point final exam from 20 students from an unknown college, with the average Final Exam score of 77.7 points and the standard deviation of 8.44 points. The Research Question will be: ***How many students earned 90 points or higher on the Final Exam?***

? Exercise 4.6.1

1. Compute the z-score for X (90 points, from the Research Question).
2. Find the p-value from the z Table for that z-score.
3. To answer the Research Question, do we need to subtract this from 1 (or 100%)? Why or why not? If so, do it.
4. You're not done yet! The Research Question asks *how many students*, not what proportion (percentage).
5. Write a concluding sentence for the Research Question. Round to the nearest whole person.

Answer

1. Compute the z-score for X (90 points, from the Research Question).

$$z = \frac{x - \overline{X}}{s} = \frac{90 - 77.70}{8.44} = \frac{12.3}{8.44} = 1.46$$

2. z=1.46 so p = 0.92785

3. Yes, we do need to subtract because the p-value is telling us that 92.79% of the Final Exam scores are *lower* than 90 points (z = 1.46), but the Research Question asked how many students scored *higher* then 90 points.

$$1 – 0.92785 = 0.07215 (or 100\% - 92.79\% = 7.21\%)$$

4. We have 20 students (N = 20). To find the amount of students, we multiply the N by the proportion (before it is a percentage):

$$Amount = 20 \times 0.07215 = 1.44$$

5. Based on the mean, standard deviation, and size of this sample, 1 of the 20 students should earn 90 points or higher on the Final Exam.

We're Getting There!

The Standard Normal Curve and z-scores are a small piece of the puzzle to get to where we're heading, to statistically comparing sample means. It's okay if you're not quite on board with z-scores yet, we'll keep working on the concept! We use z-scores to compare one sample to compare with another sample. It's okay if you're still confused about the Standard Normal Curve, too. These are building blocks to get us to the exciting statistical comparisons that will be coming up.

This page titled 4.6: Predicting Amounts is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

4.7: Summary of z Scores

Although this section has some math and the z-score formula, it's really to hit the main points of z-scores.

Summary of z-scores

Suppose Dr. Navarro's friend is putting together a new questionnaire intended to measure "grumpiness". The survey has 50 questions, which you can answer in a grumpy way or not. Across a big sample (hypothetically, let's imagine a million people or so!) the data are fairly normally distributed, with the mean grumpiness score being 17 out of 50 questions answered in a grumpy way, and the standard deviation is 5. In contrast, when Dr. Navarro takes the questionnaire, she answer 35 out of 50 questions in a grumpy way. So, how grumpy is she? One way to think about would be to say that I have grumpiness of 35/50, so you might say that she's 70% grumpy. But that's a bit weird, when you think about it. If Dr. Navarro's friend had phrased her questions a bit differently, people might have answered them in a different way, so the overall distribution of answers could easily move up or down depending on the precise way in which the questions were asked. So, I'm only 70% grumpy *with respect to this set of survey questions*. Even if it's a very good questionnaire, this isn't very a informative statement.

Can we standardize?

A simpler way around this is to describe Dr. Navarro's grumpiness by comparing me to other people. Shockingly, out of the friend's sample of 1,000,000 people, only 159 people were as grumpy as Dr. Navarro (which Dr. Navarro believes is realistic), suggesting that she's in the top **0.016%** of people for grumpiness. This makes much more sense than trying to interpret the raw data. This idea – that we should describe a person's grumpiness in terms of the overall distribution of the grumpiness of humans – is the idea that standardization attempts to get at. One way to do this is to do exactly what I just did, and describe everything in terms of percentiles. However, the problem with doing this is that "it's lonely at the top". Suppose that the friend had only collected a sample of 1000 people (still a pretty big sample for the purposes of testing a new questionnaire, I'd like to add), and this time gotten a mean of 16 out of 50 with a standard deviation of 5, let's say. The problem is that almost certainly, not a single person in that sample would be as grumpy as me.

However, all is not lost. A different approach is to convert my grumpiness score into a ***standard score***, also referred to as a z-score. The standard score is defined as the number of standard deviations above the mean that my grumpiness score lies. To phrase it in "pseudo-math" the standard score is calculated like this:

$$\text{standard score} = \frac{\text{raw score} - \text{mean}}{\text{standard deviation}}$$

In actual math, the equation for the z-score is:

$$z = \frac{X - \bar{X}}{s}$$

So, going back to the grumpiness data, we can now transform Dr. Navarro's raw grumpiness into a standardized grumpiness score. If the mean is 17 and the standard deviation is 5 then her standardized grumpiness score would be

$$z = \frac{X - \bar{X}}{s} = \frac{35 - 17}{5} = \frac{18}{5} = 3.6$$

To interpret this value, recall that 99.7% of values are expected to lie within 3 standard deviations of the mean. So the fact that my grumpiness corresponds to a z score of 3.6 (3.6 standard deviations above the mean) indicates that I'm very grumpy indeed.

So now that we know that we can compare one person's individual raw score to the whole distribution of scores, what else can the z-score do?

Compare Across Distributions

In addition to allowing you to interpret a raw score in relation to a larger population (and thereby allowing you to make sense of variables that lie on arbitrary scales), standard scores serve a second useful function. Standard scores can be compared to one another in situations where the raw scores can't. Suppose, for instance, my friend also had another questionnaire that measured extraversion using a 24 items questionnaire. The overall mean for this measure turns out to be 13 with standard deviation 4; and I scored a 2. As you can imagine, it doesn't make a lot of sense to try to compare my raw score of 2 on the extraversion questionnaire to my raw score of 35 on the grumpiness questionnaire. The raw scores for the two variables are about fundamentally different

things, so this would be like comparing apples to oranges. BUT, if we calculate the z-scores, we get $z = \frac{(35-17)}{5} = 3.6$ for grumpiness and $z = \frac{(2-13)}{4} = -2.75$ for extraversion. These two numbers *can* be compared to each other. Dr. Navarro is much less extraverted than most people ($z = -2.75$) (the negative sign means that Dr. Navarro is 2.75 standard deviations below the average extraversion of the sample) and much grumpier than most people (z=3.6): but the extent of my unusualness is much more extreme for grumpiness (since 3.6 is a bigger number than 2.75). Because each standardized score is a statement about where an observation falls *relative to its own population*, it *is* possible to compare standardized scores across completely different variables.

This ability to compare scores across different distributions is the foundation of statistical analyses. It allows us to compare the sample that we have to a probability sample, and then make predictions!

Proportions & Amounts

But that's not all! Using this z-score formula to turn a raw score into a standardized z-score, we can then use what we know about the proportions of scores in standard normal curves to predict how likely that score was. And if we know the sample size, we can even predict how many people from the sample will score above (or below) that initial raw score. Amazing!

This all is pretty cool, although you probably don't think so yet. Or maybe never will. But the way that the Standard Normal Distribution and these standardized scores (z-scores) function allows much of the rest of statistics to exist.

Next: Write-Ups

You are in a social science class, not a statistics class, so finding the number is almost never the end of your task. The next section will describe the important components of a good concluding sentence or concluding paragraph.

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)

This page titled 4.7: Summary of z Scores is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Danielle Navarro.

4.8: The Write-Up

Through the practice examples, I hope that you have realized that when conducting statistics for the social sciences, the answer is never just the number. We do the statistics to answer questions, to the final answer needs enough information to answer that question, and to let other statisticians know a little bit about the sample and the calculations. Based on what we've learned so far, here's what you might include in a concluding sentence, as well as what should be included in a paper describing a distribution.

Concluding Sentence

For any conclusion, you should include the results of your calculations, what was measured, and the answer to the original research question. Sometimes, this might be as simple as:

- Research Question: What is the average final exam score?
- Conclusion: The average final exam score was 77.7 points.

The Research Question from Exercise 5.4.1 was:

- Research Question: How many students earned 90 points or higher on the Final Exam?

So the Conclusion should be:

- Conclusion: Based on the mean, standard deviation, and size of this sample, 1 of the 20 students should earn 90 points or higher on the Final Exam.

The important pieces of information to include in these concluding sentences are the research question (rephrase as an answer), the calculation results, and what was measured. For Exercise 5.4.1, those are:

1. Research Question: "How many students earned 90 points or higher on the Final Exam?" was turned in to "1 student should earn 90 points or higher on the Final Exam."
2. Calculation: This is the "1 of the 20 students" part of the conclusion.
3. What was measured: This is your DV, your outcome variable. In this Exercise, it was points earned on the Final Exam.

For something a little more advanced you will need to include more information. We'll cover that a little later!

Paper Describing a Distribution

For a full paper to describe a distribution, you will combine the conclusion for everything that we've covered so far. This should include:

1. Describing who and what was measured. This should include:
    1. Naming the sample. Who provided the data? How many participants were there?
    2. Naming who you think the population could be. In other words, name who is the biggest group that the sample can represent?
    3. Naming what was measured (quantitative DV).
2. Interpreting what the measures of central tendency mean.
    1. Make sure that you calculate the mean, median, and mode correctly!
    2. To interpret the measures of central tendency, describe what does knowing the mean, median, and mode *collectively* tell you? Maybe answering these questions will help: Are the mean, median, and mode similar? What could that tell us about the shape of the distribution? Is one smaller or bigger? What could that tell us about the shape of the distribution? Are they all very different? What could that tell us about the shape of the distribution?
3. Interpreting what the standard deviation can tell us.
    1. Make sure that you calculated the standard deviation correctly!
    2. An interpretation of standard deviation should include:
        1. An evaluation of whether one standard deviation above and below the mean really includes about 68% of the scores, like it should if the data was normally distributed.

2. Use the standard deviation to predict the shape of the distribution (tall/narrow, medium/normal, or wide/flat), then compare the predicted shape to the actual shape.
3. If the standard deviation seems large, you would expect a platykurtic distribution (wide and flat). For example, if you look at your frequency chart, does the shape seem wide and flat? Small samples are often not normally distributed, so what we might expect based on the standard deviation is not the actual shape of the distribution. Plus, outliers can cause skewed shapes and large standard deviations. The standard deviation gives us a general idea of how different each score is from the mean, but there's nothing better than looking at the actual distribution.

4. Providing and describing/interpreting appropriate frequency charts.
    1. Don't forget to format the chart number and title in appropriate APA Style!
    2. You should mention the chart (by Figure by number) in the paper body.
    3. Just like we did for each different types of chart, you should describe/interpret the frequency chart by saying something about what you see or what the chart makes you wonder.

And don't forget that a paper in your statistics class is still a paper! You should have an introduction with some sort of hook (Why is this topic interesting?), the body (which should include everything above), and a concluding paragraph. Concluding paragraphs often include why this topic is important (which may refer back to the hook from the introduction), or who would want to know this information. A one- or two-sentence summary of what was found could also be included, but don't get too hung up on that.

The next chapter will discuss how to format this paper describing a data set in APA Style.

This page titled 4.8: The Write-Up is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

CHAPTER OVERVIEW

5: Distributions

5.1: Introduction to Distributions
5.2: What is Probability
5.3: The Law of Large Numbers
5.4: Normal Distributions and Probability Distributions
5.5: Putting it All Together
5.6: Summary- The Bigger Picture

5: Distributions is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

5.1: Introduction to Distributions

Refresher

Let's remind ourselves three important ideas we learned about last chapter: distribution, central tendency, and variance. These terms are similar to their everyday meanings (although I suspect most people don't say central tendency very often).

Distribution

When you order something from Amazon, where does it come from, and how does it get to your place? That stuff comes from one of Amazon's distribution centers. They distribute all sorts of things by spreading them around to your doorstep. "To Distribute"" is to spread something. Notice, the data in the histogram is distributed, or spread across the bins. We can also talk about a distribution as a noun. The histogram is a distribution of the frequency counts across the bins. Distributions are **very, very, very, very, very** important. They can have many different shapes.

We will start this chapter on distributions talking about frequency distributions, which show the frequency of scores. When you hear "distribution," "distribution of data," or "frequency distribution," you should be imagining a histogram or line graph (frequency polygon).

- The X-axis has all possible scores.
- The Y-axis has frequencies of occurrences.

As we learn about other types of distributions, the Y-axis might be measured differently (like in percentages or probability), but the idea is the same: How often did that score occur?

Measures of Central Tendency

Central Tendency is all about sameness: What is common about some numbers? For example, if we had a distribution in which most of scores were near 0, we could say that there is a tendency for most of the numbers to be centered near 0. Notice we are being cautious about our generalization about the numbers. We are not saying they are all 0. We are saying there is a tendency for many of them to be near zero. There are lots of ways to talk about the central tendency of some numbers. There can even be more than one kind of tendency. For example, if lots of the numbers were around -1000, and a similar large amount of numbers were grouped around 1000, we could say there was two tendencies. The three common measures of central tendency that we discussed last chapter are:

- Mode
- Median
- Mean

? Exercise 5.1.1

Which measures of central tendency can be used with quantitative data?

Answer

Add texts here. Do not delete this text first. All three measures of central tendency can be used with quantitative data. The mode is the only measure of central tendency that can be used with qualitative varaibles, though.

Measures of Variance

Variance (or variability) is all about differentness: What is different about some numbers? For example, is there anything different about all of the numbers in the histogram? YES!!! The numbers are not all the same! When the numbers are not all the same, they must vary. So, the variance in the numbers refers to how the numbers are different. There are many ways to summarize the amount of variance in the numbers, (see Measures of Varability).

The most common measure of variance that we will use is standard deviation. Some researchers also like to know the range, or the maximum and minimum scores (highest score and lowest score).

Why?

As we begin talking about different kinds of special distributions, you might be asking yourself why these matter. We will go through some of these special distributions to show why statistical analyses can be used to predict information about a population from information about a sample. This will, ultimately, let us test research hypotheses. I could explain why, but it wouldn't make sense because you don't know about the special distributions yet! So, for now, let's learn the basics about some of these special distributions so that you can interpret your results later.

Let's begin!

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 5.1: Introduction to Distributions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew J. C. Crump via source content that was edited to the style and standards of the LibreTexts platform.

- **2.3: Important Ideas - Distribution, Central Tendency, and Variance** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

5.2: 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:

$$\text{Probability of an event } = \frac{\text{number of outcomes that satisfy our criteria}}{\text{total number of possible outcomes}} \tag{5.2.1}$$

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

$$P(A) = \frac{\text{number of outcomes that count a } A}{\text{total number of possible outcomes}} \tag{5.2.2}$$

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 5.2: 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.

5.3: The Law of Large Numbers

If we want our sample statistics to be much closer to the population parameters, what can we do about it?

The answer is to collect more data. Larger samples are a much better approximation to the true population distribution than smaller samples. I feel a bit silly saying this, but the thing I want you to take away from this is that large samples generally give you better information. It does feel a bit obvious that more data will give you better answers.

The question is, why is this so? Not surprisingly, this intuition that we all share turns out to be correct, and statisticians refer to it as the ***law of large numbers***. The law of large numbers is a mathematical law that applies to many different sample statistics, but the simplest way to think about it is as a law about averages. The sample mean is the most obvious example of a statistic that relies on averaging (because that's what the mean is… an average), so let's look at that. When applied to the sample mean, what the law of large numbers states is that as the sample gets larger, the sample mean tends to get closer to the true population mean. Or, to say it a little bit more precisely, as the sample size "approaches" infinity (written as $N \rightarrow \infty$) the sample mean approaches the population mean ($\bar{X} \rightarrow \mu$).

I don't intend to subject you to a proof that the law of large numbers is true, but it's one of the most important tools for statistical theory. The law of large numbers is the thing we can use to justify our belief that collecting more and more data will eventually lead us to the truth. For any particular data set, the sample statistics that we calculate from it will be wrong, but the law of large numbers tells us that if we keep collecting more data those sample statistics will tend to get closer and closer to the true population parameters.

It's okay if this isn't obvious to you. I'm just mentioning it because it will start to feel obvious, and explains lots about why statistical formulas are the way they are. I mean, we've already run across that in how the formula for the standard deviation of a sample ($\sqrt{\frac{\sum(X-\bar{X})^2}{N-1}}$) is different from the formula for the standard deviation of the population ($\sqrt{\frac{\sum(X-\bar{X})^2}{N}}$).

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)

This page titled 5.3: The Law of Large Numbers is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Danielle Navarro.

5.4: Normal Distributions and Probability 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

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.

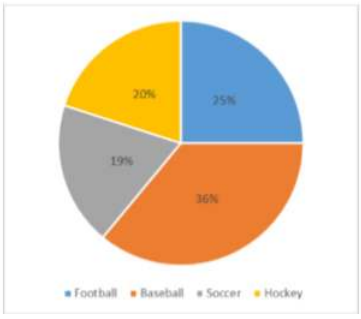


Figure 5.4.1: Favorite Sports (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Take, for example, the pie chart in Figure 5.4.1 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: $36\% + 25\% + 20\% = 81\%$. 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.

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 5.4.1 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 5.4.1).

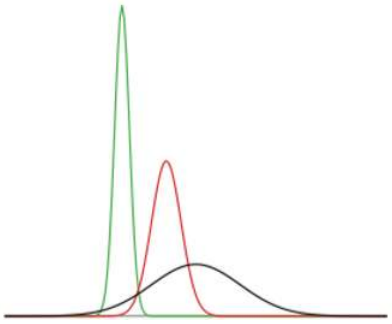


Figure 5.4.1: Symmetrical Distributions with Different Means and Standard Deviations (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Standard Normal Distribution

Important 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. In Standard Normal Curves, the mean, median, and mode are all 0.
3. The area under the normal curve is equal to 1.0 (or 100% of all scores will fall somewhere in the distribution).
4. Normal distributions are denser in the center and less dense in the tails (bell-shaped).
5. There are known proportions of scores between the mean and each standard deviation.
    1. One standard deviation- 68% of the area of a normal distribution is within one standard deviation of the mean (one standard deviation below the mean through one standard deviation above the mean).
    2. Two standard deviations- Approximately 95% of the area of a normal distribution is within two standard deviations of the mean (two standard deviations below the mean through two standard deviations above the mean).

These properties enable us to use the normal distribution to understand how scores relate to one another within and across a distribution.

Probability in Normal Distributions

Play with this applet to learn about probability distributions (website address: http://www.rossmanchance.com/applets/OneProp/OneProp.htm?candy=1)

Probability distributions are what make statistical analyses work. Each distribution of events from any sample can be “translated” into a probability distribution, then used to predict the probability of that event happening!

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. We know that one standard deviation below the mean to one standard deviation above the mean contains 68% of the area under the curve because we know the properties of standard normal curves.

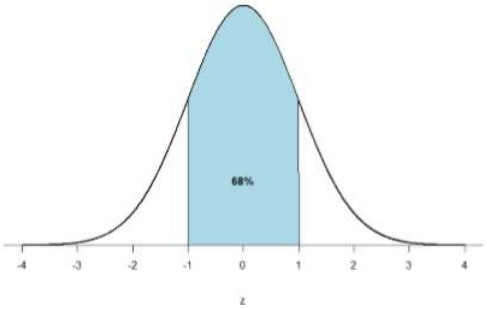


Figure 5.4.2: There is a 68% chance of selection a Score from the shaded region (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

This time, let’s find the area corresponding to the extreme tails. For standard normal curves, 95% of scores should be in the middle section of Figure 5.4.3, and a total of 5% in the shaded areas. Since there are two shaded areas, that's about 2.5% on each side.

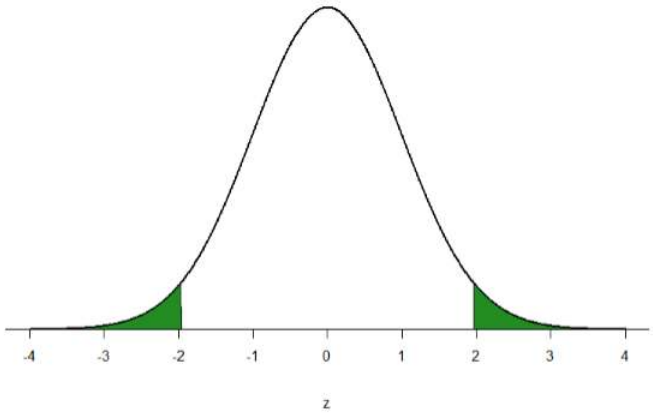


Figure 5.4.3: Area Beyond 2 Standard Deviations (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Those important characteristics of a Standard Normal Distribution are deceptively simple, so let's put some of these ideas together to see what we get.

Law of Large Numbers: I know through the Law of Large Numbers that if I have enough data, then my sample will begin to become more and more "normal' (have the characteristics of a normally distributed population, including the important characteristics above).

Probability Distributions: I know that I can make predictions from probability distributions. I can use probability distributions to understand the likelihood of specific events happening.

Standard Normal Curve: I know that there should be specific and predictable proportions (percentages) of scores between the mean and each standard deviation away from the mean.

Together, these features will allow you to:

- Predict the probability of events happening.
- Estimate how many people might be affected based on sample size.
- Test Research Hypotheses about different groups.

This is the magic of the Standard Normal Curve!

Non-Parametric Distributions

One last thing...

There's a lot of heavy lifting for the Standard Normal Distribution. But what if your sample, or even worse, the population, is not normally distributed? That's when non-parametric statistics come in.

A parameter is a statistic that describes the population. Non-parametric statistics don’t require the population data to be normally distributed. Most of the analyses that we'll conduct compare means (a measure of central tendency) of different groups. But if the data are not normally distributed, then we can’t compare means because there is no center! Non-normal distributions may occur when there are:

- Few people (small N)
- Extreme scores (outliers)
- There’s an arbitrary cut-off point on the scale. (Like if a survey asked for ages, but then just said, “17 and below”.)

The reason that I bring this up is because sometimes you just can't run the statistics that we're going to learn about because the population's data is not normally distributed. Since the Standard Normal Curve is the basis for most types of statistical inferences, you can't use these (parametric) analyses with data that they don't fit. We'll talk about alternative analyses for some of the common (parametric) statistical analyses, but you heard it here first! Non-parametric distributions are when the population is not normally distributed.

Okay, onward to learning more about how the Standard Normal Curve is important...

This page titled 5.4: Normal Distributions and Probability Distributions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

5.5: Putting it All Together

So, we've spend a lot of time going over different distributions. Why? Because they can answer research questions!

Standard Normal Curve

This is where it all comes together! What we can start doing with these distributions is *comparing them*. This might not sound like much, but it's the foundation of statistics, and allows us to answer research questions and test hypotheses.

- Probability Distributions: I know that I can make predictions from probability distributions. I can use probability distributions to understand the likelihood of specific events happening.
- Law of Large Numbers: I know through the Law of Large Numbers that enough of samples, their scores will be normally distributed (with all of the important characteristics that includes).
- Central Limit Theorem: I know through the Central Limit Theorem that if I get the means of enough samples, that the sampling distribution of means will be normally distributed.
- Standard Normal Distributions: I know that the mean is converted to always be zero, and the standard deviation is standardized to always be 1 in Standard Normal Distributions.

In Figure 5.5.1, the x-axis corresponds to the value of some variable, and the y-axis tells us something about how likely we are to observe that value. Notice that the y-axis is labeled “Probability Density” and not “Probability”. There is a subtle and somewhat frustrating characteristic of continuous distributions that makes the y-axis behave a bit oddly: the height of the curve here isn’t actually the probability of observing a particular x value. On the other hand, it *is* true that the heights of the curve tells you which x values are more likely (the higher ones!). This will be discussed in the Continuous Variable section.

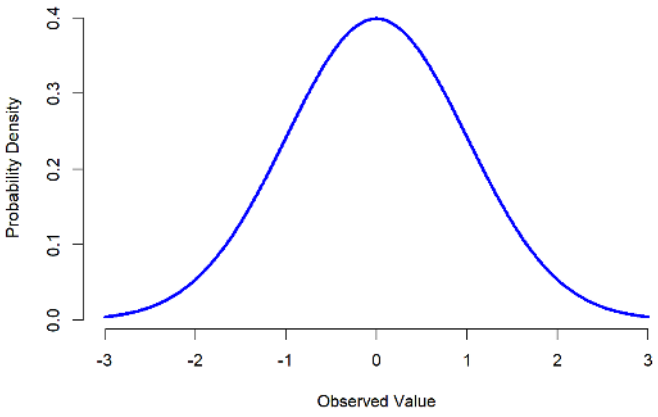


Figure 5.5.1- Standard Normal Curve (CC-BY-SA Danielle Navarro from Learning Statistics with R)

You can see where the name “bell curve” comes from in Figure 5.5.1 it looks a bit like a bell. Notice that, unlike the plots that I drew to illustrate the binomial distribution, the picture of the normal distribution in Figure 5.5.1 shows a smooth curve instead of “histogram-like” bars. This isn’t an arbitrary choice: the normal distribution is continuous, whereas the binomial is discrete. For instance, in the die rolling example, it was possible to get 3 skulls or 4 skulls, but impossible to get 3.9 skulls. The figures that I drew in the previous section reflected this fact. Continuous quantities don’t have this constraint. For instance, suppose we’re talking about the weather. The temperature on a pleasant Spring day could be 23 degrees, 24 degrees, 23.9 degrees, or anything in between since temperature is a continuous variable, and so a normal distribution might be quite appropriate for describing Spring temperatures.

- Normal Distributions: I know that there are predictable portions of scores between the mean and standard deviation.

The last important piece is learning that a special feature of normal distributions is that we know the proportions (percentages) of cases that should (probability) fall within each standard deviation around the mean. Irrespective of what the actual mean and standard deviation are, 68.3% of the area falls within 1 standard deviation of the mean. Similarly, 95.4% of the distribution falls within 2 standard deviations of the mean, and 99.7% of the distribution is within 3 standard deviations. This idea is illustrated in the follow Figures.

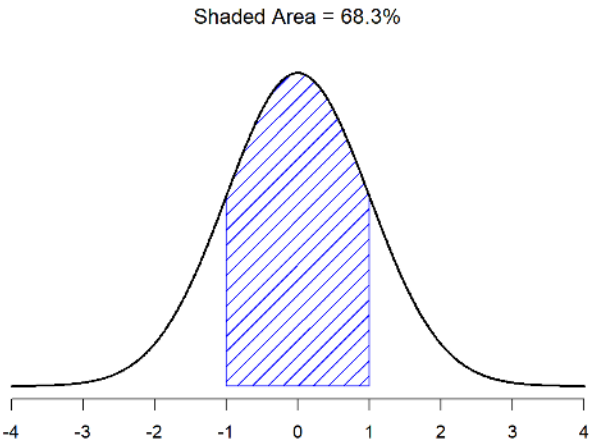


Figure **5.5.2**- About 68% of Scores are One Standard Deviation Below through One Standard Deviation Above the Mean (CC-BY-SA Danielle Navarro from Learning Statistics with R)

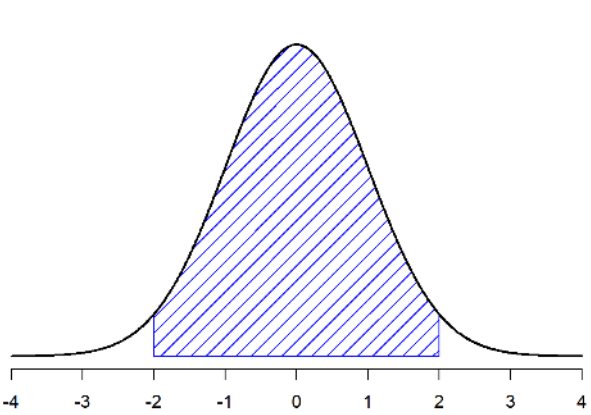


Figure **5.5.3**- About 95% of Scores are Two Standard Deviations Below through Two Standard Deviations Above the Mean (CC-BY-SA Danielle Navarro from Learning Statistics with R)

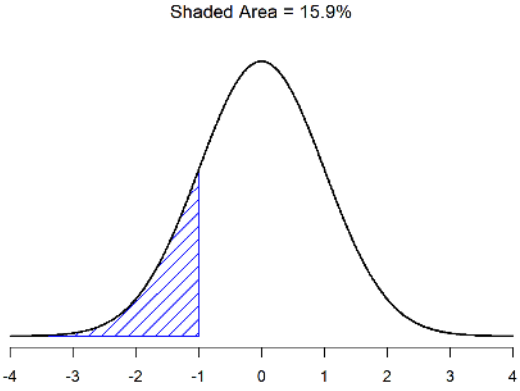


Figure **5.5.4**- About 16% of Scores Should Be *Less Than* Two Standard Deviations Below the Mean (CC-BY-SA Danielle Navarro from Learning Statistics with R)

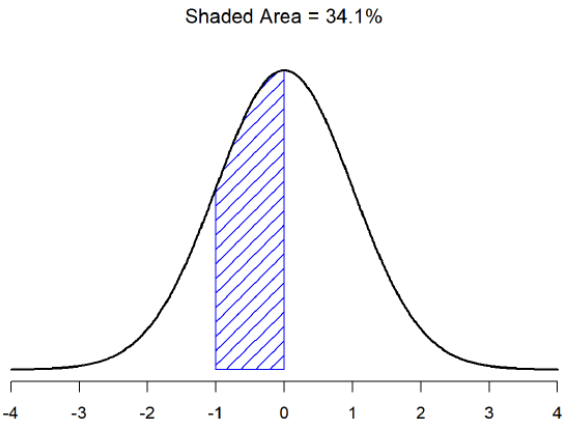


Figure 5.5.5- About 34% of Scores Should Be One Standard Deviations Below the Mean (CC-BY-SA Danielle Navarro from Learning Statistics with R)

There is a 68.3% chance that a score from any sample will fall within the shaded area of Figure 5.5.2, and a 95.4% chance that an observation will fall within two standard deviations of the mean (the shaded area in Figure 5.5.3). Similarly, there is a 15.9% chance that an observation will be below one standard deviation below the mean. There is a 34.1% chance that the observation is greater than one standard deviation below the mean but still below the mean (Figure 5.5.5). Notice that if you add these two numbers together you get 15.9+34.1=50. *For normally distributed data, there is a 50% chance that an observation falls below the mean*. And of course that also implies that there is a 50% chance that it falls above the mean.

Continuous Variables

There's something Dr. Navarro was trying to hide throughout this discussion of the normal distribution, but she just couldn't do it. Many introductory textbooks omit this completely; they might be right to do so: this "thing" that I'm hiding is weird and counterintuitive even by the admittedly distorted standards that apply in statistics. Fortunately, it's not something that you need to understand at a deep level in order to do basic statistics: rather, it's something that starts to become important later on when you move beyond the basics. So, if it doesn't make complete sense, don't worry: I'm mostly going over it to help think about the differences between qualitative and quantitative variables, and how quantitative variables can be discrete or continuous (even though we act like they are continuous).

Throughout the discussion of the normal distribution, there's been one or two things that don't quite make sense. Perhaps you noticed that the y-axis in these figures is labeled "Probability Density" rather than density. Let's spend a little time thinking about what it really *means* to say that X is a continuous variable. Let's say we're talking about the temperature outside. The thermometer tells me it's 23 degrees, but I know that's not really true. It's not *exactly* 23 degrees. Maybe it's 23.1 degrees, I think to myself. But I know that that's not really true either, because it might actually be 23.09 degrees. But, I know that… well, you get the idea. The tricky thing with genuinely continuous quantities is that you never really know exactly what they are.

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

The point of this discussion is to make clear that, when we're talking about continuous distributions, it's not meaningful to talk about the probability of a specific value. However, what we *can* talk about is the probability that the value lies within a particular range of values. To find out the probability associated with a particular range, what you need to do is calculate the "area under the

curve". We've seen this concept already: in Figure 5.5.2, the shaded areas shown depict genuine probabilities (e.g., in Figure 5.5.2 it shows the probability of observing a value that falls within 1 standard deviation of the mean).

Okay, so that explains part of the story. I've explained a little bit about how continuous probability distributions should be interpreted (i.e., area under the curve is the key thing). In terms of the plots we've been drawing, probability density corresponds to the *height* of the curve. The densities themselves aren't meaningful in and of themselves: but they're "rigged" to ensure that the *area* under the curve is always interpretable as genuine probabilities. To be honest, that's about as much as you really need to know for now.

We are about to finish up learning abstract theory about distributions, and moving on to actually using them!

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)

This page titled 5.5: Putting it All Together is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

5.6: Summary- 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 standard deviations around the mean. The probability of randomly getting a score in the specified region can then be found. 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 5.6: Summary- 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.

CHAPTER OVERVIEW

6: Inferential Statistics and an Introduction to Hypothesis Testing

6.1: Growth Mindset
6.2: People, Samples, and Populations
6.3: The Sampling Distribution of Sample Means
6.4: Using Standard Error for Probability
6.5: Sampling Distribution, Probability and Inference
6.6: Logic and Purpose of Hypothesis Testing
6.7: The Probability Value
6.8: Descriptive versus Inferential Statistics
6.9: The Null Hypothesis
6.10: The Alternative Hypothesis
6.11: Null Hypothesis Significance Testing
6.12: Critical Values, p-values, and Significance
6.12.1: Critical Values
6.12.2: Summary of p-values and NHST
6.13: Steps of the Hypothesis Testing Process
6.14: Movie Popcorn
6.15: Effect Size
6.16: Office Temperature
6.17: The Two Errors in Null Hypothesis Significance Testing
6.17.1: Power and Sample Size
6.17.2: The p-value of a Test

6: Inferential Statistics and an Introduction to Hypothesis Testing is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

6.1: Growth Mindset

In the first set of chapters, when we learned about descriptive statistics, some examples and data sets were based on academics and ways to be better students (remember Dr. Chew's videos?). In this next set of chapters we'll focus on a specific idea that can help you succeed: growth mindset.

Growth Mindset

Dr. Carol Dweck described mindset as the way that we approach learning and finding solutions (Dweck, 2006) in her book *Mindset: The New Psychology of Success*.

Growth mindset is when people realize that there are many good ways to find solutions, and that you can always learn more and become better. According to Dweck's research, if you keep trying new solutions, and learn from your "mistakes," you will improve. On the opposite side, *fixed mindset* is when people assume that everyone only has a specific amount of talent or ability, and that there's only one "best way" to accomplish tasks. There is a lot more research showing that growth mindset leads to academic success, and explains some of the reasons why that's true, at the science section of the growth mindset website called MindsetWorks.com.

This concept of growth and fixed mindset has real-world consequences for learning. For example, Cimpian, Arce, Markman, and Dweck (2007) found that young children who were told that they "did a good job drawing" (growth mindset: focus on effort) were happier and more likely to choose to draw than students who were told that they were "good drawers" (fixed mindset: focus on inherent ability). Similarly, Mueller and Dweck (1998) also found that children whose effort was praised valued learning more than those children who were praised based on their intelligence.

> **Note**
>
> How can knowing about growth mindset help you succeed in this class?

As we move back into understanding new concepts, and move away from the math and numbers for a bit, continue to think about what you are doing that is helping you learn, and what other study strategies you can use to do even better!

References

Cimpian, A., Arce, H-M. C., Markman, E. M., & Dweck, C. S. (2007). Subtle linguistic cues affect children's motivation. *Psychology Science, 18*(4), 314-316.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random House.

Mueller, C. M. & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology, 75*(1), 33-52.

This page titled 6.1: Growth Mindset is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.2: 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 6.2: People, Samples, and Populations is shared under a CC BY-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.

6.3: 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:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \tag{6.3.1}$$

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 **6.3.1** displays the principles stated here in graphical form.

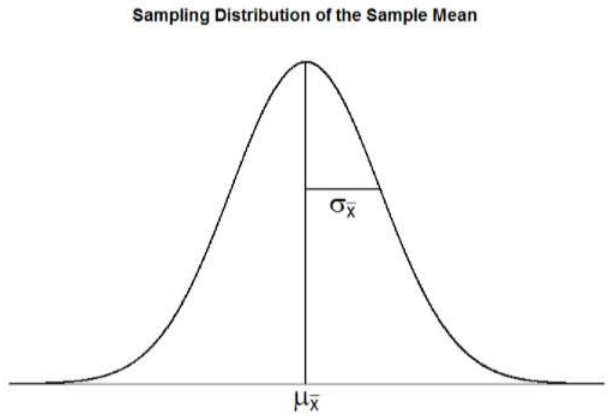


Figure **6.3.1**: The sampling distribution of sample means

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 6.3.1

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 **6.3.2** and **6.3.3**.

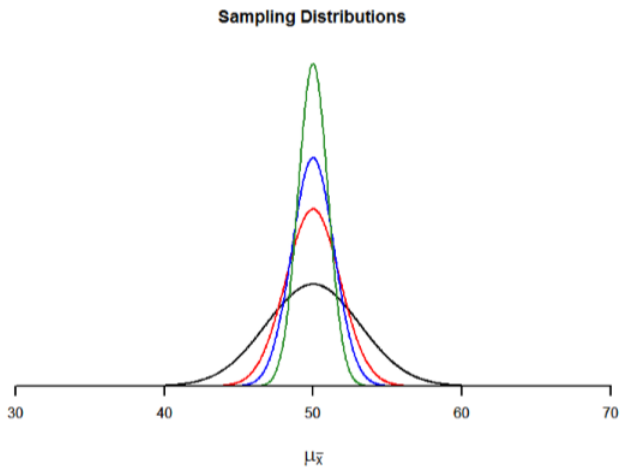


Figure **6.3.2**: Sampling distributions from the same population with μ = 50 and σ = 10 but different sample sizes (N = 10, N = 30, N = 50, N = 100)

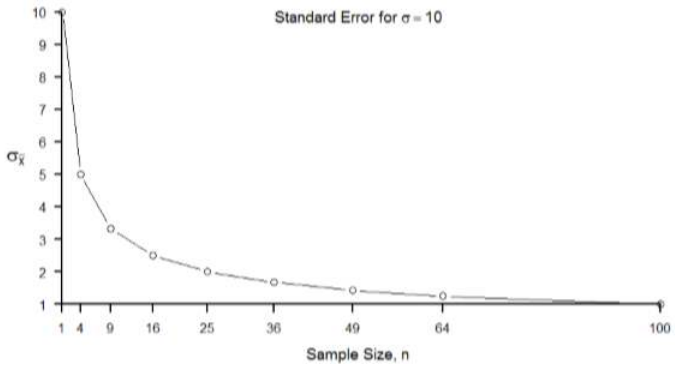


Figure **6.3.3**: Relation between sample size and standard error for a constant $\sigma = 10$

This page titled 6.3: 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.

6.4: Using Standard Error for Probability

We saw in chapter 6 that we can use z-scores to split up a normal distribution and calculate the proportion of the area under the curve in one of the new regions, giving us the probability of randomly selecting a z-score in that range. We can follow the exact sample process for sample means, converting them into z-scores and calculating probabilities. The only difference is that instead of dividing a raw score by the standard deviation, we divide the sample mean by the standard error.

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{6.4.1}$$

Let's say we are drawing samples from a population with a mean of 50 and standard deviation of 10 (the same values used in Figure 2). What is the probability that we get a random sample of size 10 with a mean greater than or equal to 55? That is, for n = 10, what is the probability that $\overline{X} \geq 55$? First, we need to convert this sample mean score into a z-score:

$$z = \frac{55 - 50}{\frac{10}{\sqrt{10}}} = \frac{5}{3.16} = 1.58$$

Now we need to shade the area under the normal curve corresponding to scores greater than z = 1.58 as in Figure **6.4.1**:

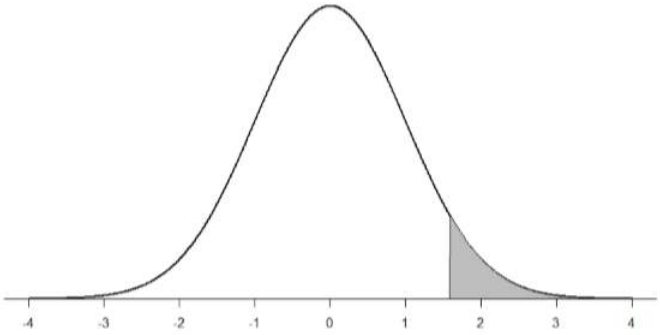


Figure **6.4.1**: Area under the curve greater than z = 1.58

Now we go to our z-table and find that the area to the left of z = 1.58 is 0.9429. Finally, because we need the area to the right (per our shaded diagram), we simply subtract this from 1 to get 1.00 – 0.9429 = 0.0571. So, the probability of randomly drawing a sample of 10 people from a population with a mean of 50 and standard deviation of 10 whose sample mean is 55 or more is p = .0571, or 5.71%. Notice that we are talking about means that are 55 or more. That is because, strictly speaking, it's impossible to calculate the probability of a score taking on exactly 1 value since the "shaded region" would just be a line with no area to calculate.

Now let's do the same thing, but assume that instead of only having a sample of 10 people we took a sample of 50 people. First, we find z:

$$z = \frac{55 - 50}{\frac{10}{\sqrt{50}}} = \frac{5}{1.41} = 3.55 \tag{6.4.2}$$

Then we shade the appropriate region of the normal distribution:

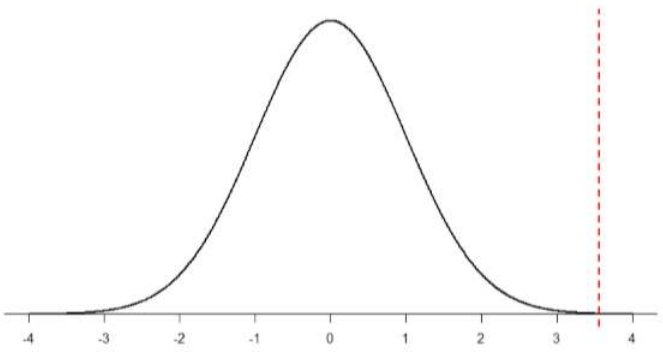


Figure **6.4.2**: Area under the curve greater than z = 3.55

Notice that no region of Figure **6.4.2** appears to be shaded. That is because the area under the curve that far out into the tail is so small that it can't even be seen (the red line has been added to show exactly where the region starts). Thus, we already know that the probability must be smaller for $N = 50$ than $N = 10$ because the size of the area (the proportion) is much smaller.

We run into a similar issue when we try to find $z = 3.55$ on our Standard Normal Distribution Table. The table only goes up to 3.09 because everything beyond that is almost 0 and changes so little that it's not worth printing values. The closest we can get is subtracting the largest value, 0.9990, from 1 to get 0.001. We know that, technically, the actual probability is smaller than this (since 3.55 is farther into the tail than 3.09), so we say that the probability is $p < 0.001$, or less than 0.1%.

This example shows what an impact sample size can have. From the same population, looking for exactly the same thing, changing only the sample size took us from roughly a 5% chance (or about 1/20 odds) to a less than 0.1% chance (or less than 1 in 1000). As the sample size n increased, the standard error decreased, which in turn caused the value of z to increase, which finally caused the p -value (a term for probability we will use a lot in Unit 2) to decrease. You can think of this relation like gears: turning the first gear (sample size) clockwise causes the next gear (standard error) to turn counterclockwise, which causes the third gear (z) to turn clockwise, which finally causes the last gear (probability) to turn counterclockwise. All of these pieces fit together, and the relations will always be the same:

$$\mathrm{n} \uparrow \sigma_{\overline{X}} \downarrow \mathrm{z} \uparrow \mathrm{p} \downarrow \tag{6.4.3}$$

Let's look at this one more way. For the same population of sample size 50 and standard deviation 10, what proportion of sample means fall between 47 and 53 if they are of sample size 10 and sample size 50?

We'll start again with $n = 10$. Converting 47 and 53 into z-scores, we get $z = -0.95$ and $z = 0.95$, respectively. From our z-table, we find that the proportion between these two scores is 0.6578 (the process here is left off for the student to practice converting $\overline{X}$ to z and z to proportions). So, 65.78% of sample means of sample size 10 will fall between 47 and 53. For $n = 50$, our z-scores for 47 and 53 are ±2.13, which gives us a proportion of the area as 0.9668, almost 97%! Shaded regions for each of these sampling distributions is displayed in Figure **6.4.3**. The sampling distributions are shown on the original scale, rather than as z-scores, so you can see the effect of the shading and how much of the body falls into the range, which is marked off with dotted line.

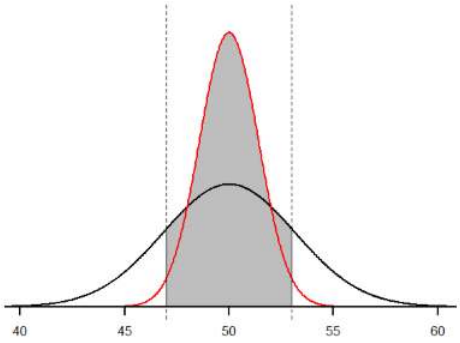


Figure **6.4.3**: Areas between 47 and 53 for sampling distributions of $n = 10$ and $n = 50$

This page titled 6.4: Using Standard Error for 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.

- **6.3: Using Standard Error for Probability** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

6.5: 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 6.5: 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.

6.6: Logic and Purpose of Hypothesis Testing

The statistician R. Fisher explained the concept of hypothesis testing with a story of a lady tasting tea. Here we will present an example based on James Bond who insisted that martinis should be shaken rather than stirred. Let's consider a hypothetical experiment to determine whether Mr. 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 6.6: 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.

6.7: 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 6.7: 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.

6.8: Descriptive versus Inferential Statistics

Now that we understand the nature of our data, let's turn to the types of statistics we can use to interpret them. There are 2 types of statistics: descriptive and inferential.

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 **6.8.1**. The table shows the average salaries for various occupations in the United States in 1999.

Table **6.8.1**: Average salaries for various occupations in 1999.

Occupation	Salary
Pediatricians	$112,760
Dentists	$106,130
Podiatrists	$100,090
Physicists	$76,140
Architects	$53,410
School, clinical, and counseling psychologists	$49,720
Flight attendants	$47,910
Elementary school teachers	$39,560
Police officers	$38,710
Floral designers	$18,980

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 **6.8.2**. It shows the number of unmarried men per 100 unmarried women in U.S. Metro Areas in 1990. From this table we see that men outnumber women most in Jacksonville, NC, and women outnumber men most in Sarasota, FL. You can see that descriptive statistics can be useful if we are looking for an opposite-sex partner! (These data come from the Information Please Almanac.)

Table **6.8.2**: Number of Unmarried Men per 100 Unmarried Women in U.S. Metro Areas in 1990.

Cities with Mostly Men	Men per 100 Women	Cities with Mostly Women	Men per 100 Women
1. Jacksonville, NC	224	1. Sarasota, FL	66
2. Killeen-Temple, TX	123	2. Bradenton, FL	68
3. Fayetteville, NC	118	3. Altoona, PA	69
4. Brazoria, TX	117	4. Springfield, IL	70
5. Lawton, OK	116	5. Jacksonville, TN	70
6. State College, PA	113	6. Gadsden, AL	70
7. ClarksvilleHopkinsville, TN-KY	113	7. Wheeling, WV	70
8. Anchorage, Alaska	112	8. Charleston, WV	71
9. Salinas-SeasideMonterey, CA	112	9. St. Joseph, MO	71
10. Bryan-College Station, TX	111	10. Lynchburg, VA	71

NOTE: Unmarried includes never-married, widowed, and divorced persons, 15 years or older.

These descriptive statistics may make us ponder why the numbers are so disparate in these cities. One potential explanation, for instance, as to why there are more women in Florida than men may involve the fact that elderly individuals tend to move down to the Sarasota region and that women tend to outlive men. Thus, more women might live in Sarasota than men. However, in the absence of proper data, this is only speculation.

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).

Table **6.8.3**: Winning Olympic Marathon Times for 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

Table **6.8.4** shows the same statistics, but for men.

Table **6.8.4**: Winning Olympic Marathon Times for 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

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 **6.8.3** and Table **6.8.4** 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 is $23.87. 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.

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.

Summary

In simpler terms, we *infer* characteristics of the population based on characteristics of the sample.

Definition: Descriptive Statistics

Used to describe or summarize the data from the sample.

Definition: Inferential Statistics

Used to make generalizations from the sample data to the population of interest.

Let's practice!

Example 6.8.1

Use the following to decide which option is **inferential**.

- Target Population -- All Psychology majors in the U.S.
- Sample -- 30 students from a Research Methods course
- Data -- 18 want to become Clinical Psychologists (60%)

Which of the following statements is descriptive of the sample and which is making an inference about the population? *Why?*

1. 60% of American Psychology majors want to be clinical psychologists.
2. 60% of the students in the sample want to be clinical psychologists.

Solution

#1 is inferential because it is using the information from the sample of one class to infer about all Psychology majors in the U.S..

Your turn!

? Exercise 6.8.1

Use the following to decide which option is **inferential**.

- Target Population -- California college students
- Sample -- 300 students from all 115 California community colleges
- Data – 150 are from the central valley, 150 are from outside of the valley

Which of the following statements is descriptive of the sample and which is making an inference about the population? *Why?*

1. 50% of California community college students are from the central valley.
2. 50% of the students in the sample are from the central valley.

Do you see anything wrong with this data?

Answer

#1 is inferential because it is using the information from the sample of 300 students to infer about all California college students.

If you know anything about the geography of California, you know that 50% of the state's population does not live in the central valley, so the sampling is suspect.

This page titled 6.8: Descriptive versus Inferential Statistics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **1.6: Types of Statistical Analyses** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.
- **7.2.2: Descriptive versus Inferential Statistics** by Michelle Oja is licensed CC BY-NC-SA 4.0.

6.9: 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:

$$H_0 : \mu_{obese} - \mu_{average} = 0 \tag{6.9.1}$$

The null hypothesis in a correlational study of the relationship between high school grades and college grades would typically be that the population correlation is 0. This can be written as

$$H_0 : \rho = 0 \tag{6.9.2}$$

where ρ is the population correlation, which we will cover in chapter 12.

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:

$$H_0 : \mu = 7.47$$

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 $\overline{X}$. This is for two reasons: 1) once we collect data, we know what the value of $\overline{X}$ 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 6.9: 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.

6.10: 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:

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

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:

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

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 6.10: The Alternative Hypothesis is shared under a CC BY-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.

6.11: Null Hypothesis Significance Testing

Null Hypotheses and Research Hypotheses

So far, so good? We develop a directional research hypothesis that names our groups, the DV (the outcome that was measured), and indicates a direction (which group will be higher). And we have a null hypothesis that says that the groups will have similar means on the DV. It's at this point that things get somewhat counterintuitive. Because the null hypothesis seems to correspond to the opposite of what I want to believe, and then we focus exclusively on that, almost to the neglect of the thing I'm actually interested in (the research hypothesis). In our growth mindset example, the null hypothesis is that the sample of junior high students with high beliefs in growth mindset will have similar average study times compared to the population of all junior high students. But for Blackwell, Trzseniewski, and Dweck (2007), and, really, any teacher ever, we actually want to believe that the understanding that your intelligence and abilities can always improve (high belief in growth mindset) will result in working harder and spending more time on homework. So the *alternative* to this null hypothesis is that those junior high students with higher growth mindset scores will spend more time on their math homework than those from the population of junior high students. **The important thing to recognize is that the goal of a hypothesis test is *not* to show that the research hypothesis is (probably) true; the goal is to show that the null hypothesis is (probably) false.** Most people find this pretty weird.

The best way to think about it, in my experience, is to imagine that a hypothesis test is a criminal trial… *the trial of the null hypothesis*. The null hypothesis is the defendant, the researcher is the prosecutor, and the statistical test itself is the judge. Just like a criminal trial, there is a presumption of innocence: the null hypothesis is *deemed* to be true unless you, the researcher, can prove beyond a reasonable doubt that it is false. You are free to design your experiment however you like, and your goal when doing so is to maximize the chance that the data will yield a conviction… for the crime of being false. The catch is that the statistical test sets the rules of the trial, and those rules are designed to protect the null hypothesis – specifically to ensure that if the null hypothesis is actually true, the chances of a false conviction are guaranteed to be low. This is pretty important: after all, the null hypothesis doesn't get a lawyer. And given that the researcher is trying desperately to prove it to be false, *someone* has to protect it.

Notation

Okay, so the null hypothesis always states that there's no difference. In our examples so far, we've been saying that there's no difference between the sample mean and population mean. But we don't *really* expect that, or why would we be comparing the means? The purpose of null hypothesis significance testing is to be able to **reject** the *expectation* that the means of the two groups are the *same*.

- Reject the null hypothesis: The sample mean is different from the population mean.
  - Rejecting the null hypothesis means that $\bar{X} \neq \mu$.
  - Rejecting the null hypothesis doesn't automatically mean that the research hypothesis is supported.
- Retain the null hypothesis: The sample mean is similar to the population mean.
  - Retaining the null hypothesis means that $\bar{X} = \mu$.
  - This means that our research hypothesis cannot be true.

We only reject or retain the null hypothesis. If we reject the null hypothesis (which says that everything is similar), we are saying that some means are statistically different from some other means. We only support the research hypothesis if the means are in the direction that we said. For example, if we rejected the null hypothesis that junior highers with high growth mindset spend as much time on homework as all junior highers, we can't automatically say that junior high students with high growth mindset study more than the population of junior high students. Instead, we'd have to look at the actual means of each group, and then decide if the research hypothesis was supported or not.

I hope that it's obvious that you don't have to look at the group means if the null hypothesis is retained?

In sum, you reject or retain the null hypothesis, and your support or or don't support the research hypothesis.

Why predict that two things are similar?

Because each sample's mean will vary around the population mean (see the first few sections of this chapter to remind yourself of this), we can't tell if our sample's mean is within a "normal" variance. But we *can* gather data to show that this sample's mean is different (enough) from the population's mean. This is *rejecting* the null hypothesis.

We use statistics to determine the probability of the null hypothesis being true.

? Exercise 6.11.1

Does a true null hypothesis say the sample mean and the population mean are similar or different?

Answer

A null hypothesis always says that the means are similar (or that there is no relationship between the variables).

Why can't we prove that the mean of our sample is different from the mean of the population? Remember the first few sections of this chapter, that showed how different samples from the same population have different means and standard deviations. Researchers are a conservative bunch; we don't want to stake our reputation on a sample mean that could be fluke, one of the extreme handfuls of green gumballs even when the mean difference between hands was zero.

But what we can show is that our sample is so extreme that it is statistically unlikely to be similar to the population.

Null hypothesis significant testing is like how courts decide if defendants are Guilty or Not Guilty, not their Guilt v. Innocent. Similarly, we decide if the sample is similar to the population or not.

Summary

This is a tough concept to grasp, so we'll keep working on it. And if you never get it, that's okay, too, as long as you remember the pattern of rejecting or retaining the null hypothesis, and supporting or not supporting the research hypothesis.

This page titled 6.11: Null Hypothesis Significance Testing is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **11.1: A Menagerie of Hypotheses** by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.
- **7.4: Null Hypothesis Significance Testing** by Michelle Oja is licensed CC BY-SA 4.0.

6.12: Critical Values, p-values, and Significance

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 $\alpha = 0.05$.

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 $\alpha = .05$, 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 **6.12.1**.

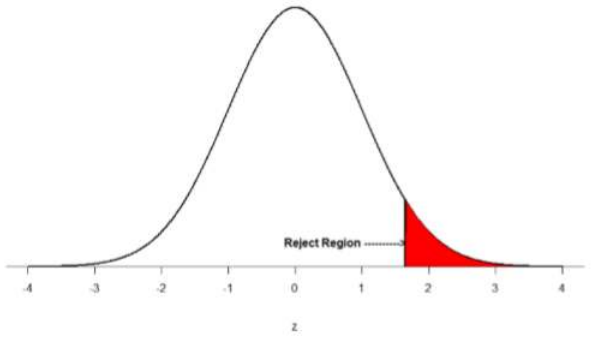


Figure **6.12.1**: The rejection region for a one-tailed test. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 0.0505 and z = -1.65 corresponds to 0.0497, so .05 is between them) if look at the proportion below the z-score. It would be 1.645 are looking for the top 5% of scores. The direction must be determined by your alternative hypothesis; drawing then shading the distribution is helpful for keeping directionality straight.

We will only talk about non-directional hypotheses when discussing Confidence Intervals, so here is a brief introduction. For non-directional research hypothesis, the critical region must be split between 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 $\alpha = .05$, this means 2.5% of the area is in each tail, which, based on the z-table, corresponds to critical values of $z* = \pm 1.96$. This is shown in Figure **6.12.2**.

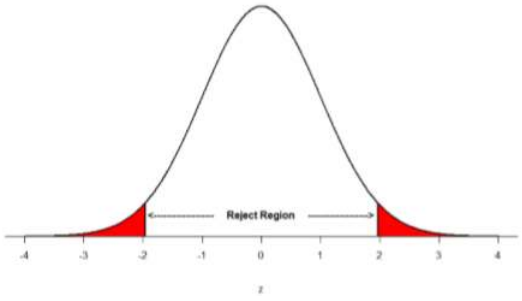


Figure **6.12.2**: Two-tailed rejection region. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Thus, any z-score falling outside ±1.96 (greater than 1.96 in absolute value) falls in the rejection region. Remember that as z gets larger (bigger standard deviations), the corresponding area under the curve *beyond* z gets smaller. Thus, if the area is smaller, the probability gets smaller that the number that we have from our sample is similar to the population. Specifically, the probability of obtaining that result, or a more extreme result, under the condition that the null hypothesis is true (that there really is no difference between the mean of the sample and the mean of the population) gets smaller (so there probably is a difference between the mean of the sample and the mean of the population).

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 statistic falls farther out into the tail than the critical value that corresponds to α, so the proportion (p-value) for thatstatistic will be smaller. Conversely, if we fail to reject, we know that the proportion will be larger than α because the statistic will not be as far into the tail. This is illustrated for a one-tailed test in Figure **6.12.3**.

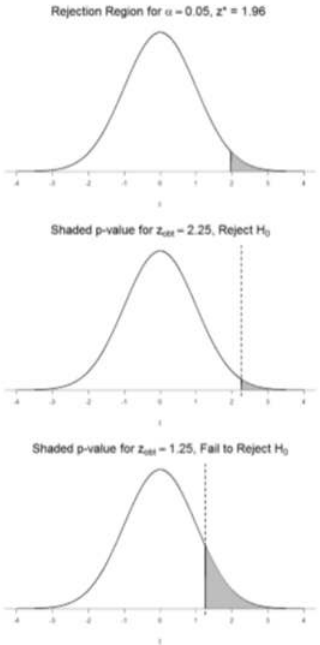


Figure **6.12.3**: Relation between α, z_{obt}, and p (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

When the null hypothesis is rejected, the effect is said to be statistically significant.

Statistical Significance

If we reject the null hypothesis, we can state that the means are different, or, in other words that the means are statistically significant. It is very important to keep in mind that statistical significance means only that the null hypothesis (of no mean differences) is rejected; it does not mean that the effect is important, which is what “significant” usually means. When an effect is

statistically significant, you can have confidence the effect is not exactly zero. Finding that an effect is statistically 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 meant that the statistics signified that the effect is real and not due to chance. Over the years, the meaning of “significant” changed, leading to the potential misinterpretation. In writing research reports, you should use the word "substantial" if you are not talking about a statistically significant calculation to avoid any confusion.

Still confused? The next section covers this information slightly differently, then there's a summary that discusses it all differently again. These are tough concepts, so try your best and keep at it!

This page titled 6.12: Critical Values, p-values, and Significance is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.12.1: Critical Values

Okay, this whole chapter is full of complex theoretical ideas. Critical values and null hypothesis significance testing is a TOUGH concept to get. Here's another description of critical values, p-values, and significance. Everyone learns differently, so hopefully this slightly different explanation will help understand the prior section.

Hypotheses

We understand that we need a research hypothesis that predicts the relationship between two groups on an measured outcome (DV), and that each research hypothesis has a null hypothesis that says that there is no relationship between the two groups on the measured outcome (the means of the DV are similar). If we reject the null hypothesis (which says that the means are similar), we are saying that the means are different; this may or may not be in the direction that we predicted in the research hypothesis so we may or may not support the research hypothesis. If we retain (fail to reject) the null hypothesis, we are saying that the means are similar and then we cannot support the research hypothesis. Got it?

But how do we decide again to retain or reject the null hypothesis? We compare a calculated statistic (there are many, depending on your variables, that we'll cover for the rest of this textbook) to a critical value from a table. The table uses probability (p-values) to tell us what calculated values are so extreme to be absolutely unlikely.

Critical Regions and Critical Values

The critical region of any test corresponds to those values of the test statistic that would lead us to reject null hypothesis (which is why the critical region is also sometimes called the rejection region). How do we find this critical region? Well, let’s consider what we know:

- The test statistic should be very big or very small (extreme) in order to reject the null hypothesis.
- If α=.05 (α is "alpha," and is just another notation for probability; we'll talk about it more in the section on errors), the critical region must cover 5% of a Standard Normal Distribution.

It’s only important to make sure you understand this last point when you are dealing with non-directional hypotheses (which we will only do for Confidence Intervals). The critical region corresponds to those values of the test statistic for which we would reject the null hypothesis, and the Standard Normal Distribution describes the probability that we would obtain a particular value if the null hypothesis (that the means are similar) were actually true. Now, let’s suppose that we chose a critical region that covers 20% of the Standard Normal Distribution, and suppose that the null hypothesis is actually true. What would be the probability of *incorrectly* rejecting the null (saying that there is a difference between the means when there really isn't a difference)? The answer is 20%. And therefore, we would have built a test that had an α level of 0.2. If we want α=.05, the critical region only covers 5% of the Standard Normal Distribution.

Huh? Let's draw that out. Figure **6.12.1.1** shows the critical region associated with a non-directional hypothesis test (also called a "two-sided test" because the calculated value might be in either tail of the distribution). Figure **6.12.1.1** itself shows the sampling distribution of X (the scores we got). The grey bars correspond to those values for which we would retain the null hypothesis. The blue (dark) bars show the critical region: those values for which we would reject the null. In this example, the research hypothesis was non-directional, so the critical region covers both tails of the distribution. To ensure an α level of .05, we need to ensure that each of the two regions encompasses 2.5% of the sampling distribution.

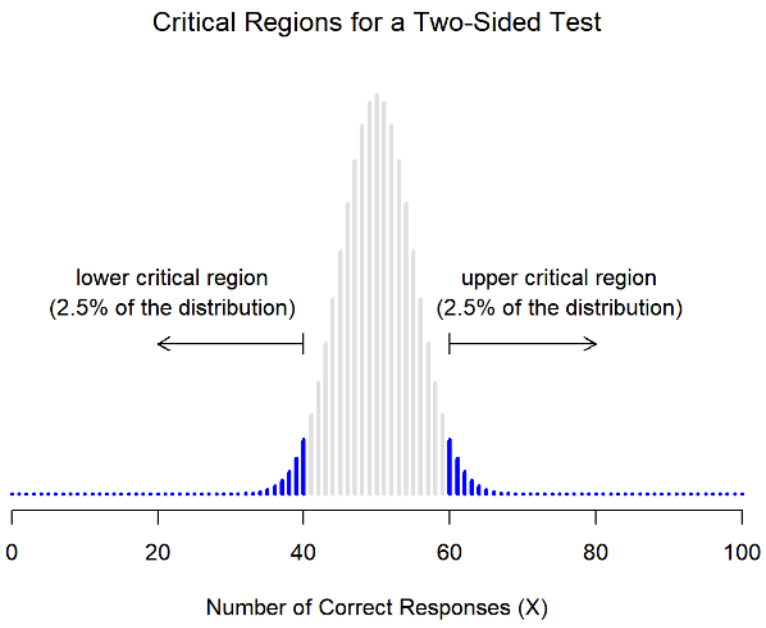


Figure **6.12.1.1**- Example of Critical Regions for a Non-Directional Research Hypothesis (CC-BY-SA- Danielle Navarro from Learning Statistics with R)

Our critical region consists of the most *extreme values*, known as the *tails* of the distribution.

At this point, our hypothesis test is essentially complete: (1) we choose an α level (e.g., α=.05, (2) come up with some test statistic (more on this step later) that does a good job (in some meaningful sense) of comparing the null hypothesis to the research hypothesis, (3) calculate the critical region that produces an appropriate α level, and then (4) calculate the value of the test statistic for the real data and then compare it to the critical values to make our decision. If we reject the null hypothesis, we say that the test has produced a ***significant*** result.

A note on statistical "significance"

> *Like other occult techniques of divination, the statistical method has a private jargon deliberately contrived to obscure its methods from non-practitioners.*
>
> *– Attributed to G. O. Ashley**

A very brief digression is in order at this point, regarding the word "significant". The concept of statistical significance is actually a very simple one, but has a very unfortunate name. If the data allow us to reject the null hypothesis, we say that "the result is *statistically significant*", which is often shortened to "the result is significant". This terminology is rather old, and dates back to a time when "significant" just meant something like "indicated", rather than its modern meaning, which is much closer to "important". As a result, a lot of modern readers get very confused when they start learning statistics, because they think that a "significant result" must be an important one. It doesn't mean that at all. All that "statistically significant" means is that the data allowed us to reject a null hypothesis. Whether or not the result is actually important in the real world is a very different question, and depends on all sorts of other things.

Directional and Non-Directional Hypotheses

There's one more thing to point out about the hypothesis test that we've just constructed. In statistical language, this is an example of a non-directional hypothesis, also known as a two-sided test. It's called this because the alternative hypothesis covers the area on both "sides" of the null hypothesis, and as a consequence the critical region of the test covers both tails of the sampling distribution (2.5% on either side if α=.05), as illustrated earlier in Figure **6.12.1.1**.

However, that's not the only possibility, and not the situation that we'll be using with our directional research hypotheses (the ones that predict which group will have a higher mean). A directional research hypothesis would only cover the possibility that p>.5, and as a consequence the null hypothesis now becomes p≤.5. When this happens, we have what's called a *one-sided test*, and when this happens the critical region only covers one tail of the sampling distribution. This is illustrated in Figure **6.12.1.2**. In this case, we

would only reject the null hypothesis for large values of of our test statistic (values that are more extreme but only in one direction). As a consequence, the critical region only covers the upper tail of the sampling distribution; specifically the upper 5% of the distribution. Contrast this to the two-sided version earlier in Figure **6.12.1.1**.

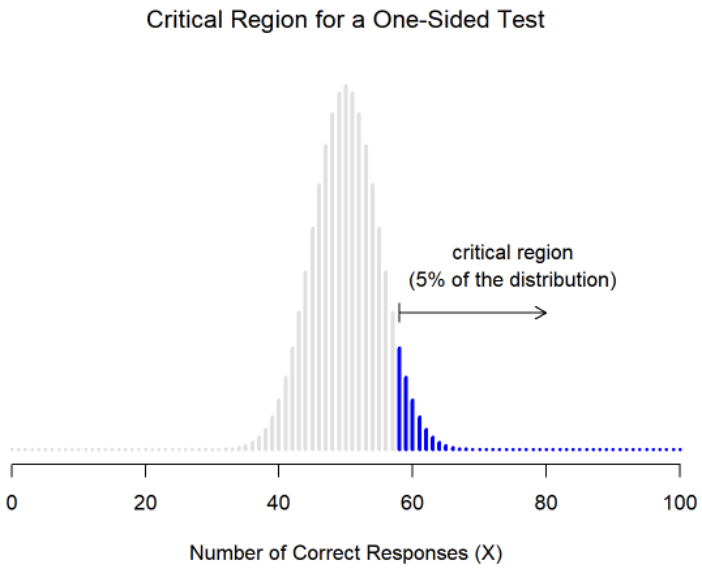


Figure **6.12.1.2**- Critical Region for Directional Research Hypotheses (CC-BY-SA- Danielle Navarro from Learning Statistics with R)

Clear as mud? Let's try one more way to describe how research hypotheses, null hypotheses, p-values, and null hypothesis significance testing work.

References

*The internet seems fairly convinced that Ashley said this, though I can't for the life of me find anyone willing to give a source for the claim.

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)

This page titled 6.12.1: Critical Values is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.12.2: Summary of p-values and NHST

We use p-values (probability) to determine if this could have happened by chance. We are comparing the mean of our sample to the mean of the population to ask “Is it likely that my sample is from the population?”

Since we can’t ever collect data from the whole population, we’re forced to make inferences from the available sample data, and the probability of events happening that we know from the Standard Normal Distribution. We know that any score that is more 3 or more standard deviations from the mean (z = 3) is very rare. So, we compare the mean of a sample to the population mean (after standardizing so that we can use the Standard Normal Distribution) to see if the mean of the sample is extreme enough to say that the probability of the sample being from the population is below 5%.

The p-value tells us *the **probability** of getting an effect this different if the sample is the same as the population*.

If the probability (p-value) is small enough (p< .05), then we conclude that the sample mean probably is from a different population.

? Exercise 6.12.2.1

What does the null hypothesis say?

Answer

The null hypothesis says that there is no difference between the groups, that the mean of the sample is similar to the mean of the pouplation.

✔ Example 6.12.2.1

If we’re saying that the sample is probably different from the population, are we **rejecting** or **retaining (failing to reject)** the null hypothesis?

Solution

If we’re saying that the sample is probably different from the population, we **reject**the null hypothesis.

Interpreting p-values

Without having to understand everything about probability distributions and the Standard Normal Distribution, what do the p-values tell us?

Small p-values

A small p-value means a small probability that the two means are similar.

Suggesting that the means are different…

We conclude that:

- The means are different.
- The sample is *not* from the population.

Large p-values

A large p-value means a large probability that the two means are similar.

We conclude that

- The means are similar.
- The sample *is* from the population.

Note

In other words:

$$\textbf{Reject null} = (\bar{X} \neq \mu) = p < .05$$

$$\textbf{Retain null} = (\bar{X} = \mu) = p > .05$$

If the probability is less than 5% that you would get a sample mean that is *this different* from the population mean if the sample really is from the population, the sample is probably *not* from that population. If you took 100 samples from a population, less than 5 of the samples' means would be *this different* from the population's mean if the samples were different in reality. So, the sample is probably not from the population. Probably.

You have a 5% chance that your results are wrong. You have a (small) chance that your sample mean is *this different* from the population mean, but the sample is still actually from the population.

Statisticians are okay with being wrong 5% of the time!

Let's practice!

Example 6.12.2.2

For each:

1. Determine whether to report "p<.05 " or "p>.05"
2. Determine whether to retain or reject the null hypothesis.
3. Determine whether the sample mean and population mean are similar or different.

Hint:

$$\textbf{Reject null} = (\bar{X} \neq \mu) = p < .05$$

$$\textbf{Retain null} = (\bar{X} = \mu) = p > .05$$

1. p < .05
2. p = .138
3. p = .510

Solution

1. p < .05
    1. "p<.05 "
    2. Reject the null hypothesis.
    3. The sample mean and population mean are different.
2. p = .138
    1. "p>.05" (because 0.138 > 0.05, 0.138 is bigger than 0.05)
    2. Retain the null hypothesis.
    3. The sample mean and population mean are similar.
3. p = .510
    1. "p>.05" (We're comparing to 0.05, or 5%, not .50, or 50%)
    2. Retain the null hypothesis.
    3. The sample mean and population mean are similar.

Now, try it yourself:

Exercise 6.12.2.2

For each:

1. Determine whether to report “p<.05 ” or “p>.05”
2. Determine whether to retain or reject the null hypothesis.
3. Determine whether the sample mean and population mean are similar or different.

Hint:

$$\textbf{Reject null} = (\bar{X} \neq \mu) = p < .05$$

$$\textbf{Retain null} = (\bar{X} = \mu) = p > .05$$

1. p > .05
2. p = .032
3. p = .049

Answer

1. p > .05
    1. “p>.05 ”
    2. Retain the null hypothesis.
    3. The sample mean and population mean are similar.
2. p = .032
    1. “p<.05 ” or “p>.05” (because 0.32 < .05, 0.032 is less than 0.05)
    2. Reject the null hypothesis.
    3. The sample mean and population mean are different.
3. p = .049
    1. “p<.05 ” (even though it’s close, .049 is smaller than 0.05)
    2. Reject the null hypothesis.
    3. The sample mean and population mean are different.

This page titled 6.12.2: Summary of p-values and NHST is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.13: 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 research 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. 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 calculated 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.

We will talk more about what is included in the write-up that explains the interpretation of the decision in relation to the research question, but remember that your answer is never just a number in behavioral statistics. And in Null Hypothesis Significance Testing, your answer is probably at least several sentences explaining the groups, what was measured, the results. and how it all relates to the research hypothesis.

This page titled 6.13: Steps of the Hypothesis Testing Process is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.14: 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.

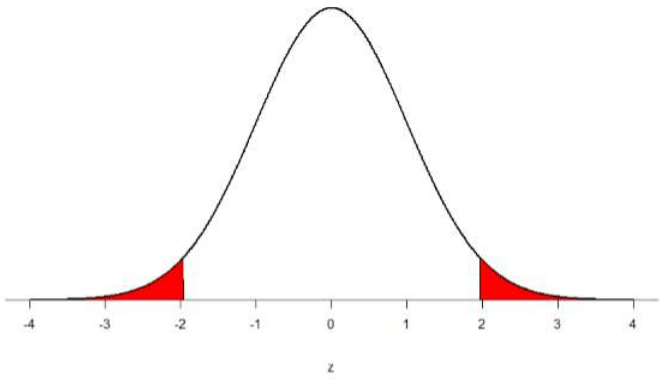


Figure **6.14.1**: Rejection region for $z*$ = ±1.96

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:

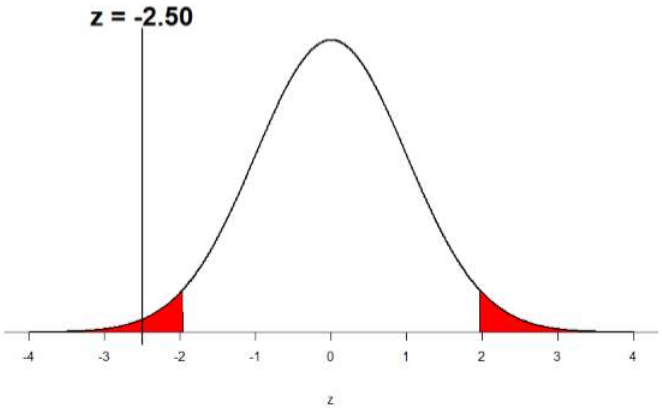


Figure 6.14.2: Test statistic location

Step 4: Make the Decision Looking at Figure 6.14.2, 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 6.14: 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.

6.15: 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***:

$$d = \frac{\overline{X} - \mu}{\sigma} \tag{6.15.1}$$

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 6.15: 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.

6.16: 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:

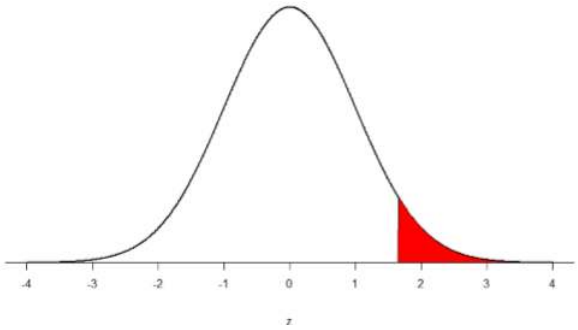


Figure 6.16.1: Rejection region

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

Table 6.16.1: 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!

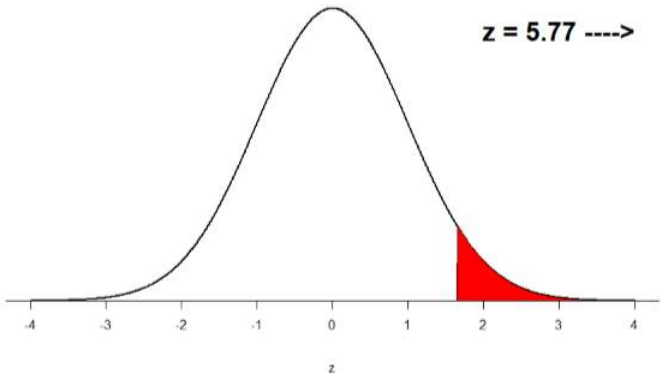


Figure 6.16.2: Obtained z-statistic

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 6.16: 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.

6.17: The Two Errors in Null Hypothesis Significance Testing

Before going into details about how a statistical test is constructed, it's useful to understand the philosophy behind it. We hinted at it when pointing out the similarity between a null hypothesis test and a criminal trial, but let's be explicit.

Ideally, we would like to construct our test so that we never make any errors. Unfortunately, since the world is messy, this is **never possible**. Sometimes you're just really unlucky: for instance, suppose you flip a coin 10 times in a row and it comes up heads all 10 times. That feels like very strong evidence that the coin is biased, but of course there's a 1 in 1024 chance that this would happen even if the coin was totally fair. In other words, in real life we *always* have to accept that there's a chance that we made the wrong statistical decision. The goal behind statistical hypothesis testing is not to *eliminate* errors, because that's impossible, but to *minimize* them.

At this point, we need to be a bit more precise about what we mean by "errors". Firstly, let's state the obvious: it is either the case that the null hypothesis is true, or it is false. The means are either similar or they ar not. The sample is either from the population, or it is not. Our test will either reject the null hypothesis or retain it. So, as the Table **6.17.1** illustrates, after we run the test and make our choice, one of four things might have happened:

Table **6.17.1**- Statistical Decision Versus Reality

Reality Versus Your Sample	Reality: Means are Different (Null Hypothesis is False)	Reality: Means are Similar (Null Hypothesis is True)
Your Sample: Means are Different (Reject Null Hypothesis)	Correct! :)	Error (Type I) :(
Your Sample: Means are Similar (Retain Null Hypothesis)	Error (Type II) :(	Correct! :)

As a consequence there are actually *two* different types of error here. If we reject a null hypothesis that is actually true, then we have made a ***type I error***. On the other hand, if we retain the null hypothesis when it is in fact false, then we have made a ***type II error***. Note that this does not mean that you, as the statistician, made a mistake. It means that, even when all evidence supports a conclusion, just by chance, you might have a wonky sample that shows you something that isn't true.

Errors in Null Hypothesis Significance Testing

Type I Error

- Reject a true null hypothesis.
  - The sample is from the population, but we say that it's not (rejecting the null).
- Saying there is a mean difference when there really isn't one!
- alpha (α, a weird a)
- False positive

Type II Error

- Retain a false null hypothesis.
  - The sample is from a different population, but we say that the means are similar (retaining the null).
- Saying there is not a mean difference when there really is one!
- beta (β, a weird B)
- Missed effect

Why the Two Types of Errors Matter

Null Hypothesis Significance Testing (NHST) is based on the idea that large mean differences would be rare if the sample was from the population. So, if the sample mean is different enough (greater than the critical value) then the effect would be rare enough (< .05) to reject the null hypothesis and conclude that the means are different (the sample is not from the population).

However, about ***5%*** of the times when we reject the null hypothesis, saying that the sample is from a different population, because ***we are wrong***. Null Hypothesis Significance Testing is not a "sure thing." Instead, we have a known error rate (5%). Because of this, replication is emphasized to further support research hypotheses. For research and statistics, "replication" means that we do many experiments to test the same idea. We do this in the hopes that we might get a wonky sample 5% of the time, but if we do enough experiments we will recognize the wonky 5%.

Remember how statistical testing was kind of like a criminal trial? Well,a criminal trial requires that you establish "beyond a reasonable doubt" that the defendant did it. All of the evidentiary rules are (in theory, at least) designed to ensure that there's (almost) no chance of wrongfully convicting an innocent defendant. The trial is designed to protect the rights of a defendant: as the English jurist William Blackstone famously said, it is "better that ten guilty persons escape than that one innocent suffer." In other words, a criminal trial doesn't treat the two types of error in the same way: punishing the innocent is deemed to be much worse than letting the guilty go free. A statistical test is pretty much the same: the single most important design principle of the test is to *control* the probability of a type I error, to keep it below some fixed probability (we use 5%). This probability, which is denoted α, is called the ***significance level*** of the test (or sometimes, the *size* of the test).

Introduction to Power

So, what about the type II error rate? Well, we'd also like to keep those under control too, and we denote this probability by β (beta). However, it's much more common to refer to the ***power*** of the test, which is the probability with which we reject a null hypothesis when it really is false, which is 1−β. To help keep this straight, here's the same table again, but with the relevant numbers added:

Table **6.17.2-** Statistical Decision Versus Reality with Alpha and Beta

Reality Versus Your Sample	Reality: Means are Different (Null Hypothesis is False)	Reality: Means are Similar (Null Hypothesis is True)
Your Sample: Means are Different (Reject Null Hypothesis)	Correct! :) $1-\beta$ (power of the test)	Error (Type I) :(α (type I error rate)
Your Sample: Means are Similar (Retain Null Hypothesis)	Error (Type II) :(β (type II error rate)	Correct! :) $1-\alpha$ (probability of correct retention)

"powerful" hypothesis test is one that has a small value of β, while still keeping α fixed at some (small) desired level. By convention, scientists usually use 5% (p = .05, α levels of .05) as the marker for Type I errors (although we also use of lower α levels of 01 and .001 when we find something that appears to be really rare). The tests are designed to ensure that the α level is kept small (accidently rejecting a null hypothesis when it is true), but there's no corresponding guarantee regarding β (accidently retaining the null hypothesis when the null hypothesis is actually false). We'd certainly like the type II error rate to be small, and we try to design tests that keep it small, but this is very much secondary to the overwhelming need to control the type I error rate. As Blackstone might have said if he were a statistician, it is "better to retain 10 false null hypotheses than to reject a single true one". To add complication, some researchers don't agree with this philosophy, believing that there are situations where it makes sense, and situations where I think it doesn't. But that's neither here nor there. It's how the tests are built.

Can we decrease the chance of Type I Error *and* decrease the chance of Type II Error? Can we make fewer false positives *and* miss fewer real differences?

Unfortunately, no. If we want fewer false positive, then we will miss more real effects. What we can do is increase the power of finding any real differences. We'll talk a little more about Power in terms of statistical analyses next.

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)

This page titled 6.17: The Two Errors in Null Hypothesis Significance Testing is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.17.1: Power and Sample Size

Large N and Small Effects

There is an intriguing relationship between N (sample-size) and power. As N increases, so does power to detect an effect. Additionally, as N increases, a design is capable of detecting smaller and smaller effects with greater and greater power. Let's think about what this means.

Imagine a drug company told you that they ran an experiment with 1 billion people to test whether their drug causes a significant change in headache pain. Let's say they found a significant effect (with power =100%), but the effect was very small, it turns out the drug reduces headache pain by less than 1%, let's say 0.01%. For our imaginary study we will also assume that this effect is very real, and not caused by chance.

Clearly the design had enough power to detect the effect, and the effect was there, so the design did detect the effect. However, the issue is that there is little practical value to this effect. Nobody is going to buy a drug to reduce their headache pain by 0.01%, even if it was "scientifcally proven" to work. This example brings up two issues. First, increasing N to very large levels will allow designs to detect almost any effect (even very tiny ones) with very high power. Second, sometimes effects are meaningless when they are very small, especially in applied research such as drug studies.

These two issues can lead to interesting suggestions. For example, someone might claim that large N studies aren't very useful, because they can always detect really tiny effects that are practically meaningless. On the other hand, large N studies will also detect larger effects too, and they will give a better estimate of the "true" effect in the population (because we know that larger samples do a better job of estimating population parameters).

Additionally, although really small effects are often not interesting in the context of applied research, they can be very important in theoretical research. For example, one theory might predict that manipulating X should have no effect, but another theory might predict that X does have an effect, even if it is a small one. So, detecting a small effect can have theoretical implication that can help rule out false theories. Generally speaking, researchers asking both theoretical and applied questions should think about and establish guidelines for "meaningful" effect-sizes so that they can run designs of appropriate size to detect effects of "meaningful size".

Small N and Large Effects

> **Note**
>
> All other things being equal, would you trust the results from a study with small N or large N?

This isn't a trick question, but sometimes people tie themselves into a knot trying to answer it. We already know that large sample-sizes provide better estimates of the distributions the samples come from. As a result, we can safely conclude that we should trust the data from large N studies more than small N studies.

At the same time, you might try to convince yourself otherwise. For example, you know that large N studies can detect very small effects that are meaningless in real life. You also know that small N studies are only capable of reliably detecting very large effects. So, you might reason that a small N study is better than a large N study because if a small N study detects an effect, that effect must be big and meaningful; whereas, a large N study could easily detect an effect that is tiny and meaningless.

This line of thinking needs some improvement. First, just because a large N study can detect small effects, doesn't mean that it *only* detects small effects. If the effect is large, a large N study will easily detect it. Large N studies have the power to detect a much wider range of effects, from small to large. Second, just because a small N study detected an effect, does not mean that the effect is real, or that the effect is large. For example, small N studies have more variability, so the estimate of the effect size will have more error. Also, there is 5% (or alpha rate) chance that the effect was spurious. Interestingly, there is a pernicious relationship between effect-size and type I error rate.

Type I errors: Convincing with Small Samples?

So what is this pernicious relationship between Type I errors and effect-size? Mainly, this relationship is pernicious for small N studies. Imagine a situation in which the null hypothesis is false, that there really is no mean differences between the groups. This is true, for example, on math performance by gender for children before puberty; girls and boys do equally well on math before the age of 13.

We know that under the null, researchers will find differences between groups that are similar about 5% of the time or less ($p<.05$), that is the definition. So, if a researcher measured math scores on 10-year olds in 100 experiments, they would expect to find a significant difference for 5 of their experiments by random chance. If the sample size is small, we are more likely to accidently find a gender difference because only a few girls that have more experience with math could shift the mean of the girls' math scores higher. This is the pernicious aspect. When you make a type I error for small N, your data will make you think there is no way it could be a type I error because the effect is just so big! When N is very large, like 1000 (say, 500 boys and 500 girls), it is very unlikely that you'd get a wonky sample when there is no differences between the groups. It would be very unlikely to get enough girls with more experience in math to actually shift a mean based on a 500 scores. So if you find a difference between the two groups when you have a large sample, it means that there probably is a real difference between the groups.

Probably.

Summary

Power: The chance of detecting a real difference, if there is one, between the sample mean and the population mean.

The easiest way to increase power is to increase the size of the sample (N).

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 6.17.1: Power and Sample Size is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **12.4: Some considerations** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

6.17.2: The p-value of a Test

There are two somewhat different ways of interpreting a p-value ("p" standing for "probability"), one proposed by Sir Ronald Fisher and the other by Jerzy Neyman. Both versions are legitimate, though they reflect very different ways of thinking about hypothesis tests. Most introductory textbooks tend to give Fisher's version only, but that's a bit of a shame. Neyman's version seems cleaner, and actually better reflects the logic of the null hypothesis test. You might disagree though, so both are included . We'll start with Neyman's version…

Neyman: A Softer View of Decision-Making

One problem with the hypothesis testing procedure that has been described so far is that it makes no distinction at all between a result this "barely significant" and those that are "highly significant". For instance, imagine that we tested whether people who say they have extra-sensory perception (ESP) really do by naming the pictures on cards, and found that people were able to correctly identify the picture 62 times out of 100 observations, the calculated value falls just inside the critical region and we get a statistically significant effect. However, it's so close to the "retaining" section of the critical value table that it is pretty nearly nothing. In contrast, suppose a different study found 97 out of the 100 participants got the answer right. This would obviously be significant too, but my a much larger margin. The procedure that we're using makes no distinction between the two. When we adopt the standard convention of allowing α=.05 as the acceptable Type I error rate, then both of these are significant results (62 correct answers and 97 correct answers are both statistically significant).

This is where the p value comes in handy. To understand how it works, let's suppose that we ran lots of hypothesis tests on the same data set: but with a different value of α in each case. When we do that for my original ESP data of 62 correct answers, what we'd get is something like this:

- If 0.05 is the chosen critical value ($\alpha = 0.05$), then the null hypothesis would be rejected.
- If 0.04 is the chosen critical value ($\alpha = 0.04$), then the null hypothesis would be rejected.
- If 0.03 is the chosen critical value ($\alpha = 0.03$), then the null hypothesis would be rejected.
- If 0.02 is the chosen critical value ($\alpha = 0.02$), then the null hypothesis would NOT be rejected.
- If 0.01 is the chosen critical value ($\alpha = 0.01$), then the null hypothesis would NOT be rejected.

hen we test ESP data (X=62 successes out of N=100 observations) using α levels of .03 and above, we'd always find ourselves rejecting the null hypothesis. For α levels of .02 and below, we always end up retaining the null hypothesis. Therefore, somewhere between .02 and .03 there must be a smallest value of α that would allow us to reject the null hypothesis for this data. This is the p value; as it turns out the ESP data has p=.021. In short:

In effect, p is a summary of all the possible hypothesis tests that you could have run, taken across all possible α values. And as a consequence it has the effect of "softening" our decision process. For those tests in which $p \leq \alpha$ you would have rejected the null hypothesis, whereas for those tests in which $p > \alpha$ you would have retained the null. The ESP study obtained X=62, with p=.021. So the error rate I have to tolerate is 2.1%. In contrast, suppose my experiment had yielded X=97. It becomes is a tiny, tiny Type I error rate. For this second case we would reject the null hypothesis with a lot more confidence that the sample really represents the reality of ESP in the population because we only have to be "willing" to tolerate a Type I error rate of about 1 in 10 trillion trillion in order to justify my decision to reject.

> p is defined to be the smallest Type I error rate (α) that you have to be willing to tolerate if you want to reject the null hypothesis.
>
> If it turns out that p describes an error rate that you find intolerable, then you must retain the null. If you're comfortable with an error rate equal to p, then it's okay to reject the null hypothesis in favor of your preferred alternative.

Fisher: The Probability of Extreme Data

The second definition of the p-value comes from Sir Ronald Fisher, and it's actually this one that you tend to see in most introductory statistics textbooks. Notice how the critical region corresponds to the *tails* (i.e., extreme values) of the sampling distribution? That's not a coincidence: almost all "good" tests have this characteristic (good in the sense of minimizing our Type II error rate, β). The reason for that is that a good critical region almost always corresponds to those values of the test statistic that are

least likely to be observed if the null hypothesis is true. If this rule is true, then we can define the p-value as the probability that we would have observed a test statistic that is at least as extreme as the one we actually did get. In other words, if the data are extremely implausible according to the null hypothesis, then the null hypothesis is probably wrong. You'll read different versions of this idea throughout this textbook.

A Common Mistake

Okay, so you can see that there are two rather different but legitimate ways to interpret the p value, one based on Neyman's approach to hypothesis testing and the other based on Fisher's. Unfortunately, there is a third explanation that people sometimes give, especially when they're first learning statistics, and it is *absolutely and completely wrong*. This mistaken approach is to refer to the p value as "the probability that the null hypothesis is true". It's an intuitively appealing way to think, but it's wrong. We won't get into why it's wrong because that gets into the philosophy of probability and statistics, but no hypotheses are every proved true or not true. Research hypotheses are support or not, and null hypotheses or retained or rejected. When a null hypothesis is retained, we're saying that our sample suggests that there's no differences between these groups in the population. It's easy fall into this phrasing, though. It might even be in this book somewhere!

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)

This page titled 6.17.2: The p-value of a Test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

CHAPTER OVERVIEW

7: Introducing t-tests (One-sample designs)

7.1: Introduction to One-Sample t-tests

7.2: Hypothesis Testing with t

7.3: Confidence Intervals

7.4: t table of critical values

7.5: Reporting Results

7.5.1: Descriptive and Inferential Calculations and Conclusion Example

---

7: Introducing t-tests (One-sample designs) is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

7.1: Introduction to One-Sample t-tests

You've learned about z-scores, but they aren't commonly used because they rely on knowing the populations standard deviation, σ, which is rarely the case. Instead, we will estimate that parameter σ using the sample statistic s in the same way that we estimate μ using $\overline{X}$ (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 is used for testing whether a sample mean is probably similar to the population mean. This statistical test is called a one-sample t-test because we are comparing one sample to a population.

Here we get back to math! The formula for a one-sample t-test is*:

$$t = \frac{(\bar{X} - \mu)}{\left(\frac{s}{\sqrt{n}}\right)}$$

What this is is the value of a sample mean compared (by subtracting) to what we expect of the population. But what about the denominator? The denominator of $\frac{s}{\sqrt{N}}$ is called the standard error. What's happening there? Notice that we are sorta finding an average of the standard deviation (which is already sorta an average of the distances between each score and the mean) by dividing the sample's standard deviation by the square root of N. The mathematical reasoning why we divide by the square root of N isn't really important. The focus here is that we are comparing the mean of the sample to the mean of the population (through subtraction), then turning that into a type of standardized (because of dividing by the standard deviation) average (because the standard deviation is divided by the something related to the size of the sample (the square root of N)).

Table of t-scores

One more thing before we practice calculating with this formula: Recall that we learned that the formulae for sample standard deviation and population standard deviation differ by one key factor: the denominator for the parameter of the population is N but the denominator for the sample 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. We can't use the table of z-scores anymore because critical values for t-scores are based on the size of the sample, so there's a different critical value for each N. For t-tests, we will use the t-distribution and t-table to find these values (see 8.3.1).

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 the sample size, 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 **7.1.1** 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. 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.

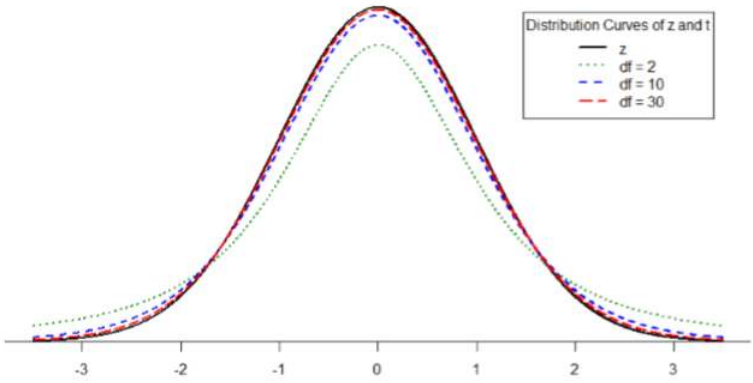


Figure **7.1.1**: Distributions comparing effects of degrees of freedom. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

---

* Sometimes you'll see the one-sample t-test formula look like the following:

$$t = \frac{(\bar{X} - \mu)}{s_{\bar{X}}}$$

The $s_{\bar{X}}$ means the same as $s/\sqrt{n}$, so let's just learn one formula and not get confused with something that looks like it might be a standard deviation but is not.

Contributors and Attributions

- Foster et al. (University of Missouri-St. Louis, Rice University, & University of Houston, Downtown Campus)
- Dr. MO (Taft College)

This page titled 7.1: Introduction to One-Sample t-tests is shared under a not declared license and was authored, remixed, and/or curated by Michelle Oja.

7.2: 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 7.2.1: our critical value is $t*$ = 2.353

t-distribution Table

df	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.01	1-tailed α 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 7.2.1: t-table

We can then shade this region on our t-distribution to visualize our rejection region

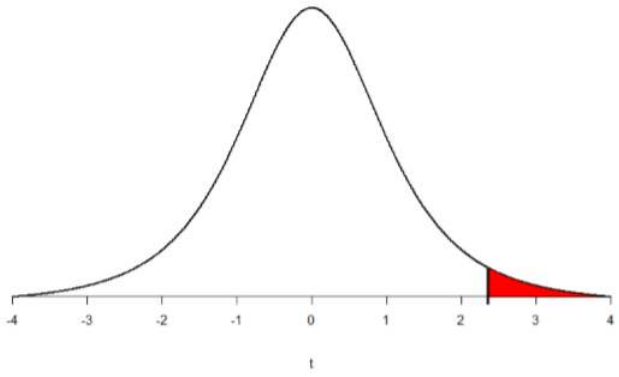


Figure 7.2.2: Rejection Region

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 7.2.1:

Table 7.2.1: 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 7.2: 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.

7.3: 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:

$$\text{Margin of Error } = t^*(s/\sqrt{n}) \tag{7.3.1}$$

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:

$$\begin{aligned} \text{Upper Bound} &= \bar{X} + \text{Margin of Error} \\ \text{Lower Bound } &= \bar{X} - \text{Margin of Error} \end{aligned} \tag{7.3.2}$$

Or simply:

$$\text{Confidence Interval } = \overline{X} \pm t^*(s/\sqrt{n}) \tag{7.3.3}$$

To write out a confidence interval, we always use soft brackets and put the lower bound, a comma, and the upper bound:

$$\text{Confidence Interval } = \text{ (Lower Bound, Upper Bound)} \tag{7.3.4}$$

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.

Hypothesis Testing with Confidence Intervals

As a function of how they are constructed, we can also use confidence intervals to test hypotheses. However, we are limited to testing two-tailed hypotheses only, because of how the intervals work, as discussed above.

Once a confidence interval has been constructed, using it to test a hypothesis is simple. The range of the confidence interval brackets (or contains, or is around) the null hypothesis value, we fail to reject the null hypothesis. If it does not bracket the null hypothesis value (i.e. if the entire range is above the null hypothesis value or below it), we reject the null hypothesis. The reason for this is clear if we think about what a confidence interval represents. Remember: a confidence interval is a range of values that we consider reasonable or plausible based on our data. Thus, if the null hypothesis value is in that range, then it is a value that is plausible based on our observations. If the null hypothesis is plausible, then we have no reason to reject it. Thus, if our confidence interval brackets the null hypothesis value, thereby making it a reasonable or plausible value based on our observed data, then we have no evidence against the null hypothesis and fail to reject it. However, if we build a confidence interval of reasonable values based on our observations and it does not contain the null hypothesis value, then we have no empirical (observed) reason to believe the null hypothesis value and therefore reject the null hypothesis.

Let's see an example. You hear that the national average on a measure of friendliness is 38 points. You want to know if people in your community are more or less friendly than people nationwide, so you collect data from 30 random people in town to look for a difference. We'll follow the same four step hypothesis testing procedure as before.

Step 1: State the Hypotheses We will start by laying out our null and alternative hypotheses:

H_0: There is no difference in how friendly the local community is compared to the national average

$$H_0 : \mu = 38$$

H_A: There is a difference in how friendly the local community is compared to the national average

$$H_A : \mu \neq 38$$

Step 2: Find the Critical Values We need our critical values in order to determine the width of our margin of error. We will assume a significance level of α = 0.05 (which will give us a 95% CI). From the t-table, a two-tailed critical value at α = 0.05 with 29 degrees of freedom (N – 1 = 30 – 1 = 29) is $t*$ = 2.045.

Step 3: Calculations Now we can construct our confidence interval. After we collect our data, we find that the average person in our community scored 39.85, or $\overline{X}$= 39.85, and our standard deviation was s = 5.61. First, we need to use this standard deviation, plus our sample size of N = 30, to calculate our standard error:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{5.61}{5.48} = 1.02$$

Now we can put that value, our point estimate for the sample mean, and our critical value from step 2 into the formula for a confidence interval:

$$95\%CI = 39.85 \pm 2.045(1.02)$$

$$\text{Upper Bound} = 39.85 + 2.045(1.02)$$
$$UB = 39.85 + 2.09$$
$$UB = 41.94$$

$$\text{Lower Bound} = 39.85 - 2.045(1.02)$$
$$LB = 39.85 - 2.09$$
$$LB = 37.76$$

$$95\%CI = (37.76, 41.94)$$

Step 4: Make the Decision Finally, we can compare our confidence interval to our null hypothesis value. The null value of 38 is higher than our lower bound of 37.76 and lower than our upper bound of 41.94. Thus, the confidence interval brackets our null hypothesis value, and we fail to reject the null hypothesis:

> Fail to Reject H_0. Based on our sample of 30 people, our community not different in average friendliness ($\overline{X}$= 39.85) than the nation as a whole, 95% CI = (37.76, 41.94).

Note that we don't report a test statistic or p-value because that is not how we tested the hypothesis, but we do report the value we found for our confidence interval.

An important characteristic of hypothesis testing is that both methods will always give you the same result. That is because both are based on the standard error and critical values in their calculations. To check this, we can calculate a t-statistic for the example above and find it to be t = 1.81, which is smaller than our critical value of 2.045 and fails to reject the null hypothesis.

Confidence Intervals using z Confidence intervals can also be constructed using z-score criteria, if one knows the population standard deviation. The format, calculations, and interpretation are all exactly the same, only replacing $t*$ with $z*$ and $s_{\overline{X}}$ with $\sigma_{\overline{X}}$.

This page titled 7.3: 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.

7.4: t table of critical values

7.4: t table of critical values is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

7.5: Reporting Results

Through the practice examples, I hope that you have realized that when conducting statistics for the social sciences, the answer is never just the number. We do the statistics to answer questions, so the final answer needs enough information to answer that question, and to let other statisticians know a little bit about the sample and the calculations.

Reporting the Results

There's no point in designing and running an experiment and then analyzing the data if you don't tell anyone about it! So let's now talk about what you need to do when reporting your analysis. Let's practice with an example with playing cards.

Scenario

Imagine this scenario: After playing a game of solitaire on my phone 10 times, I found that I won 6 times. I felt like I wasn't doing well, so I found that, on average, folks win solitary 43% of the time. To conduct a t-test, I asked my ten closest friends to play solitaire 10 times and let me know how many times they won. I guess I don't have 10 close friends, because only 4 people replied back. If I found the sample mean ($\bar{X} = .55$ (my friends and I won 55% of the time), my research hypothesis is that the sample had a higher win rate than the population ($\bar{X} > \mu$). With a standard deviation (say, s=0.18), I could conduct a t-test! We'll skip how to do that part for now, and move on to how to write up your results.

Write-Up

Here's a sample way to report this would be to write something like this:

> *The average win rate for the five participants in the experiment was 55% ($\bar{X} = .55$, while the population average win rate was 43% ($\mu = .43$. A one-sample t-test was conducted to test whether the sample represents the population. The results were not significant (t(3) = 1.49, p > 0.05), suggesting that the sample's win rate is similar to the population's win rate. This does not support the research hypothesis that the sample would have a higher win rate than the population.*

What to Include:

This is pretty straightforward, and hopefully it seems pretty unremarkable. That said, there's a few things that you should note about this description:

- *The statistical test is preceded by the descriptive statistics (means)*. That is, I told the reader something about what the data look like before going on to do the test. In general, this is good practice: always remember that your reader doesn't know your data anywhere near as well as you do. So unless you describe it to them properly, the statistical tests won't make any sense to them, and they'll get frustrated and cry.
- *The description tells you what the research hypothesis being tested is*. To be honest, writers don't always do this, but it's often a good idea since it might be a long ways and a lot of time from when the research hypothesis was first presented. Also, notice that the research hypothesis is in words, not in maths. That's perfectly acceptable. You can describe it in symbols and mathematical notation if you like, but since most readers find words easier to read than symbols, most writers tend to describe the hypotheses using words if they can. For help knowing how to write numbers in your paragraph, check out this page on Reporting Statistics in APA Style.
- *A "statistical sentence" showing the results is included*. When reporting the results of the test itself, I didn't just say that the result was no statistically significant, I included a "statistical sentence" (i.e., the dense mathematical-looking part in the parentheses), which reports all the statistical results. For the t-test, the information that gets reported is the test statistic result (that the calculated t-score was 1.49), the information about the distribution used in the test (the "t"), the Degrees of Freedom (which helps understand the sample size), and then the information about whether the result was significant or not (in this case p>.05). The general principle is that you should always provide enough information so that the reader could check the test results themselves if they really wanted to. Writing t(4)=1.49 is essentially a highly condensed way of writing "the sampling distribution of the t-test statistic with degrees of freedom of 4, and the value of the calculated t-score is 1.49". This page on

Reporting Statistics in APA Style (website address: https://my.ilstu.edu/~jhkahn/apastats.html) also shows how to write these "statistical sentences."

- *The results are interpreted*. In addition to indicating that the result was significant, I provided an interpretation of the result (i.e., that the mean of the sample was similar to the mean of the population), and whether or not the research hypothesis was supported. If you don't include something like this, it's really hard for your reader to understand what's going on.

What NOT to Include:

One thing to notice is that the null hypothesis and the critical value is NOT included. That information is for you to use to make the decision, but readers *should* be able to figure out what happened by seeing the p-value, and whether it's $p>0.05$ or $p<0.05$.

In Closing,

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

Note

- *The statistical test is preceded by the descriptive statistics (means).*
- *The description tells you what the research hypothesis being tested is.*
- *A "statistical sentence" showing the results is included.*
- *The results are interpreted in relation to the research hypothesis.*

Okay, once more, with feeling! Let's do a full practice problem, with the calculations and the write-up and everything!

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)

This page titled 7.5: Reporting Results is shared under a not declared license and was authored, remixed, and/or curated by Michelle Oja.

7.5.1: Descriptive and Inferential Calculations and Conclusion Example

Let's use data that Dr. MO's student research group collected on growth mindset from students at a community college. We'll start by interpreting the descriptive statistics and graph of the sample from the output of statistical software, then go step-by-step through a one-sample t-test.

Here we go!

Scenario

Students in Dr. MO's research group conducted an experiment in which several faculty members used videos, readings, and in-class discussion to try to improve the growth mindset (see 7.1) of their students. We will start by looking at the pretest scores of all 126 community college students in this study. In other words, we are looking at the students' scores on an assessment of their mindset before they had learned anything about growth mindset. The survey score could be from 0 points (totally fixed mindset) to 60 points (totally growth mindset), with any score above 33 points having more growth mindset beliefs.

Let's describe what we know already!

? Exercise 7.5.1.1

1. Based on the scenario, who is the sample? *Your answer should include the number of people in the sample.*
2. Based on this sample, who could be the population? Why?
3. What is the DV (effect)? In other words, what was the outcome that was measured?

Answer

1. 126 community college students
2. Any reasonable population based on the sample, but probably community college students in the U.S.
3. Scores on the mindset assessment.

Now, let's learn more about this sample, and interpret what these descriptive statistics can show us!

Descriptive Statistics

Measures of Central Tendency

The measures of central tendency are shown in Table **7.5.1.1**.

Table **7.5.1.1**- Measures of Central Tendency for Mindset Pretest

Measure of Central Tendency	Result
N	126
Mean	42.06
Median	41.00
Mode	39

What can we learn from these measures of central tendency?

? Exercise 7.5.1.2

Knowing the mean, median, and mode, what could be the “center” of this distribution of mindset scores? In other words, what one number might represent a typical score? *Why* do you think that?

Answer

The mean ($\bar{X} = 42.06 points$) and median (Md = 41 points)) are relatively close, suggesting that the "center" of this distribution is 41 or 42 points. However, because the mode is lower (Mode = 39 points), the center is probably closer to that lower number of 41 points.

As you know, measures of central tendency don't tell the whole story about a distribution. Let's look at varibility next.

Measure of Variability

The main measure of variability that we look at is standard deviation. The standard deviation of this sample of mindset scores is 6.65 points. We can also look at the highest and lowest scores to see the range. The lowest mindset score in this sample was 28 points, and the highest score was the maximum of 60 points. What can this tell us about the distribution?

? Exercise 7.5.1.3

Based on the mean and standard deviation only, do you think that the distribution of mindset scores will be tall and narrow, short and wide, or a medium bell-curve? Why?

Answer

I think that a standard deviation of 6.65 is medium, so I think that the distribution will be a bell-shaped curve. However, since no once scored below 28 points, maybe it's negatively skewed (a tail pointing to the left with the bulk of scores to the right.

I literally was just guessing in Exercise **7.5.1.3**. Let's look at the actual frequencies to see if I guessed correctly!

Frequency

Table **7.5.1.2** is a frequency table for this sample's mindset scores, and Figure **7.5.1.1** shows the same information in a line graph.

Table **7.5.1.2**- Frequency of Mindset Scores

X (Mindset Score)	Frequency
28	1
30	1
31	2
32	3
33	6
34	2
35	3
36	7
37	9
38	6
39	12
40	7
41	5
42	7
43	8
44	3
45	8
46	3
47	7
48	5
49	5
51	4
52	2
53	2
54	3
55	1
56	1
57	1
60	2

And Figure 7.5.1.1:

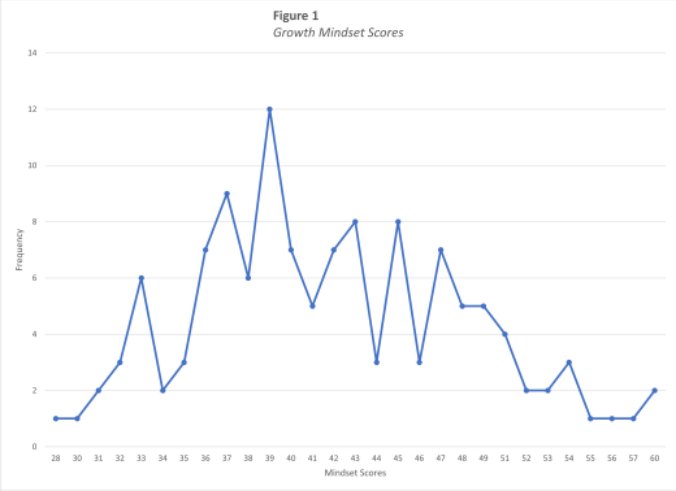


Figure **7.5.1.1**: Frequency Line Graph of Mindset Scores. (CC-BY-NC-SA; Michelle Oja)

Let's start unpacking Figure **7.5.1.1**.

Exercise 7.5.1.4

1. What kind of graph is Figure **7.5.1.1**?
2. What does the x-axis measure in Figure **7.5.1.1**?
3. What does the y-axis measure in Figure **7.5.1.1**?
4. Is Figure **7.5.1.1** skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
5. What do you notice from Figure **7.5.1.1**? What pops out to you?
6. What does Figure **7.5.1.1** make you wonder about?
7. What is a catchy headline for Figure **7.5.1.1**?
8. How could you summarize the info in Figure **7.5.1.1** into one sentence?
9. Who might want to know the information in Figure **7.5.1.1**?

Answer

1. What kind of graph is Figure **7.5.1.1**?
    1. Figure **7.5.1.1** is a line graph showing frequencies.
2. What does the x-axis measure in Figure **7.5.1.1**?
    1. The x-axis measures mindset scores.
3. What does the y-axis measure in Figure **7.5.1.1**?
    1. The y-axis measures frequencies of each mindset score.
4. Is Figure **7.5.1.1** skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?
    1. Although the mode isn't quite in the center, I would say that this frequency distribution is not skewed. It seems to be a medium normal curve (not tall/narrow or wide/flat).
5. What do you notice from Figure **7.5.1.1**? What pops out to you?
    1. I was actually surprised by how symmetrical the distribution was! I also noticed that no one scored below 28 points, even though someone could score all the way down to zero points on this mindset assessment.
6. What does Figure **7.5.1.1** make you wonder about?
    1. This makes me wonder why no one scored low; the cut-off for having a fixed mindset is 33 points, so it's surprising that few students in this sample have a fixed mindset.
7. What is a catchy headline for Figure **7.5.1.1**?
    1. Community College Students Start with Growth!
8. How could you summarize the info in Figure **7.5.1.1** into one sentence?

1. From this sample, I might infer that 90% of community college students hold growth mindset views.

9. Who might want to know the information in Figure **7.5.1.1**?
    1. Definitely community college faculty and adminstrators, as well as researchers who want to learn about growth mindset or improve student performance.

Okay, now that we've looked at Figure **7.5.1.1** closely, let's think about it in relation to the measures of central tendencies.

? Exercise 7.5.1.5

First, look back at your answer about whether the distribution would be tall/narrow, flat/wide, or medium/bell-shaped. Were you correct? Reflect on what about this distribution of data made it easier or more difficult to predict the shape based on these descriptive statistics.

Second, knowing the mean, median, mode, standard deviation, and what the distribution looks like, do you think that community college students tend to have fixed mindsets or growth mindsets?

Answer

First, yes, I believe that I estimated correctly. I said that the distribution should be medium and bell-shaped, and I think that it is. I think it was easier to correctly estimate the shape because the mean and median were so close. But that lower mode was a little confusing!

Second, knowing the mean, median, mode, standard deviation, and what the distribution looks like, think that community college students tend to have growth mindsets. The measures of central tendency are all above the cut-off of 33 points for a fixed mindset, and the graph shows that only a few students had fixed mindsets. Also, the standard deviation wasn't too high, which suggests that there are no outliers (extreme scores), which the graph confirms.

You might wonder why we went through all of this. The answers to these questions will be what you use to for a future paper analyzing a data set. Plus, now you know more about growth mindset of community college students!

Note

Now that you have the full data set, you are encouraged to calculate the mean, median, mode, and standard deviation yourself, and create your own graph as a refresher! (If you do, note that the data is in a frequency table, it is not a list of all of the scores individually...)

But this chapter isn't about growth mindset or descriptive statistics. This chapter is about one-sample t-tests, so let's get to that!

Null Hypothesis Significance Testing Steps

Step 1: State the Hypotheses

Let's start with a research hypothesis.

When the student researchers first started on this research project, we were very surprised by how high the mindset scores start out. Other research suggests that the cut-off of 33 points is a good estimate of the general population's mindset scores. This leads to:

- Research Hypothesis: Community college students' average mindset scores are higher than the general population's mindset scores.
- Symbols: $\bar{X} > \mu$

? Exercise 7.5.1.6

Based on the research hypothesis, what is the null hypothesis? Write it out in words, then in symbols.

Answer

- Null Hypothesis: Community college students' average mindset scores as similar to the general population's mindset scores.
- Symbols: $\bar{X} = \mu$

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. Using the table of critical values for the t-distribution, what critical value do we use?

Example 7.5.1.1

First, what Degrees of Freedom would you use?

Second, what is the critical t-score?

Solution

First, $df = N - 1$, so:

$$126 - 1 = 125$$

Second, for df=126, we use df=100 from the table (because it is both the closest whether you round up or down). The critical t-score for df=100 is 1.660.

To keep track of the directionality of the test and rejection region, you draw out your distribution:

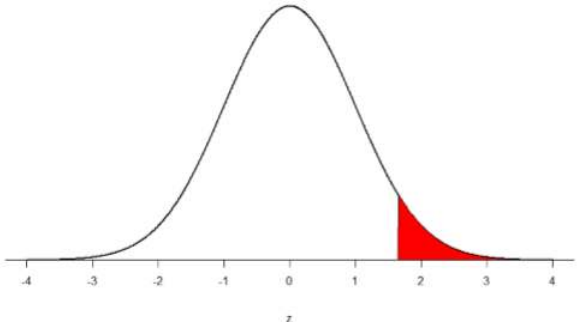


Figure 7.5.1.2: Rejection region (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Step 3: Calculate the Test Statistic

Okay, you have everything you need to calculate the one-sample t-test! To make sure it's all in one place, Table 7.5.1.3 puts all of the numbers in one place. This might be a good practice everytime that you conduct a statistical analysis...

Table 7.5.1.3- Necessary Information for One-Sample t-test

Necessary Information	Numbers
N	126
Mean of the Sample	42.06
Std. Deviation	6.65
Mean of the Population	33.00

$$t = \frac{(\bar{X} - \mu)}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{(42.06 - 33)}{\left(\frac{6.65}{\sqrt{126}}\right)} = \frac{(9.06)}{\left(\frac{6.65}{11.22}\right)} = \frac{9.06}{0.59} = 15.29$$

Step 4: Make the Decision

You compare your calculated t-statistic (t = 15.29), to the critical value, t_{crit} = 1.6600, and find that the calculated score is bigger than the critical value.

Remember:

> $(Critical < |Calculated|) = \textbf{Reject null} = (\bar{X} \neq \mu) = p < .05$
> $(Critical > |Calculated|) = \textbf{Retain null} = (\bar{X} = \mu) = p > .05$

So, we reject the null hypothesis, and write up our conclusion.

Write-Up

Remember the four components that should be in a complete conclusion?

? Exercise 7.5.1.7

What should your write-up of the t-test results look like?

Answer

I hypothesized that the sample's mindset scores ($\bar{X} = 42.06 points$ would be higher than mindset scores of the population ($\mu = 33 points$). This research hypothesis was supported (t(125) = 15.29, p< 0.05); the sample of community college students did have higher mindset scores than the general population.

Cool, huh?!

This page titled 7.5.1: Descriptive and Inferential Calculations and Conclusion Example is shared under a not declared license and was authored, remixed, and/or curated by Michelle Oja.

CHAPTER OVERVIEW

8: Two Sample Designs

8.1: Repeated Measures

8.1.1: Change and Differences

8.1.2: Hypotheses of Change and Differences

8.1.3: Increasing Satisfaction at Work

8.1.4: Bad Press

8.2: Independent Samples

8.2.1: Difference of Means

8.2.2: Research Questions about Independent Means

8.2.3: Hypotheses and Decision Criteria

8.2.4: Independent Samples t-statistic

8.2.5: Standard Error and Pooled Variance

8.2.6: Movies and Mood

8.2.7: Effect Sizes and Confidence Intervals

8.2.8: Homogeneity of Variance

8: Two Sample Designs is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

SECTION OVERVIEW

8.1: Repeated Measures

8.1.1: Change and Differences

8.1.2: Hypotheses of Change and Differences

8.1.3: Increasing Satisfaction at Work

8.1.4: Bad Press

This page titled 8.1: 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.

8.1.1: 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 **8.1.1.1**: 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 **8.1.1.1** 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 **8.1.1.1**, we calculate our difference scores by taking the time 2 score and subtracting the time 1 score. That is:

$$\mathbf{X_d = X_{T2} - X_{T1}} \tag{8.1.1.1}$$

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 8.1.1: 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.

8.1.2: 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.

$$\overline{X_D} = \frac{\Sigma X_D}{n} \tag{8.1.2.1}$$

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:

$$s_D = \sqrt{\frac{\sum \left(X_D - \overline{X_D}\right)^2}{n-1}} = \sqrt{\frac{SS}{df}} \tag{8.1.2.2}$$

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:

$$s_{\overline{X}_D} = {}^{s_D}/\sqrt{n} \tag{8.1.2.3}$$

Finally, our test statistic t has the same structure as well:

$$t = \frac{\overline{X_D} - \mu_D}{s_{\overline{X}_D}} \tag{8.1.2.4}$$

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 8.1.2: 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.

8.1.3: 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 8.1.3.1.

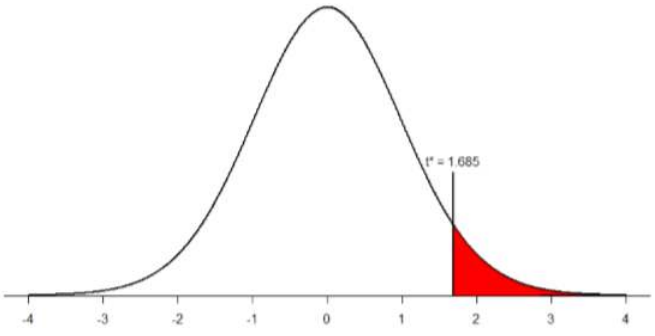


Figure 8.1.3.1: Critical region for one-tailed t-test at α = 0.05

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 8.1.3: 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.

8.1.4: 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 8.1.4.1. 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 8.1.4.1. As a reminder, the difference scores are calculated as Time 2 – Time 1.

Table 8.1.4.1: 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 8.1.4.2: 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 8.1.4: 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.

SECTION OVERVIEW

8.2: Independent Samples

8.2.1: Difference of Means

8.2.2: Research Questions about Independent Means

8.2.3: Hypotheses and Decision Criteria

8.2.4: Independent Samples t-statistic

8.2.5: Standard Error and Pooled Variance

8.2.6: Movies and Mood

8.2.7: Effect Sizes and Confidence Intervals

8.2.8: Homogeneity of Variance

This page titled 8.2: 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.

8.2.1: 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 8.2.1: 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.

8.2.2: 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 8.2.2: 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.

8.2.3: 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 8.2.3: 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.

8.2.4: 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:

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{s_{\overline{X}_1 - \overline{X}_2}} \tag{8.2.4.1}$$

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:

$$t = \frac{(\overline{X_1} - \overline{X_2})}{s_{\overline{X}_1 - \overline{X}_2}} \tag{8.2.4.2}$$

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 8.2.4: 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.

8.2.5: 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:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} \tag{8.2.5.1}$$

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:

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} \tag{8.2.5.2}$$

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:

$$S_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \tag{8.2.5.3}$$

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:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}} \tag{8.2.5.4}$$

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 8.2.5: 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.

8.2.6: 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

$$\mathrm{H_A}: \mu_1 - \mu_2 > 0$$

or

$$\mathrm{H_A}: \mu_1 > \mu_2$$

Notice that in the first formulation of the alternative hypothesis we say that the first mean minus the second mean will be greater than zero. This is based on how we code the data (higher is better), so we suspect that the mean of the first group will be higher. Thus, we will have a larger number minus a smaller number, which will be greater than zero. Be sure to pay attention to which group is which and how your data are coded (higher is almost always used as better outcomes) to make sure your hypothesis makes sense!

Step 2: Find the Critical Values Just like before, we will need critical values, which come from 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 **8.2.6.1** shows the values for the Comedy group, and Table **8.2.6.2** 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 **8.2.6.1**: Raw scores and Sum of Squares for Group 1

x	(x − x̄)	(x − x̄)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 8.2.6.2: 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 $\mu_1 - \mu_2$ 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 8.2.6: 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.

8.2.7: 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:

$$d = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{s_p^2}} \tag{8.2.7.1}$$

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:

$$\text{Confidence Interval } = (\overline{X_1} - \overline{X_2}) \pm t^* \left(s_{\overline{X_1} - \overline{X_2}}\right) \tag{8.2.7.2}$$

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:

$$s^2 = (s)^2 = \frac{SS}{df} \tag{8.2.7.3}$$

To get the Sum of Squares, we just multiply both sides of the above equation to get:

$$s^2 * df = SS \tag{8.2.7.4}$$

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 8.2.7: 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.

8.2.8: 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 8.2.8: 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.

CHAPTER OVERVIEW

9: Analysis of Variance- Independent Samples

9.1: Why ANOVA?

9.1.1: Observing and Interpreting Variability

9.1.2: Ratio of Variability

9.2: Introduction to ANOVA's Sum of Squares

9.2.1: Summary of ANOVA Summary Table

9.3: Hypotheses in ANOVA

9.4: Practice with Job Applicants

9.4.1: Table of Critical F-Scores

9.5: Introduction to Pairwise Comparisons

9.5.1: Pairwise Comparison Post Hoc Tests for Critical Values of Mean Differences

9.6: Practice on Mindset Data

9.7: On the Relationship Between ANOVA and the Student t Test

9.8: Non-Parametric Analysis Between Multiple Groups

This page titled 9: Analysis of Variance- Independent Samples is shared under a not declared license and was authored, remixed, and/or curated by Michelle Oja.

9.1: Why ANOVA?

We've done all that we can do with t-tests; it's now time to learn a new test statistic! Introducing ANOVA (ANalysis Of VAriance).

Type I Error

You may be wondering why we do not just conduct a t-test to test our hypotheses about three or more groups. After all, we are still just looking at group mean differences. The answer is that our t-statistic formula can only handle up to two groups, one minus the other. 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.

So, why is that a bad thing? Statistical software could run that in a jiffy. The real issue is our probability of committing a Type I Error. Remember Type I and Type II Errors? Briefly, a Type I error is a false positive; 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, we have the same Type I error rate for each analysis. The Type I error rate is cumulative, it is added together with each each new analysis. The more calculations we run, the more our Type I error rate increases. This raises 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.

Variability

The ANOVA is a cool analysis for other reasons, as well. As the name suggests, it looks at variances (meaning, the squared average difference between each score and the mean), but it looks at more than just the variance for each group (IV level).

What's also cool about ANOVAs is that we can use them to analyze mean differences on a DV when we have more than one IV, not just three or more groups (or levels) of one IV! We'll get to that later. For now, let's learn more about what ANOVAs do with variability.

This page titled 9.1: Why ANOVA? is shared under a not declared license and was authored, remixed, and/or curated by Michelle Oja.

- **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.

9.1.1: 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 **9.1.1.1**, 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).

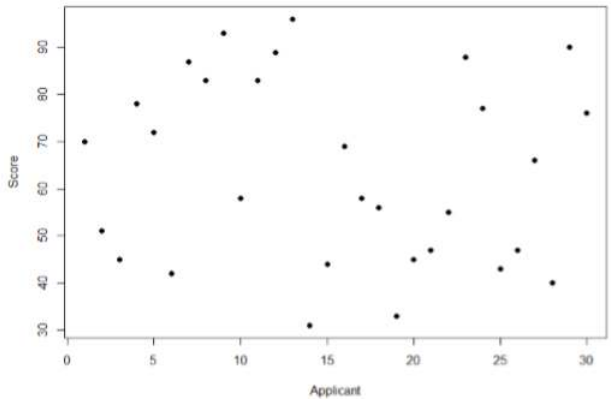


Figure **9.1.1.1**: Scores on a job test (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 a new kind of analysis to test if these groups are actually different. The new analysis is called ANOVA, standing for ANalysis Of VAriance. Figure **9.1.1.2** 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 **9.1.1.2**.

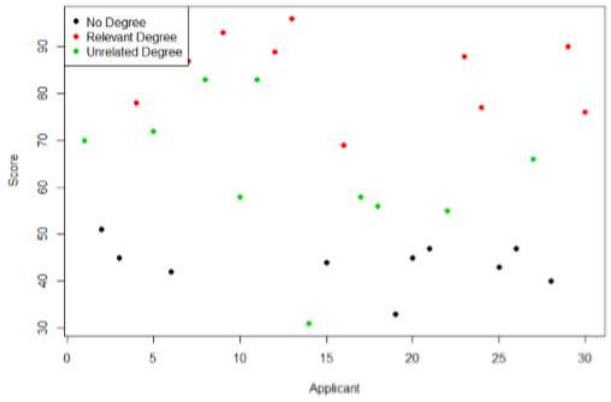


Figure **9.1.1.2**: Applicant scores coded by degree earned (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

This pattern is even easier to see when the applicants are sorted and organized into their respective groups, as shown in Figure 9.1.1.3.

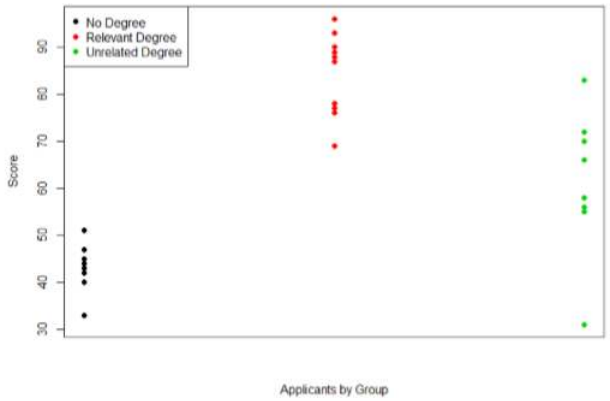


Figure 9.1.1.3: Applicant scores by group (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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.

An ANOVA is a great way to be able to compare the differences (variance) *between* the groups, while taking into account the differences (variances) *within* each group: ANalyis Of VAriance.

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.

Before we get into the calculations themselves, we must first lay out some important terminology and notation.

Variables & Notation

IV Levels

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 previous example, our grouping variable was education, which had three levels, so $K = 3$. When we report any descriptive value (e.g. mean, sample size, standard deviation) for a specific group or IV level, we will use a subscript 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.

To be more informative, it can be easier to use letters representing the group level names as subscripts instead of numbers. For example, if we have a high, medium, and low group, we could represent their means as: $\overline{X}_H$, $\overline{X}_M$, and $\overline{X}_L$. This makes

interpretation easier, as well, because you don't have to keep going back to see if the subscript of "1" was for the high group or if "1" represented the low group.

DV

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 $\overline{X_G}$, sometimes call the mean of the total ($\overline{X_T}$). These different means – the individual group means and the overall grand mean – will be how we calculate our sums of squares.

Samples

Finally, we now have to differentiate between several different sample sizes. Our data will now have sample sizes for each group, and we will often 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. You don't have to get too caught up with lower case and upper case issues, as long as you use the subscripts to accurately identify which group you are talking about. If you see an upper case N with a subscript (N_L), you can be pretty confident that it's talking about the sample size of the Low group, even though it should probably be a lower case n for the group. To make things clear, you can also include a subscript for the total sample size (N_T) instead of relying only on the capitalization to identify the groups.

Now that we've covered that, let's learn a little bit more about how we will use these different kinds of variability in the ANOVA.

This page titled 9.1.1: 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.

9.1.2: Ratio of Variability

Now that we know a little bit more about the different kinds of variability, or variances, let's learn what we can do with them.

ANOVA is a Ratio of Variances

The between-subjects ANOVA, is sometimes called a one-factor ANOVA, an independent factor ANOVA, or a one-way ANOVA (which is a bit of a misnomer). The critical ingredient for a one-factor (one IV), between-subjects (multiple groups) ANOVA, is that you have one independent variable, with at least two levels (at least two different groups in that one IV). You might be thinking, "When you have one IV with two levels, you can run a t-test." And you would be correct! You could also run an ANOVA. Interestingly, they give you almost the exact same results. You will get a p-value from both tests that is identical (they are really doing the same thing under the hood). The t-test gives a t-value as the important sample statistic. The ANOVA gives you the F-value (for Fisher, the inventor of the test) as the important sample statistic. It turns out that t^2 equals F, when there are only two groups in the design. They are the same test. Side-note, it turns out they are all related to Pearson's r too (which we'll discuss much later in this textbook).

Remember that t is the mean difference divided by the standard error of the sample. The idea behind F is the same basic idea that goes into making t. Here is the general idea behind the formula, it is again a ratio (division) of the effect we are measuring (in the numerator), and the variation associated with the effect (in the denominator).

$$\mathrm{F} = \frac{\text{measure of effect}}{\text{measure of error}}$$

This idea is the same as with the t-test. The difference with F, is that we use variances (how different, on average, each score is from the sample mean) to describe both the measure of the effect and the measure of error. So, F is a ratio of two variances.

When the variance associated with the effect is the same size as the variance associated with sampling error, we will get two of the same numbers, this will result in an F-value of 1 (a number divided by itself equals 1). When the variance due to the effect is larger than the variance associated with sampling error, then F will be greater than 1. When the variance associated with the effect is smaller than the variance associated with sampling error, F will be less than one. Let's rewrite in plainer English. We are talking about two concepts that we would like to measure from our data. 1) A measure of what we want to explain using our IV levels, and 2) a measure of error, or stuff about our data we can't explain by our IV levels. So, the F formula looks like this:

$$\mathrm{F} = \frac{\text{Can Explain by IV}}{\text{Can't Explain by IV}}$$

When what we can explain based on what group participants are in is as much as we can't explain, F = 1. This isn't that great of a situation for us to be in. It means we have a lot of uncertainty. When we can explain much more than we can't explain, we are doing a good job and F will be greater than 1. When we can explain less than what we can't, we really can't explain very much, F will be less than 1. That's the concept behind making F.

If you saw an F in the wild, and it was .6. Then you would automatically know the researchers couldn't explain much of their data. If you saw an F of 5, then you would know the researchers could explain 5 times more than the couldn't, that's pretty good. And the point of this is to give you an intuition about the meaning of an F-value, even before you know how to compute it.

Computing the F-value

Fisher's ANOVA is considered very elegant. It starts us off with a big problem that we always have with data. We have a lot of numbers, and there is a lot of variation in the numbers, what to do? Wouldn't it be nice to split up the variation into to kinds, or sources. If we could know what parts of the variation were being caused by our experimental manipulation (IV), and what parts were being caused by sampling error (wiggly samples), we would be making really good progress. We would be able to know if our IV was causing more change in the data than sampling error, or chance alone. If we could measure those two parts of the total variation, we could make a ratio, and then we would have an F value. This is what the ANOVA does. It splits the total variation in the data into two parts. The formula is:

$$\text{Total Variation} = \text{Variation due to IV levels (groups)} + \text{Variation due to sampling error}$$

This is a nice idea, but it is also vague. We haven't specified our measure of variation. What should we use?

Remember the sums of squares ? That's what we use. Let's take another look at the formula, using sums of squares for the measure of variation:

$$SS_{\text{Total}} = SS_{\text{Effect}} + SS_{\text{Error}}$$

We'll look at the exact formula for each sum of squares in the next section. Now, let's look at how we get from these sums of squares to the final calculated F-value.

Introduction to the ANOVA Table

We will go into more detail into the ANOVA Summary Table again, but here is your first introduction to understand the basic components and how they relate to each other. An ANOVA Summary Tables is provided in Table **9.1.2.1** for an example with three groups, A, B, and C. For example, we might have three scores in each group. To get from raw data the calculated F-value, you plug in the sums of squares for each kind of variation (between the groups, within each group, and the total) that we'll learn how to calculate next, with some the degrees of freedom for each type of variance (with new formulas!) to get the Mean Square for each type of variance. Then the ratio of the Mean Square for the Between Groups (MS_B) divided by the Mean Square for Within Groups (sometimes called "Error", MS_W) gives the final calculated F-value. All of these little pieces are conveniently organized in ANOVA Summary Tables.

Table **9.1.2.1**- Example ANOVA Summary Table

Source	SS	DF	MS	F
Between Groups	72	2	36.00	0.94
Within Groups (Error)	230	6	38.33	N/A
Total	302	8	N/A	N/A

There isn't anything special about the ANOVA table, it's just a way of organizing all the pieces. After conducting an ANOVA, you would provide this summary table in your write-up so that readers can see important information about your samples, error, and the effect of your IV.

Let's look through each column.

Sum of Squares

The SS column stands for Sum of Squares. We will get to the equations in the next section, but the basic idea is the same that we've idea that we've talked about since standard deviations. In general, sum of squares look at the sum of the difference from the mean, but squared before the summing to get rid of the negative values.

Degrees of freedom

DFs can be fairly simple when we are doing a relatively simple ANOVA like this one, but they can become complicated when designs get more complicated. Notice that each source has a difference degree of freedom, which means that you will need to calculate the DF for each source of variance (Between Groups, Within Groups or Error, and Total).

The formula for the degrees of freedom for SS_{BG} is $df_{\text{BG}} = \text{k} - 1$, where k is the number of groups in the design. In the example in Table **9.1.2.1**, there were three groups, so the $DF$is 3-1 = 2.

The formula for the degrees of freedom for $SS_{\text{WG or Error}}$ is $df_{\text{WG or Error}} = \text{N} - \text{k}$, or the number of scores minus the number of groups. We have 9 scores and 3 groups, so our df for the error term is 9-3 = 6.

The formula for the degrees of freedom for SS_{T} is $df_{\text{Total}} = \text{N} - 1$, or the number of scores minus 1; this is the degrees of freedom that you are used to. We have 9 scores, so our Total df is 9-1 = 8.

We are lucky to gave another computation check here because $df_{\text{Total}} = df_{\text{BG}} + df_{\text{WG or Error}}$. You will be tempted to not calculate one of these, and just use addition (or subtraction) to figure out the other ones but I caution you not to do this. There have been

plenty of times when I think that I have the DFs correct, but then try $df_{BG} + df_{WG\ or\ Error}$ and find that it does not equal the df_{Total}, which means that I messed up somewhere.

Mean Squared Error

The next column is MS, for Mean Square (or Mean Squared Error). To get the MS for each type of variance (between groups and within groups), we divide the SSes by their respective degrees of freedom. Remember we are trying to accomplish this goal:

$$\mathrm{F} = \frac{\text{measure of effect}}{\text{measure of error}}$$

We want to build a ratio that divides a measure of an effect by a measure of error. Perhaps you noticed that we already have a measure of an effect and error! The SS_{BG} and $SS_{WG\ or\ Error}$ both represent the variation due to the effect, and the leftover variation that is unexplained. Why don't we just do this?

$$\frac{SS_{BG}}{SS_{WG\ or\ Error}}$$

Well, of course you could do that, but the kind of number you would get wouldn't be readily interpretable like a t value or a z score. The solution is to *normalize* the SS terms. Don't worry, normalize is just a fancy word for taking the average, or finding the mean. Remember, the SS terms are all sums. And, each sum represents a different number of underlying properties.

For example, the SSBG represents the sum of variation for three means in our study. We might ask the question, well, what is the average amount of variation for each mean…You might think to divide SS_ by 3, because there are three means, but because we are estimating this property, we divide by the degrees of freedom instead (# groups - 1 = 3-1 = 2). Now we have created something new, it's called the MS_{BG}.

$$MS_{BG} = \frac{SS_{BG}}{df_{BG}}$$

$$MS_{BG} = \frac{72}{2} = 36$$

This might look alien and seem a bit complicated. But, it's just another mean. It's the mean of the sums of squares for the effect of the IV levels between the groups.

The $SS_{WG\ or\ Error}$ represents the sum of variation for nine scores in our study. That's a lot more scores, so the $SS_{WG\ or\ Error}$ is often way bigger than than SS_{BG}. If we left our SSes this way and divided them, we would almost always get numbers less than one, because the $SS_{WG\ or\ Error}$ is so big. What we need to do is bring it down to the average size. So, we might want to divide our $SS_{WG\ or\ Error}$ by 9, after all there were nine scores. However, because we are estimating this property, we divide by the degrees of freedom instead (the number of scores minus the number groups) = 9-3 = 6). Now we have created something new, it's called the $MS_{WG\ or\ Error}$.

$$MS_{WG\ or\ Error} = \frac{SS_{WG\ Error}}{df_{WG\ or\ Error}}$$

$$MS_{WG\ or\ Error} = \frac{230}{6} = 38.33$$

Calculate F

Now that we have done all of the hard work, calculating F is easy! Notice, the Mean Square for the effect (36) is placed above the Mean Square for the error (38.33) in the ANOVA Summary Table? That seems natural because we divide 36/38.33 to get the F-value!

$$\mathrm{F} = \frac{\text{measure of effect}}{\text{measure of error}}$$

$$\mathrm{F} = \frac{MS_{BG}}{MS_{WG\ or\ Error}}$$

$$\mathrm{F} = \frac{36}{38.33} = .94$$

Summary

So, that's how the ANOVA's F is a ratio of variability, and how you use the ANOVA Summary Table to calculate that ratio. We'll learn how to calculate each sum of squares next, then go back to the ANOVA Summary Table.

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 9.1.2: Ratio of Variability is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

9.2: Introduction to ANOVA's Sum of Squares

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.

Sums of Squares

Between Groups Sum of Squares

One source of variability we can identified in 11.1.3 of the previous example was differences or variability *between* the groups. That is, the groups seemed to have 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:

$$SS_B = \sum_{EachGroup} \left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 * \left(n_{group} \right) \right]$$

1. Subtract
2. Square
3. Multiply
4. Sum

I know, this looks a little extreme, but it really is what is says that it is, subtracting the mean of the total of all participants ($\overline{X}_T$) from the mean of one of the groups ($\overline{X}_{group}$), then squaring that subtraction. That gives you the difference score for that group. You then multiply that by the sample size for that group (n_{group}). You do that for each group ($\sum_{EachGroup}$). For example, if you had an IV with three levels (k = 3) that were High, Medium, and Low, the you'd do the parts in the brackets for each group, then add all three groups together for one final Between Groups Sum of Squares.

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 (Error)

The formula for this within groups 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:

$$SS_W = \sum_{EachGroup} \left[\sum \left(\left(X - \overline{X}_{group} \right)^2 \right) \right]$$

1. Subtract
2. Square
3. Sum

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. It is important to remember that the deviation score for each person is only calculated relative to their group mean; this is what ($X - \overline{X}_{group}$) is telling you to do: subtract the mean of a group from each score from that same group. You then square each of those subtractions. The sum ($\sum$) is to sum all of the individual squared scores of all of the groups.

Total Sum of Squares

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:

$$SS_T = \sum \left[\left(X - \overline{X}_T \right)^2 \right]$$

1. Subtract
2. Square
3. Sum

We can see that our Total Sum of Squares is just each individual score minus the grand mean (the mean of the totality of scores, $\overline{X}_T$). 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.

Computation Check!

An important feature of these calculations 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:

$$SS_T = SS_B + SS_W$$

This will prove to be a very convenient way to check your work! 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.

By Hand?

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 rarely be required to calculate the SS values by hand. Many professors will have you work out one problem on your own by hand, then either show you how to use statistical software or provide the Sums of Squares for you. However, you should still take the time to understand how they fit together and what each one represents to help understand the analysis itself; this will make it easier to interpret the results and make predictions (research hypotheses).

Contributors and Attributions

- Dr. MO (Taft College)

This page titled 9.2: Introduction to ANOVA's Sum of Squares is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

9.2.1: Summary of ANOVA Summary Table

Reminder

So far, we've discussed the:

- Variability between groups, showing how effective the IV was; this is shown in the SS_B
- Variability within groups, which masks how effective the IV was; this is shown in the SS_W, and is sometimes called "Error"
- Degrees of Freedom for Between Groups
- Degrees of Freedom for Within Groups
- The Mean Square for Between Groups, which is like an estimated average of the variability between the groups
- The Mean Square for Within Groups (Error), which is like an estimated average of the variability within groups
- The calculated F-score, which is a ratio of the variability between the groups and the variability within the group
- ANOVA Summary Table, which shows all of these!

ANOVA Summary Table

All of our sources of variability fit together in meaningful, interpretable ways as we saw, and the easiest way to do this is to organize them in the ANOVA Summary Table (Table **9.2.1.1**), which shows the formulas for everything other than the sum of squares, to calculate our test statistic.

Table **9.2.1.1**: ANOVA Table

Source	SS	df	MS	F
Between Groups	SS_B	$k-1$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_W}$
Within Groups (Error)	SS_W	$N-k$	$\frac{SS_W}{df_W}$	N/A
Total	SS_T	$N-1$	N/A	N/A

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 **9.2.1.1**. 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 usually 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 another 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. Let's look at hypotheses when we have three or more levels in our IV next.

You might notice a few cells in Table **9.2.1.1** that say "N/A" for "Not Applicable". In reality, these cells should be kept blank. However, to the table as accessible as possible, Dr. MO include "N/A" to indicate that you don't have to do anything for those cells. In your own ANOVA Summary Tables, you can leave them blank.

This page titled 9.2.1: Summary of ANOVA Summary Table is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

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

9.3: 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 a research hypothesis to lay out.

Research Hypotheses

Our research hypothesis for ANOVA is more complex with more than two groups. Let's take a look at it and then dive deeper into what it means.

What the ANOVA tests is whether there is a difference between any one set of means, but usually we still have expected directions of what means we think will be bigger than what other means. Let's work out an example. Let's say that my IV is mindset, and the three groups (levels) are:

- Growth Mindset
- Mixed Mindset (some Growth ideas and some Fixed ideas)
- Fixed Mindset

If we are measuring passing rates in math, we could write this all out in one sentence and one line of symbols:

- Research Hypothesis: Students with Growth Mindset with have higher average passing rates in math than students with either a mixed mindset or Fixed Mindset, but Fixed Mindset will have similar average passing rates to students with mixed mindset.
- Symbols: $\overline{X}_G > \overline{X}_M = \overline{X}_F$

But it ends up being easier to write out each pair of means:

- Research Hypothesis: Students with Growth Mindset with have higher average passing rates in math than students with a mixed mindset. Students with Growth Mindset with have higher average passing rates in math than students with a Fixed Mindset. Students with a Fixed Mindset will have similar average passing rates to students with mixed mindset.
- Symbols:
  - $\overline{X}_G > \overline{X}_M$
  - $\overline{X}_G > \overline{X}_F$
  - $\overline{X}_M = \overline{X}_F$

What you might notice is that one of these looks like a null hypothesis (no difference between the means)! And that is okay, as long as the research hypothesis predicts that at least one mean will differ from at least one other mean. It doesn't matter what order you list these means in; it helps to match the research hypothesis, but it's really to help you conceptualize the relationships that you are predicting so put it in the order that makes the most sense to you!

Why is it better to list out each pair of means? Well, look at this research hypothesis:

- Research Hypothesis: Students with Growth Mindset with have a similar average passing rate in math as students with a mixed mindset. Students with Growth Mindset with have higher average passing rates in math than students with a Fixed Mindset. Students with a Fixed Mindset will have similar average passing rates to students with mixed mindset.
- Symbols:
  - $\overline{X}_G = \overline{X}_M$
  - $\overline{X}_G > \overline{X}_F$
  - $\overline{X}_M = \overline{X}_F$

If you try to write that out in one line of symbols, it'll get confusing because you won't be able to easily show all three predictions. And if you have more than three groups, many research hypotheses won't be able to be represented in one line.

Another reason that this makes more sense is that each mean will be statistically compared with each other mean if the ANOVA results end up rejecting the null hypothesis. If you set up your research hypotheses this way in the first place (in pairs of means), then these pairwise comparisons make more sense later.

Null Hypotheses

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:

- Null Hypothesis: Students with Growth Mindset, mixed mindset, and Fixed Mindset will have similar average passing rates in math .
- Symbols: $\overline{X}_G = \overline{X}_M = \overline{X}_F$

You can list them all out, as well, but it's less necessary with a null hypothesis:

- Research Hypothesis: Students with Growth Mindset with have a similar average passing rate in math as students with a mixed mindset. Students with Growth Mindset with have a similar average passing rates in math than students with a Fixed Mindset. Students with a Fixed Mindset will have similar average passing rates to students with mixed mindset.
- Symbols:
  - $\overline{X}_G = \overline{X}_M$
  - $\overline{X}_G = \overline{X}_F$
  - $\overline{X}_M = \overline{X}_F$

Null Hypothesis Significance Testing

In our studies so far, when we've calculated an inferential test statistics, like a t-score, what do we do next? Compare it to a critical value in a table! And that's the same thing that we do with our calculated F-value. We compare our calculated value to our critical value to determine if we retain or reject the null hypothesis that all of the means are similar.

(Critical $<$ Calculated) = Reject null = At least one mean is different from at least one other mean. $= p < .05$

(Critical $>$ Calculated) = Retain null = All of the means are similar. $= p > .05$

What does Rejecting the Null Hypothesis Mean for a Research Hypothesis with Three or More Groups?

Remember when we rejected the null hypothesis when comparing two means with a t-test that we didn't have to do any additional comparisons; rejecting the null hypothesis with a t-test tells us that the two means are statistically significantly different, which means that the bigger mean was statistically significantly bigger. All we had to do was make sure that the means were in the direction that the research hypothesis predicted.

Unfortunately, with three or more group means, we do have to do additional statistical comparisons to test which means are statistically significantly different from which other means. The ANOVA only tells us that at least one mean is different from one other mean. So, rejecting the null hypothesis doesn't really tell us whether our research hypothesis is (fully) supported, partially supported, or not supported. When the null hypothesis is rejected, we will know that a difference exists somewhere, but we will not know where that difference is. Is Growth Mindset different from mixed mindset and Fixed Mindset, but mixed and Fixed are the same? Is Growth Mindset different from both mixed and Fixed Mindset? Are all three of them different from each other? And even if the means are different, are they different in the hypothesized direction? Does Growth Mindset always have a higher mean? We will come back to this issue later and see how to find out specific differences. For now, just remember that an ANOVA tests for any difference in group means, and it does not matter where that difference occurs. We must follow-up with any significant ANOVA to see which means are different from each other, and whether those mean differences (fully) support, partially support, or do not support the research hypothesis.

Table **9.3.1**- Interpreting Null Hypotheses

REJECT THE NULL HYPOTHESIS	RETAIN THE NULL HYPOTHESIS
Small p-values (p<.05)	**Large p-values (p>.05)**
A small p-value means a small probability that ***all*** of the means are similar. **Suggesting that *at least one of the means* is different from at least one other mean…**	A large p-value means a large probability that all of the means are similar.
We conclude that: • **At least one mean is different from one other mean.** • At least one group is *not* from the same population as the other groups.	We conclude that • The means for all of the groups are similar. • All of the groups *are* from the same population.
The calculated F is further from zero (more extreme) than the **critical F**. In other words, the calculated F is *bigger* than the **critical F**. (Draw the standard normal curve and mark the calculated F and the **critical F** to help visualize this.)	The calculated F is closer to zero (less extreme) than the **critical F**.In other words, the calculated F is *smaller* than the **critical F.** (Draw the standard normal curve and mark the calculated F and the **critical F** to help visualize this.)
Reject the null hypothesis (which says that all of the means are similar).	Retain (or fail to reject) the null hypothesis (which says that the all of the means are similar).
Support the Research Hypothesis? MAYBE. Look at the actual means: • *Support the Research Hypothesis if the means are in the directions that were hypothesized.* ◦ *The mean of the group that you said would be bigger, really is bigger;* ◦ *The mean of the group that you said would be smaller really is smaller;* ◦ *The means of the groups that you said would be similar are actually similar.* • *Partial support of the Research Hypothesis if some of the means are in the directions that were hypothesized, but some aren't.* • *Do not support the Research Hypothesis if none of the means are in the direction that were hypothesized.*	Do not support the Research Hypothesis (because all of the means are similar).
Statistical sentence: F(df) = F-calc, p<.05 (fill in the df and the calculated F)	Statistical sentence: F(df) = F-calc, p>.05 (fill in the df and the calculated F)

This page titled 9.3: Hypotheses in ANOVA is shared under a not declared license and was authored, remixed, and/or curated by Michelle Oja.

- **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.

9.4: Practice with Job Applicants

Let's practice our 4-step process using the ANOVA Summary Table to complete the calculations, but the Sum of Squares will be provided in this first example.

Scenario

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 in their scores on a test related to the job that they applied for, and, by doing so, see if there is a difference between our three group means.

To help understand what's going on in this scenario, let's answer some questions:

? Exercise 9.4.1

Answer the following questions to understand the variables and groups that we are working with.

1. Who is the sample?
2. Who do might be the population?
3. What is the IV (groups being compared)?
4. What is the DV (quantitative variable being measured)?

Answer

1. The sample is the 30 people with different degrees (or no degree)?
2. The population might be any job applicant?
3. The IV can be called "Degree," with the levels being:
    1. None,
    2. Related to job
    3. Unrelated to job
4. The DV is the score on the test.

Identifying the IV levels and DV helps when constructing your hypotheses.

Step 1: State the Hypotheses

Our hypotheses are concerned with the average score on the test for each of the groups based on education level, so you get to decide which groups you think will have a higher score, which groups will earn a lower average score, and which groups will have scores that are similar.

? Exercise 9.4.2

Determine the research hypothesis in words and symbols. You can fill in the following underlined spot with the symbols for greater than (>), less than (<), or equal signs. Just remember, at least one pair of means must be predicted to be different from each other.

Symbols:

- $\overline{X}_N$ _____ $\overline{X}_R$
- $\overline{X}_N$ _____ $\overline{X}_U$
- $\overline{X}_R$ _____ $\overline{X}_U$

Answer

Here's a reasonable research hypothesis. However, without the group means to guide us, your research hypothesis might be slightly different. Just remember, at least one pair of means must be predicted to be different from each other.

- Research Hypothesis: Those with No Degree will have a lower average test score than those with a Related Degree, but will have a similar average test score to those with an unrelated degree. The average test score for those with a Related Degree will also have a higher average test score compared to those with an Unrelated Degree.
- Symbols:
  - $\overline{X}_N < \overline{X}_R$
  - $\overline{X}_N = \overline{X}_U$
  - $\overline{X}_R > \overline{X}_U$

What about the null hypothesis?

? Exercise 9.4.3

State the null hypothesis in words and symbols.

Answer

- Null Hypothesis: The average test score will be similar for each group; the degree does not affect the hiring rate.
- Symbols: $\overline{X}_N = \overline{X}_U = \overline{X}_R$

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 shown in the next section.

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: for the Numerator because it is the degrees of freedom value used to calculate the Mean Square for the Between Groups source, which in turn was the numerator of our F statistic. Likewise, the df_W is the Degrees of Freedome for the Denominator because it is the degrees of freedom value used to calculate the Mean Square for the Within Groups (sometimes called Error) source, 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 fourth column, the one labeled 2, to find our critical value. To find the proper row, we 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 fourth column down to the row for df_W = 27, then find the middle row for p = 0.05, we see 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 calculated 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. We will use the Sum of Squares values that we are provided in Table 9.4.1.

Table 9.4.1: ANOVA Summary Table

Source	SS	df	MS	F
Between	8246			
Within	3020			
Total	11266			

These may seem like random numbers, but remember that they are based on the distances between the groups themselves and within each group. Figure 9.4.2 shows the plot of the data with the group means and total 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 are just trusting that these values are the correct ones.

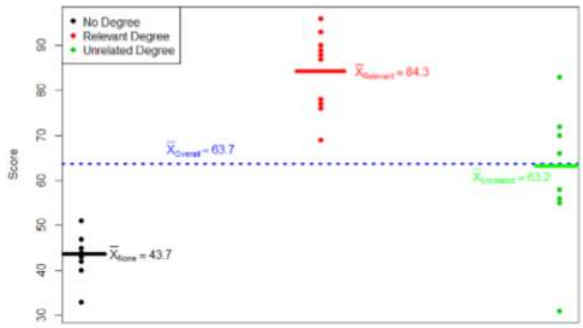


Figure **9.4.2**: Means (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Example 9.4.1

Using the information provided in the scenario and Table **9.4.1**, fill in the rest of the ANOVA Summary Table in Table **9.4.2**

Table **9.4.2**: ANOVA Summary Table

Source	SS	df	MS	F
Between	8246			
Within	3020			
Total	11266			

Solution

Using the formulas that we learned about earlier, we can complete Table **9.4.3**:

Table **9.4.3**- ANOVA Summary Table with Formulas

Source	SS	df	MS	F
Between Groups	8246	$k-1$ = 3 - 1 = 2	$\frac{SS_B}{df_B} = \frac{8246}{2} = 4123.00$	$\frac{MS_B}{MS_W} = \frac{4123.00}{111.85} = 36.86$
Within Groups (Error)	3020	$N-k$ = 30 - 3 = 27	$\frac{SS_W}{df_W} = \frac{3020}{27} = 111.85$	N/A
Total	11266	$N-1$ = 30 - 1 = 29	N/A	N/A

We leave those three empty cells blank; no information is needed from them. So, that leaves us with the final table looking like Table **9.4.4** showing calculated F-score of 36.86.

Table **9.4.4**- Completed ANOVA Summary Table

Source	SS	df	MS	F
Between Groups	8246	2	4123.00	36.86
Within Groups (Error)	3020	27	111.85	leave blank
Total	11266	29	leave blank	leave blank

We can move on to comparing our calculated value to the critical value in step 4.

Step 4: Make the Decision

Our calculated test statistic was found to be $F_{calc} = 36.86$ and our critical value was found to be $F_{crit} = 3.35$. Our calculated statistic is larger than our critical value, so we can reject the null hypothesis.

$$\textbf{(Critical < Calculated) = Reject null = At least one mean is different from at least one other mean. } = p < .05$$

$$\textbf{(Critical > Calculated) = Retain null = All of the means are similar. } = p > .05$$

Based on our three groups of 10 people, we can conclude that average 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.

Because we were only testing for any difference, we cannot yet conclude which groups are different from the others or if our research hypothesis was supported, partially supported, or not supported. We will learn about pairwise comparisons next to answer these last questions!

This page titled 9.4: Practice with Job Applicants is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

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

9.4.1: Table of Critical F-Scores

F Distribution

The tables following Figure **9.4.1.1** show the critical t-score. If the your calculated F-score is bigger (more extreme) than the critical F-score, then you reject the null hypothesis. Similarly, if the calculated F-score is to the right (in the shaded area), then the null hypothesis should be rejected. This means that there is a small probability (less than 5%, p<.05) that all of the means are similar (suggesting that at least one mean is different from one other mean). In contrast, if the calculated F-score is smaller than the critical value (to the right of the shaded area), then the null hypothesis is retained; there is a large probability (larger than 5%. p>.05) that the group means are similar.


Sample F-distribution that is a line graph that leans to the left, with the extreme section to the right of an F-score at the right tail labeled "Probability p" shaded to show the Rejection Region.Figure **9.4.1.1**- Critical Values for F (F^*) with probability p to its right. (CC-BY by Barbara Illowsky & Susan Dean (De Anza College) from OpenStax)

Note

Remember:

(Critical < Calculated) = Reject null = At least one mean is different from at least one other mean. $= p < .05$

(Critical > Calculated) = Retain null = All of the means are similar. $= p > .05$

Tables of Critical Values of F (ANOVA)

The critical values table for ANOVA is a little different than other tables of critical values because ANOVA is a ratio of variances between groups and variances within each group. This means that there are two degrees of freedom, one for the numerator and one for the denominator. Below the tables are the formulas for calculating the degrees of freedom for each (numerator and denominator) for each type of ANOVA. Because there are so many options with the combination of two degrees of freedom, there are two critical value tables.

Both tables show degrees of freedom of the numerator (which is based on the number of groups) in the columns; the options are from 1 to 10. If you have 11 or more groups (or combinations), use the column for 10. But honestly, if you have more than 10 groups, you should be using statistical software that will provide the probability so that you wouldn't have to use a table of critical values!

The p column, in italics in both tables, shows three different alphas (α), or probabilities. This textbook will always use the 5% alpha ($p = 0.05$, but your professor or your circumstances might indicate another alpha level is preferred.

Denominator 5 to 29

The first table (Table **9.4.1.1**) has degrees of freedom of the denominator for every degree of freedom 5 to 29. The degrees of freedom of the denominator are based on the sample size and the number of groups being compared.

Table **9.4.1.1**- Critical Values of F for Denominator DF's from 5 to 29

Degrees of Freedom		Degrees of Freedom for Numerator									
Degrees of Freedom for Denominator (Error)	p	1	2	3	4	5	6	7	8	9	10
5	*0.10*	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
5	*0.05*	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
5	*0.01*	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	*0.10*	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
6	*0.05*	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
6	*0.01*	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	*0.10*	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
7	*0.05*	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
7	*0.01*	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	*0.10*	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
8	*0.05*	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
8	*0.01*	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	*0.10*	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
9	*0.05*	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
9	*0.01*	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	*0.10*	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
10	*0.05*	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
10	*0.01*	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	*0.10*	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
11	*0.05*	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
11	*0.01*	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	*0.10*	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
12	*0.05*	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
12	*0.01*	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	*0.10*	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
13	*0.05*	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
13	*0.01*	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10

Degrees of Freedom		Degrees of Freedom for Numerator									
14	*0.10*	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
14	*0.05*	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
14	*0.01*	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	*0.10*	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
15	*0.05*	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
15	*0.01*	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	*0.10*	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
16	*0.05*	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
16	*0.01*	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	*0.10*	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
17	*0.05*	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
17	*0.01*	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	*0.10*	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
18	*0.05*	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
18	*0.01*	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	*0.10*	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	1.96
19	*0.05*	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	2.38
19	*0.01*	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	3.43
20	*0.10*	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
20	*0.05*	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
20	*0.01*	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	*0.10*	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
21	*0.05*	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
21	*0.01*	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	*0.10*	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
22	*0.05*	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
22	*0.01*	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	*0.10*	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
23	*0.05*	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
23	*0.01*	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	*0.10*	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88

Degrees of Freedom		Degrees of Freedom for Numerator									
24	*0.05*	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
24	*0.01*	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	*0.10*	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
25	*0.05*	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
25	*0.01*	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	*0.10*	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
26	*0.05*	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
26	*0.01*	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	*0.10*	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
27	*0.05*	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
27	*0.01*	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	*0.10*	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
28	*0.05*	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
28	*0.01*	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	*0.10*	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
29	*0.05*	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
29	*0.01*	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00

Denominator 30 and Above

The second table (Table **9.4.1.2**) has degrees of freedom of the denominator for every tenth degree of freedom from 30 t0 60, then 100, 200, and 1,000. The degrees of freedom of the denominator are based on the sample size and the number of groups being compared.

Table **9.4.1.2**- Critical Values of F for Denominator DF's 30, 40, 50, 60, 100, 200, and 1,000

Degrees of Freedom		Degrees of Freedom for Numerator									
Degrees of Freedom for Denominator (Error)	p	1	2	3	4	5	6	7	8	9	10
30	*0.10*	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
30	*0.05*	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
30	*0.01*	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	*0.10*	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
40	*0.05*	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
40	*0.01*	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
50	*0.10*	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73
50	*0.05*	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
50	*0.01*	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70
60	*0.10*	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
60	*0.05*	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
60	*0.01*	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
100	*0.10*	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66
100	*0.05*	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93
100	*0.01*	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50
200	*0.10*	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63
200	*0.05*	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88
200	*0.01*	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41
1000	*0.10*	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64	1.61
1000	*0.05*	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84
1000	*0.01*	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34

Because tables are limited by size, not all critical t-scores are listed. For example, if you had F(2, 49), then the degrees of freedom of the numerator would be 2 ($DF_N = 2$) and the degrees of freedom for the denominator would be 49 ($df_D = 49$). However, the DF for the denominator jumps from 40 to 50. There are a couple of options when your Degrees of Freedom is not listed on the table.

- One option is to use the Degrees of Freedom that is *closest* to your sample's Degrees of Freedom. For our example of F(2, 49), that would mean that we would use the df_D of 50 because 50 is closer to 49 than 40 is. That would mean that the critical F-score for F(2, 49) would be 3.18.

- Another option is to always we round down. For our example of F(2, 49), we use the df_D of 40 because it is the next lowest df_D listed. That would mean that the critical F-score for F(2, 49) would be 3.23. This option avoids inflating Type I Error (false positives).

Ask your professor which option you should use!

Whichever option you choose, your statistical sentence should include the actual degrees of freedom for the numerator and for the denominator, regardless of which number is listed in the table; the table is used to decide if the null hypothesis should be rejected or retained.

Degrees of Freedom

Between Groups ANOVA (independent groups)

- Numerator (effect of IV): k-1
- Denominator (error): N-k

Repeated Measures ANOVA (dependent groups)

- Numerator (effect of IV Between Groups): k-1
- Denominator (Within Groups or Error): $(k-1) \times (P-1)$
- Participants: P-1

Factorial ANOVA (2+ IVs)

1. Cells: $(k_1 \times k_2) - 1$
    1. Remembering that "k" is the number of groups, k_1 is the number of levels of the first IV and , k_2 is the number of levels of the other IV.
2. Between group for one variable ($IV1_1$): $k_1 - 1$
3. Between group for the other variable (IV_2): $k_2 - 1$
4. Interaction: $df_1 \times df_2$
5. Within group: $df_{Total} - df_{Cells}$
6. Total: N - 1
    1. With N being the number of scores.

Regression

- Model (numerator): 1
- Error (denominator): N-2
- Total: N-1

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.
- Dr. MO (Taft College)

This page titled 9.4.1: Table of Critical F-Scores is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by OpenStax.

9.5: Introduction to Pairwise Comparisons

Any time you run an ANOVA with more than two groups and end up with a significant effect (reject the null hypothesis), the first thing you'll want to do is see if the mean differences are in the direction that you predicted in your research hypothesis. This means that you'll have to find out which groups are actually different from one another; remember, the ANOVA only tells you that at least one mean is different from at least one other mean. In our job applicant example, our null hypothesis was that all three types of degrees (None, Related, and Unrelated) have the same average test score. But if you think about it, the null hypothesis is actually claiming *three* different things all at once here. Specifically, it claims that:

- $\overline{X}_N = \overline{X}_R$
- $\overline{X}_N = \overline{X}_U$
- $\overline{X}_R = \overline{X}_U$

If any *one* of those three claims is false, then the null hypothesis for the whole ANOVA is also false. So, now that we've rejected our null hypothesis, we're thinking that *at least* one of those things isn't true. But which ones? All three of these propositions are of interest; that's why the research hypothesis predicts how each pair of group means relates to one another. When faced with this situation, its usually helps to look at the data,. For instance, if we look at the plots in Figure 11.4.2, it's tempting to conclude that having a Related degree is better than having No degree, but it's not quite clear if having a Related degree results in a significantly higher average test score than having an Unrelated degree. If we want to get a clearer answer about this, it might help to run some tests.

Running "pairwise" t-tests

How might we go about solving our problem? Given that we've got three separate pairs of means ($\overline{X}_N$ versus $\overline{X}_R$; $\overline{X}_N$ versus $\overline{X}_U$; $\overline{X}_R$ versus $\overline{X}_U$) to compare, what we could do is run three separate t-tests and see what happens. If we go on to do this for all possible pairs of variables, we can look to see which (if any) pairs of groups are significantly different to each other. This "lots of t-tests idea" isn't a bad strategy, though as we'll see later on there are some problems with it. However, our current problem is that it's a *pain* to calculate all of these t-tests by hand. You might be asking if statistical software would do this, and the answer is yes! But still, if your experiment has 10 groups, then you would have to run 45 t-tests.

There's Always a "But..."

In the previous section it was hinted that there's a problem with just running lots and lots of t-tests. The concern is that when running these analyses, what we're doing is going on a "fishing expedition": we're running lots and lots of tests without much theoretical guidance, in the hope that some of them come up significant. This kind of theory-free search for group differences is referred to as post hoc analysis ("post hoc" being Latin for "after this").

It's okay to run post hoc analyses, but a lot of care is required. For instance, running a t-test for each pair of means is actually pretty dangerous: each *individual* t-test is designed to have a 5% Type I error rate (i.e., α=.05), and I ran three of these tests so now I have about 15% chance of rejecting the null hypothesis when it is really true (Type I Error). Imagine what would have happened if my ANOVA involved 10 different groups, and I had decided to run 45 "post hoc" t-tests to try to find out which ones were significantly different from each other! You'd expect 2 or 3 of them to come up significant *by chance alone*. As we saw in the chapter when we first learned about inferential statistics, the central organizing principle behind null hypothesis testing is that we seek to control our Type I Error rate, but now that I'm running lots of t-tests at once, in order to determine the source of my ANOVA results, my actual Type I Error rate across this whole setof tests has gotten completely out of control.

The usual solution to this problem is to introduce an adjustment to the p-value, which aims to control the total error rate across the set of tests. There are different ways to do this adjustment. We'll discuss some common analyses in this section, but you should be aware that there are many other methods out there.

Corrections of p-values with Multiple Comparisons

These first two post-hoc analysis focus on calculating a probability based on the raw p-values from analyses conducted by statistical software, although these calculations can easily be done by hand.

Bonferroni Corrections

The simplest of these adjustments is called the Bonferroni correction, and it's very very simple indeed. Suppose that my post hoc analysis consists of "m" separate tests (in which "m" is the number of pairs of means you need to compare), and I want to ensure that the total probability of making *any* Type I errors at all is a specific alpha (α), such as 0.05. If so, then the Bonferroni correction just says "multiply all your raw p-values by m". If we let p denote the original p-value, and let p_B be the corrected value, then the Bonferroni correction tells that:

$$p_{Bonferroni} = m \times p$$

If you're using the Bonferroni correction, you would reject the null hypothesis if your Bonferroni probability is smaller than the alpha $(p_{Bonferroni} < \alpha)$.

Example 9.5.1

What would this look like for our job applicant scenario if the raw p-value was .004 for the difference between No Degree and a Related Degree?

Solution

$$p_{Bonferroni} = m \times p$$

We are making three comparisons ($\overline{X}_N$ versus $\overline{X}_R$; $\overline{X}_N$ versus $\overline{X}_U$; $\overline{X}_R$ versus $\overline{X}_U$), so m = 3.

$$p_{Bonferroni} = 3 \times 0.004$$

$$p_{Bonferroni} = 0.012$$

Because our Bonferroni probability (p_B) is smaller than our typical alpha $(\alpha)(0.012 < 0.05)$, we reject the null hypothesis that this set of pairs (the one with a raw p-value of .004)

The logic behind this correction is very straightforward. We're doing m different tests; so if we arrange it so that each test has a Type I error rate of at most α/m, then the *total* Type I error rate across these tests cannot be larger than α. That's pretty simple, so much so that in the original paper, the author writes:

> *The method given here is so simple and so general that I am sure it must have been used before this. I do not find it, however, so can only conclude that perhaps its very simplicity has kept statisticians from realizing that it is a very good method in some situations (pp 52-53 Dunn 1961)*

Holm Corrections

Although the Bonferroni correction is the simplest adjustment out there, it's not usually the best one to use. One method that is often used instead is the ***Holm correction*** (Holm, 1979). The idea behind the Holm correction is to pretend that you're doing the tests sequentially; starting with the smallest raw p-value and moving onto the largest one. For the j-th largest of the p-values, the adjustment is *either*

$$p'_j = j \times p_j$$

(i.e., the biggest p-value remains unchanged, the second biggest p-value is doubled, the third biggest p-value is tripled, and so on), *or*

$$p'_j = p'_{j+1}$$

whichever one is *larger*. This might sound a little confusing, so let's go through it a little more slowly. Here's what the Holm correction does. First, you sort all of your p-values in order, from smallest to largest. For the smallest p-value all you do is multiply it by m, and you're done. However, for all the other ones it's a two-stage process. For instance, when you move to the second smallest p value, you first multiply it by m−1. If this produces a number that is bigger than the adjusted p-value that you got last

time, then you keep it. But if it's smaller than the last one, then you copy the last p-value. To illustrate how this works, consider the table below, which shows the calculations of a Holm correction for a collection of five p-values:

Table **9.5.1**-Holm Calculations and p-values

raw p	rank j (m)	p×j	Holm p
.001	5	.005	.005
.005	4	.020	.020
.019	3	.057	.057
.022	2	.044	.057
.103	1	.103	.103

Hopefully that makes things clear.

Although it's a little harder to calculate, the Holm correction has some very nice properties: it's more powerful than Bonferroni (i.e., it has a lower Type II error rate), but – counterintuitive as it might seem – it has the *same* Type I error rate. As a consequence, in practice there's never any reason to use the simpler Bonferroni correction, since it is always outperformed by the slightly more elaborate Holm correction.

Those are the types of corrections to the p-values that can be done to make sure that you don't accidently commit a Type I Error. Next, we'll cover post-hoc analyses that find a critical value for the differences between each pair of means.

Reference

Dunn, H. L. (1961). *High-level wellness: A collection of twenty-nine short talks on different aspect os the them "High Level Wellness for Man and Society"*. Arlington, VA: Beatty.

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)

This page titled 9.5: Introduction to Pairwise Comparisons is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

9.5.1: Pairwise Comparison Post Hoc Tests for Critical Values of Mean Differences

As just discussed , a post hoc test is used only after we find a statistically significant (reject the null hypothesis) result and need to determine where our differences truly came from. This implies that when the null hypothesis is retained, you do not need to conduct pairwise comparisons; if there's no differences between the means, why would you look for where a mean difference is? The term “post hoc” comes from the Latin for “after the event”.

Mean Differences

The next set of post-hoc analyses compare the difference between each pair of means, then compares that to a critical value. Let's start by determining the mean differences. Table **9.5.1.1** shows the mean test scores for the three IV levels in our job applicant scenario.

Table **9.5.1.1**- Average Test Scores by Job Applicant Degree

IV Levels	Average Test Score ($\overline{X}$)
No Degree	43.7
Related Degree	84.3
Unrelated Degree	63.2

Using the average test scores for each level of the IV (Degree), we could find the difference between each pair of means:

$$\overline{X}_N - \overline{X}_R = 43.7–84.3 = -40.6$$

$$\overline{X}_N - \overline{X}_U = 43.7–63.2 = -19.5$$

$$\overline{X}_R - \overline{X}_U = 84.3–63.2 = 21.1$$

We could avoid negative answers if we always put the bigger number first, but Dr. MO likes to follow the order from the research hypotheses. As with all critical values, we use the absolute value of the calculated statistic anyway, so the big issue is to remember is to make sure that you subtract each mean from each other mean.

That’s it on mean differences, let’s now learn about some ways to calculate a critical value to compare these mean differences to; as always:

$$Critical < |Calculated| = \textbf{Reject null} = \textbf{means are different} = p < .05$$

$$Critical > |Calculated| = \textbf{Retain null} = \textbf{means are similar} = p > .05$$

Pairwise Comparisons

For this type of post-hoc analysis, you compare each of these mean differences (that you just calculated by subtracting one mean from another mean) to a critical value. What should you do if the calculated mean difference is further from zero (bigger) than the critical value? Yep, you reject the null hypothesis and say that those two means are different from each other! Sound familiar? It’s just like when using a critical z, critical t, or critical F.

As with converting the raw p-values, the critical mean difference can be computed in different ways. Tukey’s Honestly Significant Difference will be discussed here, but just know that there are other types of pairwise comparison tests that statistical software can complete with ease.

Pairwise Comparison Steps:

1. Compute a mean difference for each pair of variables.
2. Find the critical mean difference.
3. Compare each calculated mean difference to the critical mean.
4. Decide whether to retain or reject the null hypothesis for that pair of means.

Tukey's HSD (Honestly Significant Difference)

Tukey's HSD corrects for the alpha inflation caused by doing a bunch of statistical tests by correcting for the probability of a Type I error so that all the pairwise comparisons *combined* is p<.05, sorta like the Bonferroni that we just discussed. Tukey's makes it so that each individual pairwise comparison has a much smaller p-value.

Formula:

$$HSD = q * \sqrt{\frac{MSw}{n_{group}}}$$

There are a couple things to know about this formula. First, the q is found in yet another table. You probably won't need this too much because statistical software tells you which means are different from which other means, but you can find a table of q-values at this Real-Statistics.com webpage: https://www.real-statistics.com/statistics-tables/studentized-range-q-table/; make sure that you use the Alpha = 0.05 set of tables!

To find the q-value in the table, you find the column with the total number of groups in the analysis (k) on the top, then find the Degrees of Freedom for the denominator (df_W) on the side. Remember, there are different tables for the different alpha levels so don't just use the first table.

Example 9.5.1.1

What would the critical value (Tukey's HSD) be for our job applicant scenario?

Solution

Let's first find the q-value. We know that we have 3 groups (k = 3), and 30 participants. The df_W is calculated as N – k, so that's 30 – 3 = 27. According to the Alpha = 0.05 table at the table of q-values, q = 3.506.
Now we can use the q-value we just found and information from the ANOVA Summary Table to complete this formula:

$$HSD = q * \left(\sqrt{\frac{MSw}{n_{group}}}\right)$$

$$HSD = 3.506 * \left(\sqrt{\frac{111.85}{10}}\right)$$

$$HSD = 3.506 * \left(\sqrt{11.185}\right)$$

$$HSD = 3.506 * (3.34)$$

$$HSD = 11.73$$

Okay, but what do we *do* with this number? Tukey's HSD is a critical value, so if any of your mean differences are bigger than this critical value of 3.82, then the null hypothesis is rejected and that set of means is statistically significantly different from each other. If the critical value (Tukey's HSD that you just calculated) is bigger than any of the differences between the means, then you retain the null hypothesis and say that they are not different enough to think that they are from different populations.

$$Critical < |Calculated| = \textbf{Reject null} = \textbf{means are different} = p < .05$$

$$Critical > |Calculated| = \textbf{Retain null} = \textbf{means are similar} = p > .05$$

Let's look at that in Table 9.5.1.2.

Table **9.5.1.2-** Interpreting Pairwise Comparison Results

Means Compared	Mean Difference	Is Tukey's HSD of 11.73 Smaller than the Absolute Value of the Mean Difference?	Reject or Retain the Null Hypothesis?	Are the Means Similar or Different?
$\overline{X}_N - \overline{X}_R = 43.7 - 84.3$	-40.6	Yes	Reject the null hypothesis.	These two means are different from each other.
$\overline{X}_N - \overline{X}_U = 43.7 - 63.2$	-19.5	Yes	Reject the null hypothesis.	These two means are different from each other.
$\overline{X}_R - \overline{X}_U = 84.3 - 63.2$	21.1	Yes	Reject the null hypothesis.	These two means are different from each other.

But what does it mean?

To complete this analysis of the job applicant scenario, we need to go back to the research hypothesis. We had said that the applicants with No Degree will have a lower average test score than those with a Related Degree, but those with No Degree will have a similar average test score to those with an Unrelated degree. The average test score for those with a Related Degree will also have a higher average test score compared to those with an Unrelated Degree. In symbols, this looks like:

- $\overline{X}_N < \overline{X}_R$
- $\overline{X}_N = \overline{X}_U$
- $\overline{X}_R > \overline{X}_U$

Now, let's look at our mean differences. All of them were statistically significantly different, which means that our research hypothesis was partially supported.

- RH1: $\overline{X}_N < \overline{X}_R$ = Supported; those without a degree scored statistically significantly lower than those with a related degree.
- RH2: $\overline{X}_N = \overline{X}_U$ = Not supported; the research hypothesis was that the group without a degree would have average test scores that were similar to the group with an unrelated degree, but those with an unrelated degree actually scored higher than those with no degree.
- RH3: $\overline{X}_R > \overline{X}_U$ = Supported; those with a related degree did score higher than those with an unrelated degree.

And that leads us to our final write-up!

Write-Up

? Exercise 9.5.1.1

Write up a conclusion to the job applicant scenario.

Answer

Look at your conclusion. Did you include the four components that should be in all write-ups?

- The statistical test is preceded by the descriptive statistics (means).
- The description tells you what the research hypothesis being tested is.
- A "statistical sentence" showing the results is included.
- The results are interpreted in relation to the research hypothesis.

Look at Dr. MO's conclusion.

> *The research hypothesis that the applicants with No Degree will have a lower average test score than those with a Related Degree, but will have a similar average test score to those with an unrelated degree, and that the average test score for those with a Related Degree will also have a higher average test score compared to those with an Unrelated Degree was partially supported ($F(2,27) = 36.86$, $p < 0.05$. Those with a Related Degree ($M = 84.3$) did score higher than both those with an Unrelated Degree ($M = 63.2$) and those with No Degree ($M = 43.7$). However, it was hypothesized that those with an Unrelated Degree would have similar average test scores than those with No Degree, when those with an Unrelated Degree actually scored significantly higher than those with No Degree.*

Note

Did Dr. MO include the four components that should be in all write-ups?

Make sure to include the statistical sentence for the ANOVA with both Degrees of Freedom (with the smaller one from the numerator first), but you don't have to include all of that information for each pairwise comparison of the mean differences.

Summary

There are many more post hoc tests than just the ones discussed here, 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.

So, that's it! You've learned a Between Groups ANOVA and pairwise comparisons to test the null hypothesis! Let's try one full example next!

This page titled 9.5.1: Pairwise Comparison Post Hoc Tests for Critical Values of Mean Differences is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **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.

9.6: Practice on Mindset Data

Okay, here's your time to practice the whole process of a Between Groups ANOVA, including each Sum of Squares formula and the pairwise comparisons!

Scenario

Dr. MO helped student researchers try to improve the mindset of their fellow community college students (If you need a refresher, check out the Growth Mindset section, or search online for "growth mindset"). Three different professors (all women) tried three different interventions (shown below) to see if they could improve mindset by the end of the semester. The three intervention levels were:

- No Intervention: This was the comparison control group. It was a behavioral statistics course (like you are taking right now!). The instructor had students complete the Mindset Quiz as part of class related to surveys and data collection, but the professor never explained what Growth Mindset was or how it helps students succeed. No class activities or assignments were related to Growth Mindset.
- Minimal Intervention: This was an Early Childhood Education class that explained what Growth Mindset was, and how it could help both the students themselves, and the children that they would be taking care of or teaching. The course had a few assignments and class activities related to Growth Mindset.
- Super Intervention: This was a History course. This professor explained Growth Mindset and how it could help them succeed, as well as weekly readings, discussions, and activities related to Growth Mindset.

Students who enrolled in these courses were participants in one of the three different interventions, and filled out the Mindset Quiz at the end of the semester. Mindset Quiz scores can range from 20 to 60, with higher scores showing that the student has more Growth Mindset (and less Fixed Mindset). At the end of the semester, 10 students in each course completed the Mindset Quiz and had attended most classes throughout the semester to experience any activities related to mindset (N = 30).

After reading the scenario, can you describe the sample, population, IV, and DV?

? Exercise 9.6.1

Answer the following questions to understand the variables and groups that we are working with.

1. Who is the sample?
2. Who do might be the population?
3. What is the IV (groups being compared)?
4. What is the DV (quantitative variable being measured)?

Answer

1. 30 community college students.
2. Community college students? Community college students with women professors?
3. Intervention levels: None, Minimal, Super
4. Mindset Quiz

Step 1: State the Hypotheses

Using the following the research question and the descriptive statistics in Table **9.6.1**, what could be the research hypothesis?

> **Research Question**: The student researchers were interested to know if Growth Mindset could be taught, and how much time the instructor needed to devote to these intervention activities to improve growth mindset.

Table **9.6.1**- Descriptive Statistics for each IV level and the Total

	N:	Mean:	SD:	Median:
No Intervention	10	42.70	8.18	43.00
Minimal Intervention	10	44.40	5.87	43.00
Super Intervention	10	51.10	4.93	50.50
Total	30	46.07	7.25	48.00

Note

You are encourages to use the raw data found in Table **9.6.2** in Example **9.6.2** to practice calculating the means, standard deviations, and medians yourself!

What's your research hypothesis? Remember, the research hypothesis should predict how all levels of the IV related to all other levels of the IV.

Exercise 9.6.2

What is a research hypothesis in words and symbols?

Answer

- Research Hypothesis: Students in the No Intervention group average score will be lower on the Mindset Quiz than both students in the Minimal Intervention Group and the Super Intervention group. Similarly, students in the Minimal Intervention group average score will be lower than students in the Super Intervention group.
- Symbols:
  - $\overline{X}_N < \overline{X}_M$
  - $\overline{X}_N < \overline{X}_S$
  - $\overline{X}_M < \overline{X}_S$

Your research hypothesis might be slightly different than this one. Just make sure that at least one mean is predicted to be different than one other mean.

What is the null hypothesis with this scenario?

Exercise 9.6.3

State the null hypothesis in words and symbols.

Answer

- Null Hypothesis: Students in the No Intervention group will have a similar average score on the Mindset Quiz as both students in the Minimal Intervention Group and the Super Intervention group. Similarly, students in the Minimal Intervention group will have a similar average score as students in the Super Intervention group.
- Symbols:
  - $\overline{X}_N = \overline{X}_M$
  - $\overline{X}_N = \overline{X}_S$
  - $\overline{X}_M = \overline{X}_S$

Step 2: Find the Critical Values

This step might be easier after you've completed the ANOVA Summary Table because you will have the Degrees of Freedom for both groups, but let's try it now.

? Exercise 9.6.3

Using the Critical Values of F Table by going to the page or finding the link in the Common Critical Values page at the end of the textbook, what is the critical value at α = 0.05 with three groups and a total of 30 people?

Answer

Critical F(2,27)=3.35

The first Degree of Freedom (2), for the numerator (df_B) is found through: k–1, with k being the number of groups.

The second Degree of Freedom (27), for the denominator (df_W) is found through $N–k(30–3 = 27)$. A common mistake is to use the Degrees of Freedom of the numerator instead of the number of groups..

Step 3: Compute the Test Statistic

If you will never have to calculate the Sums of Squares by hand, skip this part and just fill in the ANOVA Summary Table (Table **9.6.4**) at the end of this section. If you are practicing the Sums of Squares, each Sum of Square will have it's own Example. Heads up, to do all of these can take about an hour!

Let's go!

✔ Example 9.6.1

Calculate the Between Groups Sums of Squares.

Solution

$$SS_B = \sum_{EachGroup} \left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 * (n_{group}) \right]$$

The $\sum_{EachGroup}$ means that you do everything following that for each intervention level, then add them all together. Let's start with what's inside the brackets for the No Intervention group.

$\sum_{EachGroup}$

$$SS_B = \sum_{EachGroup} \left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 * (n_{group}) \right]$$

$$\text{No Intervention} = \left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 * (n_{group}) \right]$$

$\overline{X}_{group}$ is asking for the mean of the group that we're looking at, so the No Intervention group right now. $\overline{X}_T$ is asking for the total mean, the mean for all 30 scores. Both of these means were provided in Table **9.6.1**. The number of scores in the group that we're looking at right now is what n_{group} is asking.

So let's plug those all in!

$$\text{No Intervention} = \left[(42.70 - 46.07)^2 * (10) \right]$$

$$\text{No Intervention} = \left[(-3.37)^2 * (10) \right]$$

The mean of the group minus the mean of the sample should be negative, but that negative sign goes away when we square it:

$$\text{No Intervention} = [(11.36) * (10)]$$

$$\text{No Intervention} = [113.60]$$

And let's do that process two more times, once for the Minimal Intervention Group and once for the Super Intervention group, and then we'll add those three numbers together to get the SS_B.

$$\text{Minimal Intervention} = \left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 * (n_{group}) \right]$$

$$\text{Minimal Intervention} = \left[(44.40 - 46.07)^2 * (10)\right]$$

$$\text{Minimal Intervention} = \left[(-1.67)^2 * (10)\right]$$

$$\text{Minimal Intervention} = [(2.79) * (10)]$$

$$\text{Minimal Intervention} = [27.90]$$

Why don't you try to do it on your own for the Super Intervention group?

$$\text{Super Intervention} = \left[\left(\overline{X}_{group} - \overline{X}_T\right)^2 * (n_{group})\right]$$

$$\text{Super Intervention} = [253.10]$$

Next step, add them all together! It's easy to forget this step, but the Sum of Squares ends up to be one number, so when you get lost or forget the next step, look back at the full formula:

$$SS_B = \sum_{EachGroup} \left[\left(\overline{X}_{group} - \overline{X}_T\right)^2 * (n_{group})\right]$$

$$SS_B = \sum_{EachGroup} = [113.60 + 27.90 + 253.10]$$

$$SS_B = 394.60$$

You did it! Only two more Sums of Squares to go!

Example 9.6.2

Calculate the Within (Error) Sum of Squares.

Solution

$$SS_W = \sum \left[\left(X - \overline{X}_{group}\right)^2\right]$$

As before, complete the calculations in the brackets first for each group, then add them all together. To complete the calculations in the brackets, it's easiest to use a table with all of the values since you subtract the mean of the group that you're working with from each score. Do you remember doing this with standard deviations?

Table **9.6.2** shows the group mean subtracted from each score in the column to the right of the raw scores, then that is squared in the column to the next right. The squared values are then summed for each group.

Table **9.6.2**- Group Mean Subtracted from Each Score, Squared, then Summed

No Intervention	**minus group mean**	**squared**	Minimal Intervention	**minus group mean**	**squared**	Super Intervention	**minus group mean**	**squared**
30	-12.70	161.29	37	-7.40	54.76	43	-8.10	65.61
35	-7.70	59.29	38	-6.40	40.96	45	-6.10	37.21
36	-6.70	44.89	39	-5.40	29.16	48	-3.10	9.61
36	-6.70	44.89	42	-2.40	5.76	50	-1.10	1.21
43	0.30	0.09	42	-2.40	5.76	50	-1.10	1.21
43	0.30	0.09	44	-0.40	0.16	51	-0.10	0.01
49	6.30	39.69	48	3.60	12.96	54	2.90	8.41
49	6.30	39.69	48	3.60	12.96	56	4.90	24.01
53	10.30	106.09	53	8.60	73.96	56	4.90	24.01
53	10.30	106.09	53	8.60	73.96	58	6.90	47.61
	Sum:	**602.10**		**Sum:**	**310.40**		**Sum:**	**218.90**
$\overline{X}_N$ = **42.7**	N/A	N/A	$\overline{X}_M$ = **44.4**	N/A	N/A	$\overline{X}_S$ = **51.1**	N/A	N/A

Now that we have the sum of each squared score of the subtraction, we can finish the formula:

$$SS_W = \sum_{EachGroup} \left[\sum \left((X - \overline{X}_{group})^2\right)\right]$$

$$SS_W = \sum_{EachGroup} = [602.10 + 310.40 + 218.90]$$

$$SS_W = 1131.40$$

And, on to the final Sum of Squares!

Example 9.6.3

Calculate the Total Sum of Squares.

Solution

$$SS_T = \sum \left[(X - \overline{X}_T)^2\right]$$

This formula is also saying to subtract a mean from each score, but this time we should be subtracting the Total mean ($\overline{X}_T$ = 46.07, found in Table **9.6.1**). This is again easiest to compute in a table.

Table **9.6.3** shows the Total mean subtracted from each score in the column to the right of the raw scores, then that is squared in the column to the next right. The squared values are then summed for *all* of the scores.

Table **9.6.3**- Total Mean Subtracted from Each Score, Squared, then Summed

IV Levels	Mindset Quiz Scores	minus Total mean	squared
No Intervention	30	-16.07	258.24
No Intervention	35	-11.07	122.54
No Intervention	36	-10.07	101.40
No Intervention	36	-10.07	101.40
No Intervention	43	-3.07	9.42
No Intervention	43	-3.07	9.42
No Intervention	49	2.93	8.58
No Intervention	49	2.93	8.58
No Intervention	53	6.93	48.02
No Intervention	53	6.93	48.02
Minimal Intervention	37	-9.07	82.26
Minimal Intervention	38	-8.07	65.12
Minimal Intervention	39	-7.07	49.98
Minimal Intervention	42	-4.07	16.56
Minimal Intervention	42	-4.07	16.56
Minimal Intervention	44	-2.07	4.28
Minimal Intervention	48	1.93	3.72
Minimal Intervention	48	1.93	3.72
Minimal Intervention	53	6.93	48.02
Minimal Intervention	53	6.93	48.02
Super Intervention	43	-3.07	9.42
Super Intervention	45	-1.07	1.14
Super Intervention	48	1.93	3.72
Super Intervention	50	3.93	15.44
Super Intervention	50	3.93	15.44
Super Intervention	51	4.93	24.30
Super Intervention	54	7.93	62.88
Super Intervention	56	9.93	98.60
Super Intervention	56	9.93	98.60
Super Intervention	58	11.93	142.32
σ	N/A	$\sigma = 1525.87$	

$$SS_T = \sum \left[\left(X - \overline{X}_T \right)^2 \right] = 1525.87$$

Yay! You finished the Sum of Squares! But before we move on, let's do the computation check to see if we did the Sums of Squares correctly: $SS_T = SS_B + SS_W$

Let's check:

$$SS_T = SS_B + SS_W = 394.40 + 1131.40 = 1526.00$$

This is close to our SS_T of 1525.87, so Dr. MO checked to make sure all of the calculations were done correctly with both a calculator and with Excel; it appears that this slight discrepancy is due to rounding differences.

But we're not done yet! Now, plug those three Sums of Squares into the ANOVA Summary Table (Table 9.6.4) so that you can fill out the rest of the table to calculate the final ANOVA F-value.

Table 9.6.4- ANOVA Summary Table with SS only

Source	SS	df	MS	F
Between	394.60			
Within	1131.40			
Total	1525.87			

Example 9.6.4

Fill out the rest of the ANOVA Summary Table.

Solution

Table 9.6.5- ANOVA Summary Table with Formulas

Source	SS	df	MS	F
Between	394.60	$k - 1 = 3–1 = 2$	$\frac{SS_B}{df_B} = \frac{394.60}{2} = 197.30$	$\frac{MS_B}{MS_W} = \frac{197.30}{41.90} = 4.71$
Within	1131.40	$N - k = 30–3 = 27$	$\frac{SS_W}{df_W} = \frac{1131.40}{27} = 41.90$	leave blank
Total	1525.87	$N - 1 = 30–1 = 29$	leave blank	leave blank

If you do a computation check (which you should) of the degrees of freedom, you find that:

$$df_T = df_B + df_W = 2 + 27 = 29$$

Since we did that part correctly, we can move on to the next step of the process.

Step 4: Make the Decision

Based on the completed ANOVA Summary Table (Table 9.6.5), our calculated F-score is 4.71. If you remember all the way back to Step 2, we found our critical F-score to be 3.35.

(Critical < Calculated) = Reject null = At least one mean is different from at least one other mean. = $p < .05$

(Critical > Calculated) = Retain null = All of the means are similar. = $p > .05$

Since our critical value is smaller than our calculated value, we reject the null hypothesis that all of the means are similar, which means that at least one mean is different from at least one other mean. That's not specific enough to evaluate whether our research hypothesis is correct or not, so on to pairwise comparisons!

Pairwise Comparisons

Let's start with finding the differences between each pair of means.

Example 9.6.5

What are the mean differences for each pair of means?

Solution

$$\overline{X}_N - \overline{X}_M = 42.70–44.40 = -1.7$$

$$\overline{X}_N - \overline{X}_S = 42.7–51.10 = -8.4$$

$$\overline{X}_M - \overline{X}_S = 44.40–51.10 = -6.7$$

Great! Now, let's compute Tukey's HSD to see if any of these mean differences are big enough to say that they are statistically significantly different.

Example 9.6.6

Using the q-value for our degrees of freedom from the Alpha = 0.05 table from Real-Statistics.com's q table, compute Tukey's HSD. If you will never have to calculate a post-hoc analysis by hand, skip this part and use the calculated Tukey's HSD to make your decisions and write-up the conclusion.

Solution

From the critical q table, we find that for our study with 3 groups and the degrees of freedom for the denominator (df_W = 27), the correct q-score to plug into the formula is 3.506. The other information is in the completed ANOVA Summary Table (Table 9.6.5).

$$HSD = q * \sqrt{\frac{MSw}{n_{group}}}$$

$$HSD = 3.506 * \sqrt{\frac{41.90}{10}}$$

$$HSD = 3.506 * \sqrt{4.19}$$

$$HSD = 3.506 * 2.05$$

$$HSD = 7.18$$

Now what?

Well, you have a critical value (Tukey's HSD) and some mean differences, let's compare them!

(Critical < Calculated) = Reject null = At least one mean is different from at least one other mean. = $p < .05$

(Critical > Calculated) = Retain null = All of the means are similar. = $p > .05$

Example 9.6.7

Using the Tukey's HSD of 7.18, which means are statistically significantly different from each other?

Solution

Our pairwise comparison critical value is 7.18. Of the absolute value of the mean differences, that means that the Super Intervention was statistically significantly higher than the No Intervention group's average score on the Mindset Quiz, but none of the other pairs of means were significantly different.

$$\overline{X}_N - \overline{X}_M = 42.70–44.40 = -1.7$$

$$\overline{X}_N - \overline{X}_S = 42.7–51.10 = -8.4$$

$$\overline{X}_M - \overline{X}_S = 44.40–51.10 = -6.7$$

How does this relate to the research hypothesis?

- $\overline{X}_N < \overline{X}_M$: This was NOT supported. Although the mean of the Minimal Intervention group was larger than the group with No Intervention, it was not significantly larger (because the critical value of Tukey's HSD of 7.18 was bigger than the absolute value of the mean difference between these two groups of -1.7)
- $\overline{X}_N < \overline{X}_S$: This was supported.
- $\overline{X}_M < \overline{X}_S$: This was NOT supported. The means were not different enough to reject the null hypothesis that the means are from different populations, so we conclude that the Minimal Intervention group had similar Mindset Quiz scores as the Super Intervention group.

This means that our research hypothesis was *partially supported*.

Write-Up

Okay, here's the big finish!

? Exercise 9.6.4

Write a conclusion to describe the results of the analysis. Don't forget to include the four components necessary in any report of results.

Answer

The researchers hypothesized that students in the No Intervention group will score lower on average on the Mindset Quiz than both students in the Minimal Intervention Group and the Super Intervention group. Similarly, students in the Minimal Intervention group will score lower on average than students in the Super Intervention group.

This research hypothesis was partially supported ($F(2,27) = 4.71$, $p < 0.05$). The Super Intervention group's average Mindset Quiz score ($M = 51.10$) was higher than the No Intervention group ($M = 42.70$), but the the Minimal Intervention group's ($M = 44.40$) average Mindset Quiz scores were not statistically significantly different than either the Super Intervention group or the No Intervention group.

Did Dr. MO's write-up include all of the components?

You did it! Take a break and reward yourself!

Next up, we'll look at what to do if you don't think that your distribution is normally distributed when you have three or more groups…

This page titled 9.6: Practice on Mindset Data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

9.7: On the Relationship Between ANOVA and the Student t Test

There’s one last thing before moving on to non-parametric alternatives to ANOVA. It’s something that a lot of students ask about, and many people find the answer surprising, but it’s worth knowing about: an ANOVA with two groups is identical to the Student t-test. No, really. It’s not just that they are similar, but they are actually equivalent in every meaningful way.

You are encouraged to try this yourself! Iif you do conduct an independent samples t-test and a BG ANOVA on the same data, you will get different calculated answers but statistical software will show you that that the actual p-value is identical. What's more, if you square the calculated t-score, you should get pretty close to the calculated F-value! Math is so weird.

Dr. MO has been tempted to turn to ANOVAs because it takes into account within-group variation, but it turns out, mathematically, it doesn't matter! When you only have two groups, use whichever analysis you prefer!

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)

This page titled 9.7: On the Relationship Between ANOVA and the Student t Test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

9.8: Non-Parametric Analysis Between Multiple Groups

When you have ranked data, or you think that the distribution is not normally distributed, then you use a non-parametric analysis. When you have three or more independent groups, the Kruskal-Wallis test is the one to use! The test statistic letter for the Kruskal-Wallis is H, like the test statistic letter for a Student t-test is t and ANOVAs is F.

More on When to Use the Kruskal-Wallis:

Some people have the attitude that unless you have a large sample size and can clearly demonstrate that your data are normal, you should routinely use Kruskal–Wallis; they think it is dangerous to use a Between Groups ANOVA, which assumes normality, when you don't know for sure that your data are normal. However, a Between Groups ANOVA is generally robust (not very sensitive not meeting its assumptions, like deviations from normality). Dr. McDonald has done simulations with a variety of non-normal distributions, including flat, highly peaked, highly skewed, and bimodal, and the proportion of false positives is always around 5% or a little lower, just as it should be. For this reason, he doesn't recommend the Kruskal-Wallis test as an alternative to a Between Groups ANOVAs. However, because many people use the Kruskal-Wallis, you should be familiar with it even if Dr. McDonald convince you that it's overused.

The Kruskal-Wallis test is a non-parametric test, which also means that it does not assume that the data come from a distribution that can be completely described by two parameters, mean and standard deviation (the way a normal distribution can). Like most non-parametric tests, you perform it on ranked data, so you convert the measurement observations to their ranks in the overall data set: the smallest value gets a rank of 1, the next smallest gets a rank of 2, and so on. You lose information when you substitute ranks for the original values, which can make this a somewhat less powerful test than a a Between Groups ANOVA; this is another reason to prefer a a Between Groups ANOVA.

The other assumption of a Between Groups ANOVA is that the variation *within* the groups is equal (homoscedasticity). While Kruskal-Wallis does not assume that the data are normal, it does assume that the different groups have the same distribution, and groups with different standard deviations have different distributions. If your data are heteroscedastic, Kruskal–Wallis is no better than a Between Groups ANOVA, and may be worse. Instead, you should use Welch's ANOVA for heteoscedastic data (which is not discussed in this textbook).

The only time I recommend using Kruskal-Wallis is when your original data set actually consists of one nominal variable and one *ranked* variable.

Hypotheses

Research Hypotheses

Like a Between Groups ANOVA, the Kruskal-Wallis will have three or more groups, so the research hypothesis will describe how all of these groups relate by predicting their mean *ranks*.

Null Hypothesis

The null hypothesis of the Kruskal–Wallis test is that the mean ranks of the groups are the same.

Critical Value

H is approximately chi-square distributed. We have not discussed chi-square, but we will! The critical values of chi-square table can be found on a future page or through the Common Critical Values page at the end of this textbook. The degrees of freedom is the number of groups minus 1 (k-1).

Compute the Test Statistic

Here are some data on Wright's F_{ST} (a measure of the amount of geographic variation in a genetic polymorphism) in two populations of the American oyster, *Crassostrea virginica*. McDonald et al. (1996) collected data on F_{ST} for six anonymous DNA polymorphisms (variation in random bits of DNA of no known function) and compared the F_{ST} values of the six DNA polymorphisms to F_{ST} values on 13 proteins from Buroker (1983). The biological question was whether protein polymorphisms

would have generally lower or higher F_{ST} values than anonymous DNA polymorphisms. McDonald et al. (1996) knew that the theoretical distribution of F_{ST} for two populations is highly skewed, so they analyzed the data with a Kruskal–Wallis test.

When working with a measurement variable, the Kruskal–Wallis test starts by substituting the rank in the overall data set for each measurement value. The smallest value gets a rank of 1, the second-smallest gets a rank of 2, etc. Tied observations get average ranks; in this data set, the two F_{ST} values of -0.005 are tied for second and third, so they get a rank of 2.5.

Table 9.8.1- Oyster Data and Ranks

gene	class	FST	Rank (DNA)	Rank (Protein)
CVJ5	DNA	-0.006	1	
CVB1	DNA	-0.005	2.5	
6Pgd	protein	-0.005		2.5
Pgi	protein	-0.002		4
CVL3	DNA	0.003	5	
Est-3	protein	0.004		6
Lap-2	protein	0.006		7
Pgm-1	protein	0.015		8
Aat-2	protein	0.016		9.5
Adk-1	protein	0.016		9.5
Sdh	protein	0.024		11
Acp-3	protein	0.041		12
Pgm-2	protein	0.044		13
Lap-1	protein	0.049		14
CVL1	DNA	0.053	15	
Mpi-2	protein	0.058		16
Ap-1	protein	0.066		17
CVJ6	DNA	0.095	18	
CVB2m	DNA	0.116	19	
Est-1	protein	0.163		20

To use the following formula, there is one new notation: R_{group}. This is the sum of the ranks for that group. Otherwise, you've seen all of this before!

$$H = \left[\left(\frac{12}{(N*(N=1)}\right)*\left(\sum\frac{R^2_{group}}{n_{group}}\right)\right] - (3*(N=1))$$

You calculate the sum of the ranks for each group (R_{group}), then the test statistic, H. For the example data, You add the 6 ranks for the first group ($\sum = 60.5$) and the 14 ranks for the second group ($\sum = 149.5$), then follow the rest of the formula to get to H = 0.04.

Make the Decision

The Kruskall-Wallis uses the Table of Critical Values of Chi-Square, which can be found on a page in the Chi-Square chapter, or you can find the link in the Common Tables of Critical Values page at the end of this textbook. To use this table, the Degrees of

Freedom for Kruskal-Wallis is k-1 ($df = k - 1$) in which k is the number of groups.

In our example above, we have two groups (protein or DNA), so the Degrees of Freedom would be 2 ($df = k - 1 = 3 - 1 = 2$). The critical value for an probability of 0.05 for $H(2)$ = 5.991, so we would retain the null hypothesis and say that there is not a difference in mean ranks for these two groups.

If you do find a significant difference in the ranks and you have more than two groups, you should run a Mann-Whitney U (discussed previously in the chapter on independent samples t-test analysis) as a pairwise comparison of each pair of ranks.

Assumptions

The Kruskal–Wallis test does NOT assume that the data are normally distributed; that is its big advantage. If you're using it to test whether the medians are different, it does assume that the observations in each group come from populations with the same shape of distribution, so if different groups have different shapes (one is skewed to the right and another is skewed to the left, for example, or they have different variances), the Kruskal–Wallis test may give inaccurate results (Fagerland and Sandvik 2009). If you're interested in any difference among the groups that would make the mean ranks be different, then the Kruskal–Wallis test doesn't make any assumptions.

Heteroscedasticity is one way in which different groups can have different shaped distributions. If the distributions are heteroscedastic, the Kruskal–Wallis test won't help you; instead, you should use Welch's *t*–test for two groups, or Welch's ANOVA for more than two groups; both of these are beyond the scope of this textbook but many statistical software packages will run these analyses.

References

Fagerland, M.W., and L. Sandvik. 2009. The Wilcoxon-Mann-Whitney test under scrutiny. Statistics in Medicine 28: 1487-1497.

McDonald, J.H., B.C. Verrelli and L.B. Geyer. 1996. Lack of geographic variation in anonymous nuclear polymorphisms in the American oyster, *Crassostrea virginica*. Molecular Biology and Evolution 13: 1114-1118.

Contributor

- John H. McDonald (University of Delaware)
- Dr. MO (Taft College)

This page titled 9.8: Non-Parametric Analysis Between Multiple Groups is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **4.8: Kruskal–Wallis Test** by John H. McDonald has no license indicated. Original source: http://www.biostathandbook.com.

CHAPTER OVERVIEW

10: Analysis of Variance- Repeated Measures

10.1: Introduction to Repeated Measures ANOVA

10.1.1: Things Worth Knowing About RM ANOVAs

10.2: ANOVA Summary Table

10.2.1: Repeated Measures ANOVA Sum of Squares Formulas

10.3: Practice with RM ANOVA Summary Table

10.3.1: Practice with Mindset

10.4: Non-Parametric RM ANOVA

This page titled 10: Analysis of Variance- Repeated Measures is shared under a not declared license and was authored, remixed, and/or curated by Michelle Oja.

10.1: Introduction to Repeated Measures ANOVA

This chapter will introduce a new kind of ANOVA, so we'll know about two kinds of ANOVAs:

- Between-Groups ANOVA: We collect the quantitative variable from each participant, who is only in **one** condition of the qualitative variable.

This is what we've been doing.

Example 10.1.1

Is this more like an independent t-test or a dependent t-test?

Solution

This is more like an independent t-test because the participants in the separate groups are unrelated to each other.

- Within-Groups ANOVA: We collect the quantitative variable from each participant, who is in all conditions of the qualitative variable.

Exercise 10.1.1

Is this more like an independent t-test or a dependent t-test?

Answer

This is more like a dependent t-test because the participants are in all ofthe conditions, or are somehow linked with participants in each of the different groups (levels of the IV).

RM ANOVAs are still ANOVAs, we're still looking at the ratio of between groups variability to within groups variability. With repeated measures, we just have more information. Our total variation is comprised of:

- Within group variation which we can't measure directly,
- Individual differences, which we *can* measure indirectly, and
- Between group differences, which we can measure directly.

Comparison of BG & WG ANOVA

- Data collection is different for the two
  - BG - each participant is only in one condition/group
  - RM- each participant will be in all conditions/groups
- SAME hypotheses:
  - Research hypothesis: Mean difference in a specific direction
  - Null hypothesis: No mean differences between the groups
- Formulas and computation is different for BG & WG
- Both use an ANOVA Summary Table, but RM has an extra row for the within-participant source of variation

Partitioning Variation

Time to introduce a new name for an idea you learned about last chapter, it's called partitioning the sums of squares. Sometimes an obscure new name can be helpful for your understanding of what is going on. ANOVAs are all about partitioning the sums of squares. We already did some partitioning in the last chapter. What do we mean by partitioning?

The act of partitioning, or splitting up, is the core idea of ANOVA. To use the house analogy. We want to split our total sums of squares (SS Total) up into littler pieces. Before we partitioned SS Total using this formula:

$$SS_{\text{TOTAL}} = SS_{\text{BG Effect}} + SS_{\text{WG Error}}$$

Remember, the $SS_{\text{BG Effect}}$ was the variance we could attribute to the means of the different groups, and $SS_{\text{wG Error}}$ was the leftover variance that we couldn't explain. $SS_{\text{BG Effect}}$ and $SS_{\text{WG Error}}$ are the partitions of SS_{TOTAL}, they are the littler pieces that combine to make the whole.

In the Between Groups ANOVA, we got to split SS_{TOTAL} into two parts. What is most interesting about the Repeated Measures ANOVA (sometimes known as Within-Groups ANOVA), is that we get to split SS_{TOTAL} into three parts, there's one more little piece. Can you guess what the new partition is? Here is the new idea for partitioning SS_{TOTAL} in a Repeated Measures ANOVA:

$$SS_{\text{TOTAL}} = SS_{\text{BG Effect}} + SS_{\text{Participants}} + SS_{\text{WG Error}}$$

We've added $SS_{\text{Participants}}$ as the new idea in the formula. What's the idea here? Well, because each participant was measured in each condition (more than once), we have a new set of means. These are the means for each participant, collapsed across the conditions. For example, if we had a study in which each participant was measured three time (Before, During, After), then Participant 1 has a mean (an average of their three scores); Participant 2 has a mean (an average of their three scores); and so on. We can now estimate the portion of the total variance that is explained by the means of these participants.

We just showed you a conceptual "formula" to split up SS_{TOTAL} into three parts. All of these things have different names, though. To clarify, we will confuse you just a little bit. Be prepared to be confused a little bit.

A Rose by Any Other Name

First, the Repeated Measures ANOVA has different names. Sometimes it's called a Within-Groups ANOVA, and you could even call it an ANOVA with Dependent Groups (although I've never actually seen anyone do that). The point is, these are all the same analysis. "Repeated Measures" means that each participant is measured in each condition, like in our Before, During, and After example above. This is pretty clear, and that's why I call this ANOVA the RM ANOVA. However, sometimes it's not the same person in each condition, but three people who are related, or dependent upon each other. For example, let's say that we did a study comparing families with three kids to see if birth order matters. Each family has three different kids, but the kids are definitely related and dependent on each other. We would need an analysis that takes into account that inter-relatedness. I mean, we could use a Between Groups ANOVA, but that information about how the three kids are treated similarly or affect each other would become noise in a BG ANOVA and mask differences based on birth order.

You might notice that this is a design issue (how data is collected), not an analysis issue. That's why it doesn't really matter what it's called, as long as we all agree on how to analyze this related data. If you're a psychology or sociology major, you'll probably take a class called Research Methods which will talk about these design issues in detail.

Second, different statisticians use different words to describe parts of the ANOVA. This can be really confusing. For example, we described the SS formula for a between subjects design like this:

$$SS_{\text{TOTAL}} = SS_{\text{BG Effect}} + SS_{\text{Participants}} + SS_{\text{WG Error}}$$

To have something called "between groups" in the formula might make you think that you're conducting a Between Groups ANOVA, when it's really just describing the partition comparing the different groups. We think this is very confusing for people. Here "within groups" has a special meaning. It does not refer to a within-groups or repeated measures design. $SS_{\text{WG Error}}$ refers to the leftover variation within each group mean. Specifically, it is the variation between each group mean and each score in the group. "AAGGH, you've just used the word between to describe within group variation!". Yes! We feel your pain. Remember, for each group mean, every score is probably off a little bit from the mean. So, the scores within each group have some variation. This is the within group variation, and it is why the leftover error that we can't explain is often called $SS_{\text{WG Error}}$. It is error caused by the variation within each group.

We're getting there, but perhaps a picture will help to clear things up. Note that Figure **10.1.1** uses the outdated name for the within-participant variation as "Subjects".

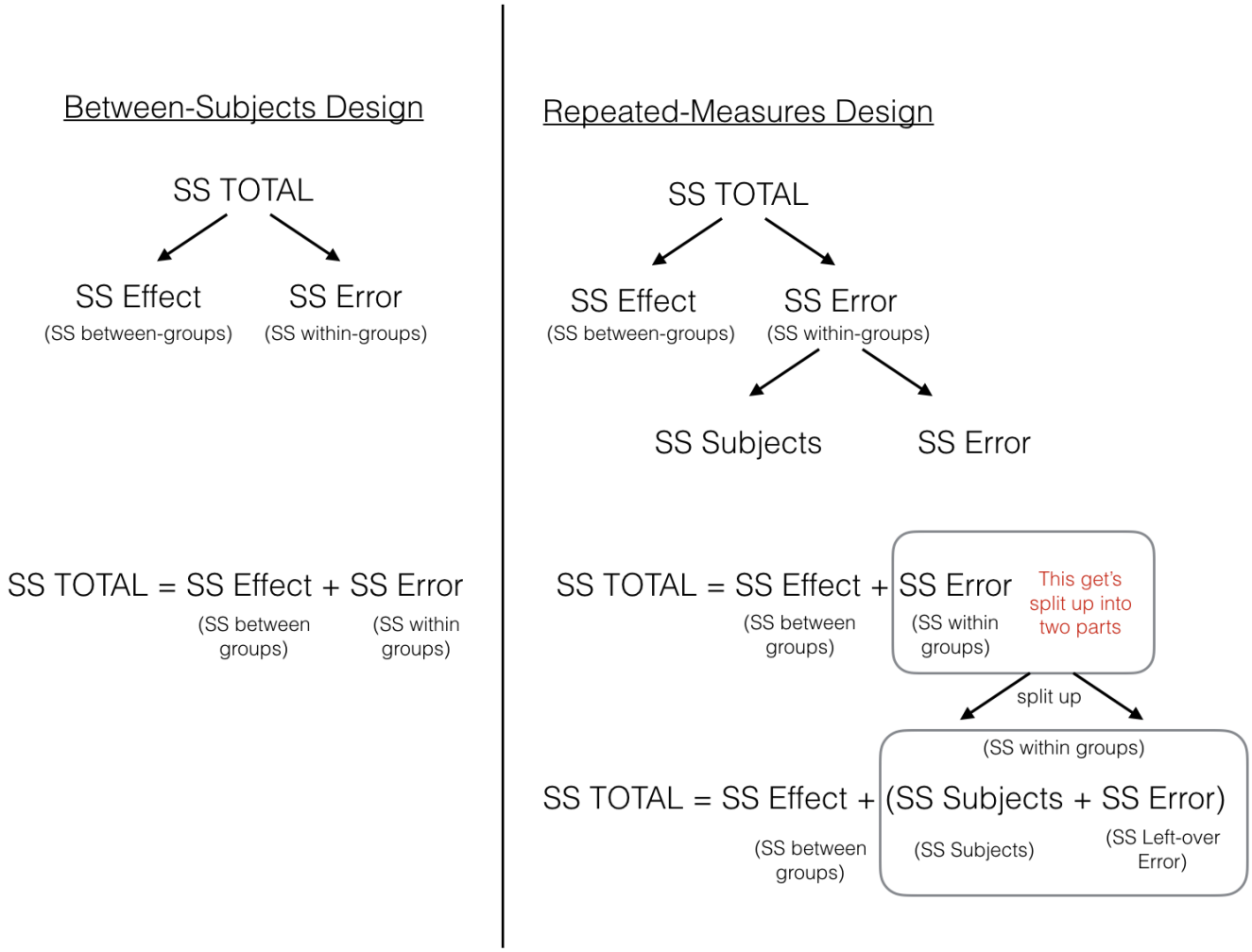


Figure **10.1.1**: Illustration showing how the total sums of squares are partitioned differently for a between versus repeated-measures design.(CC-BY-SA Matthew J. C. Crump via Answering Questions with Data)

The figure lines up the partitioning of the Sums of Squares for both Between Groups ANOVA and Repeated Measures ANOVA designs. In both designs, SS_{Total} is first split up into two pieces $SS_{\text{Effect (between-groups)}}$ and $SS_{\text{Error (within-groups)}}$. At this point, both ANOVAs are the same. In the repeated measures case we split the $SS_{\text{Error (within-groups)}}$ into two more littler parts, which we call $SS_{\text{Subjects (error variation about the subject mean)}}$ and $SS_{\text{Error (left-over variation we can't explain)}}$.

So, when we earlier wrote the formula to split up SS in the repeated-measures design, we were kind of careless in defining what we actually meant by SS_{Error}, this was a little too vague:

$$SS_{\text{TOTAL}} = SS_{\text{BG Effect}} + SS_{\text{Partipants}} + SS_{\text{WG Error}}$$

The critical feature of the Repeated Measures ANOVA, is that the $SS_{\text{WG Error}}$ that we will later use to compute the Mean Square in the denominator for the F-value, is smaller in a repeated measures design, compared to a between subjects design. This is because the $SS_{\text{Error (within-groups)}}$ is split into two parts, $SS_{\text{Subjects (error variation about the subject mean)}}$ and $SS_{\text{Error (left-over variation we can't explain)}}$.

To make this more clear, Dr. Crump made another Figure **10.1.2**:

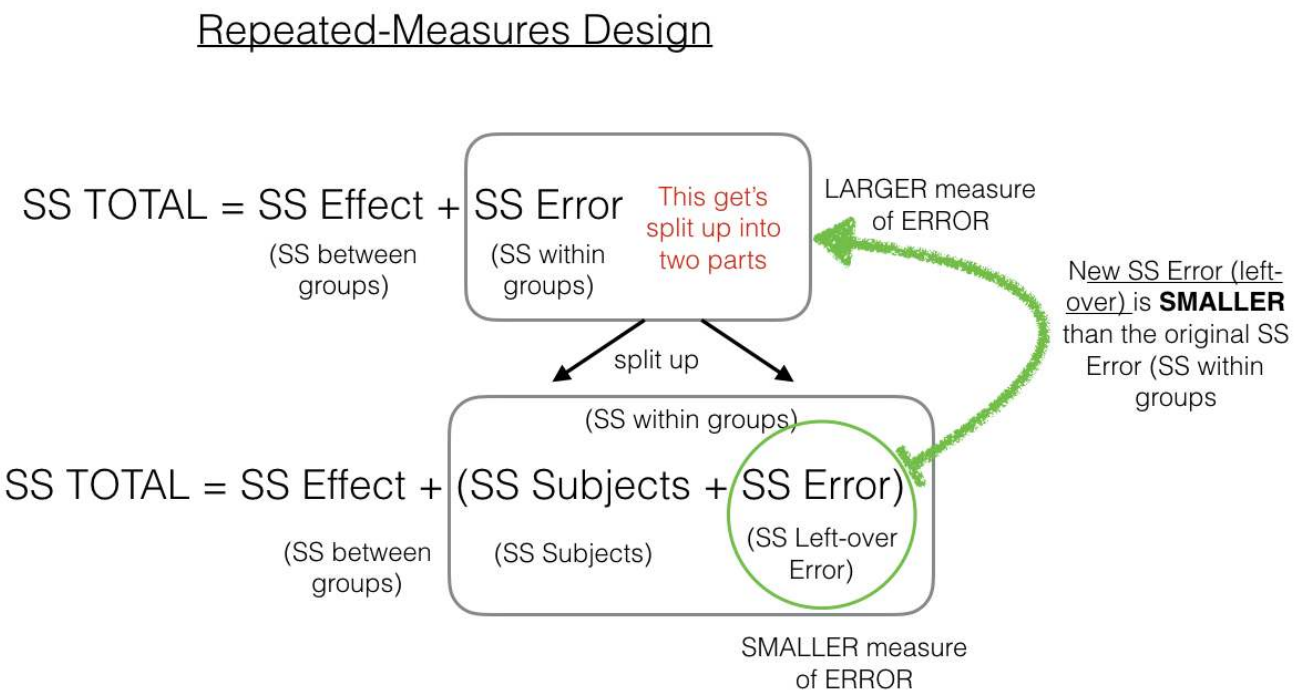


Figure **10.1.2**: Close-up showing that the Error term is split into two parts in the repeated measures design. (CC-BY-SA Matthew J. C. Crump via Answering Questions with Data)

As we point out, the $SS_{\text{Error (left-over)}}$ in the green circle will be a smaller number than the $SS_{\text{Error (within-group)}}$. That's because we are able to subtract out the SS_{Subjects} part of the $SS_{\text{Error (within-group)}}$. As we will see shortly, this can have the effect of producing larger F-values when using a repeated-measures design compared to a between-subjects design.

What does this new kind of variation look like in the ANOVA Summary Table? Let's look!

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 10.1: Introduction to Repeated Measures ANOVA is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

10.1.1: Things Worth Knowing About RM ANOVAs

Repeated Measures ANOVAs have some special properties that are worth knowing about. The main special property is that the error term used to for the F-value (the denominator of the calculated F ratio) will always be smaller than the error term (denominator) in a Between Groups ANOVA. It is smaller because we subtract out the error associated with the participant variability. This can have the consequence of generally making calculated F-values in Repeated Measures ANOVAS larger than calculated F-values in Between Groups ANOVAs because when the number in the bottom is smaller, it will make the resulting product a larger number.

Because big F values usually let us reject the idea that differences in our means are due to chance, the Repeated Measures ANOVA becomes a more sensitive test of the differences (its F-values are usually larger).

At the same time, there is a trade-off here. The Repeated Measures ANOVA uses different Degrees of Freedom, and these are typically smaller for the Within-Groups/Error in a Repeated Measures formula. The F-distributions for the Repeated Measures and Between Groups ANOVAs are actually different F-distributions, because they have different degrees of freedom.

If we generally get a bigger calculated score with Repeated Measures, but the bar is raised a little because the Degrees of Freedom are smaller, which type of research design should you try to use? Well, the answer is actually not statistical.

First, some IVs cannot be repeated. You can't teach someone to swim, and then put them in a condition in which they are expected to not know how to swim. In those cases, you must use a between groups design.

Second, with each additional condition (group), we have to get at least 30 more participants. This can get time-consuming and expensive quickly. Since repeated measures designs are more cost-effective, and they take into account each participants' individual tendencies, they are preferred.

The best of both worlds is to get the same amount of participants that we'd expect in a between groups design, but measure them repeatedly. This increases our sample size without actually increasing our number of participants!

Contributors and Attributions

Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

Dr. MO (Taft College)

This page titled 10.1.1: Things Worth Knowing About RM ANOVAs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

10.2: ANOVA Summary Table

RM ANOVAs are still ANOVAs, we're still looking at the *ratio* of between groups variability to within groups variability. With repeated measures, we have even more information: How similar that person is to themselves!

RM ANOVA Summary Table

Now that you are familiar with the concept of an ANOVA table (remember the ANOVA Summary Table from last chapter where we reported all of the parts to calculate the F-value?), we can take a look at the things we need to find out to make the ANOVA table. Table **10.2.1** presents an empty Repeated Measures ANOVA Summary Table.

Table **10.2.1**- RM ANOVA Summary Table

Source	SS	df	MS	F
Between Groups				
Participants				
Within Groups (Error)				
Total				

? Exercise 10.2.1

What is the biggest difference between the Repeated Measure ANOVA Summary Table and the Between Groups ANOVA Summary Table (the one we talked about in the prior chapter)?

Answer

The biggest difference is that there's a whole new row! We now also are taking intou account how similar each person is to themselves (each person's average response).

ANOVA Summary Table Formulas

Since the Within-Groups variability cannot only be measured indirectly, there's a wrinkle to completing the ANOVA Summary Table. To figure out the Within Groups error you need to find it with subtraction: $SS_{\text{WGerror}} = SS_{\text{tot}} - SS_{\text{b}} - SS_{\text{Ps}}$

Table **10.2.2** shows this in the Sum of Squares column, but the rest of the formulas are presented later. The formulas for Degrees of Freedom, Mean Square, and the final calculated F-score are included. There are more cells that should be blank in this version of the ANOVA Summary Table; these are labled "N/A" in the table.

Table **10.2.2**- RM ANOVA Summary Table with Formulas for df, MS, and F

Source	SS	df	MS	F
Between Groups	Formula elsewhere	k-1	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_W}$
Participants	Formula elsewhere	P-1	N/A	N/A
Within Groups (Error)	$SS_{WG} = SS_T - SS_{BG} - SS_P$	$(k-1) \times (P-1)$	$\frac{SS_W}{df_W}$	N/A
Total	Formula elsewhere	N-1	N/A	N/A

Degrees of Freedom

- **N** = the number of *scores* (not the number of participants)
- **Participants df**: The number of participants (subjects) minus 1 (S-1).

So, like in a dependent t-test, df is the number of participants minus 1 (not the number of numbers we have). But, confusingly, **N** *is* the number of scores!

Although we won't be able to do a computation check for the Sums of Squares, we can make sure that Degrees of Freedom are correct: $df_{Total} = df_{BG} + df_P + df_{WG}$

Practicing the ANOVA Summary Table

Let's practice filling in the table if:

- k (number of groups): 3
- P (number of people): 15
- N (number of scores): 45

If you are not provided the number of scores (N), you can figure it out by multiplying the number of groups (k) with the number of people (P) since all people are in all groups.

The Sum of Squares for Between Groups, Participants, and the Total are also provided.

Table **10.2.3**- Practice with RM ANOVA Summary Table

Source	SS	df	MS	F
Between Groups	4.42	$k-1=3-1=2$	Between Groups $\frac{SS}{df}=\frac{4.42}{2}$ = 2.21	$F_{calc}=\frac{MS_{BG}}{MS_{WG}}=\frac{2.21}{0.38}$ = 5.83
Participants	3.59	$P-1=15-1=14$	leave blank	leave blank
Within Groups (Error)	$SS_{WG}=SS_T-SS_{BG}-SS_P=18.63-4.42-3.59=10.62$	$(k-1)\times(S-1)=(3-1)\times(15-1)=2\times14=28$	Within Groups $\frac{SS}{df}=\frac{10.62}{28}$ = 0.38	leave blank
Total	18.63	$N-1=45-1=44$	leave blank	leave blank

Make sure to do the Computation Check to make sure that you didn't make a mistake: Total=BG+P+WG

Degrees of Freedom for $BG+P+WG=2+14+28$ should equal 44... (It does! We did it correctly!)

What you might have noticed is that we calculated the Sum of Squares for the Participants, but didn't seem to do anything with it. But we actually did! We calculated the average variation of the Participants so that we can account for it in the total variation. That's what we did when we subtracted Sum of Squares for the Particiapnts and the Between Groups from the Total, we were saying that we know that there's a certain amount of variation within each participant, and we are taking it out of our total variation so that our within groups variation (Error) is smaller so that we can see the variation between the groups easier.

Next Steps?

What we we do next? We can compare this calculated F-value to the critical F-value foudn in the page from the chapter on Between Groups ANOVAs (or look for the link in the Common Critical Value Tables at the back of this book) with the Degrees of Freedom of the numerator (Between Groups MS) and denominator (Within Groups MS) to find the critical F at p = 0.05of 3.34.

Note

Still:

(Critical < Calculated) = Reject null = At least one mean is different from at least one other mean. = $p < .05$

(Critical > Calculated) = Retain null = All of the means are similar. = $p > .05$

So we would reject the null hypothesis, say that at least one mean is different from at least one other mean, and use post-hoc analyses to find out which means differ.

But what we're going to do next is look at the new formula for Sum of Squares for Participants.

Contributors and Attributions

This page was extensively adapted by Michelle Oja (Taft College) from work by Matthew J. C. Crump (Brooklyn College of CUNY)

This page titled 10.2: ANOVA Summary Table is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **8.3: Calculating the RM ANOVA** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.
- **11.3: ANOVA Table** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

10.2.1: Repeated Measures ANOVA Sum of Squares Formulas

In a Repeated Measures ANOVA, you don't calculate the Within-Groups (Error) Sum of Squares from a formula. Instead, you calculate the Within-Groups (Error) Sum of Squares by calculating the other Sums of Squares, then subtracting the Between Groups SS and Participants SS from the Total Ss. Unfortunately, that means that there's no computational check. ☹

But in better news, that also means that you already know two of the three Sum of Squares formulas! (Well, three of the four formulas if you can the subtraction one…)

Sum of Squares Refresher

Between Groups Sum of Squares

$$SS_B = \sum_{EachGroup} \left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 * (n_{group}) \right]$$

1. Subtract
2. Square
3. Multiply
4. Sum

Within Groups Sum of Squares

A refresher from a page ago!

$$SS_{WG} = SS_T - SS_{BG} - S_P$$

1. Calculate
2. Subtract

Total Sum of Squares

$$SS_T = \sum \left[\left(X - \overline{X}_T \right)^2 \right]$$

1. Subtract
2. Square
3. Sum

New Sum of Squares Formula: Participants

The newest formula has similarities to the prior sums of squares:

$$SS_P = \left[\sum \frac{((\sum X_P)^2)}{k} \right] - \frac{((\sum X)^2)}{N}$$

In which $\sum X_P$ means that you sum all of the scores for that particular participant. For example, if you had the average album sales for three participants (Ariana, Beyonce, and Carli) at three different times periods (pre-pandemic, pandemic, post-pandemic), you would add up all of Ariana's scores for all three time periods, square them, then divide by the number of time periods (3). Then, do the same thing for the other participants. You can see how this is a mean for each individual person. The big sigma at the beginning is tell you to add all of these individual participants' averages together.

What is being subtracted from the sum of each participants' average score is the basic sum of squares average for all of the scores. You add up all of the scores, square them, then divide by the number of scores. Easy-peasy!

1. Individual participants:
    1. Add
    2. Square
    3. Divide
2. Add all of the participants' means together.

3. Subtract
1. Add the sum of squares for every score:
2. Square
3. Divide

We will practice this later on mindset data. For now, let's just leave this here for now, and practice with the RM ANOVA Summary Table without these pesky Sum of Squares formulas!

This page titled 10.2.1: Repeated Measures ANOVA Sum of Squares Formulas is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

10.3: Practice with RM ANOVA Summary Table

RLet's use a real scenario to practice with the Repeated Measures ANOVA Summary Table.

Scenario

The data are taken from a recent study conducted by Behmer and Crump (a co-author of this chapter), at Brooklyn College (Behmer & Crump, 2017).

Behmer and Crump (2017) were interested in how people perform sequences of actions. One question is whether people learn individual parts of actions, or the whole larger pattern of a sequence of actions. We looked at these issues in a computer keyboard typing task. One of our questions was whether we would replicate some well known findings about how people type words and letters.

From prior work we knew that people type words way faster than than random letters, but if you made the random letters a little bit more English-like, then people who can read English type those letter strings a little bit faster, but not as slow as random string.

In the study, 38 participants sat in front of a computer and typed five-letter strings one at a time. Sometimes the five letters made a word (Normal condition: TRUCK), sometimes they were completely random (Random condition: JWYFG), and sometimes they followed patterns like you find in English but were not actual words (Bigram condition: QUEND). What makes this repeated measures is that each participant received each condition; some trials was a word, some trials were a random string of letters, and some trials were a string a letters that looked like a word in English but was not a word. The order for each trial were randomly assigned. We measured every single keystroke that participants made, and we'll look at the reaction times (how long it took for participants to start typing the first letter in the string, in milliseconds).

Example 10.3.1

Answer the following questions to understand the variables and groups that we are working with.

1. Who is the sample?
2. Who do might be the population?
3. What is the IV (groups being compared)?
4. What is the DV (quantitative variable being measured)?

Solution

1. The sample is 38 participants.
2. Maybe anyone who types on a keyboard? English-speaker typists? There's not much info in the scenario to determine a specific population.
3. The IV is something like "word status" with the three levels being Normal (English word), Random (letter string), and Bigram (English-like letter string).
4. Reaction time (how long it took for participants to start typig the first letter) in milliseconds.

Step 1: State the Hypotheses

Based on the means from Table **10.3.1**, we can see the means look different. What could be a directional research hypothesis?

Table **10.3.1**- Descriptive Statistics for Reaction Time by Word Conditions

	N:	Mean:	SD:
Normal (English word)	38	779.00	20.40
Bigram (English-like non-word)	38	869.00	24.60
Random (non-word)	38	1037.00	29.30

? Exercise 10.3.1

Determine the research hypothesis in words and symbols. You can fill in the following underlined spot with the symbols for greater than (>), less than (<), or equal signs. Just remember, at least one pair of means must be predicted to be different from each other.

Symbols:

- $\overline{X}_N$ _____ $\overline{X}_B$
- $\overline{X}_N$ _____ $\overline{X}_R$
- $\overline{X}_B$ _____ $\overline{X}_R$

Answer

Based on the means, I might predict that the Normal condition will react fastest and will be significantly faster than the Bigram condition and the Random condition. I also might hypothesis that the Bigram condition would have a significantly shorter reaction time as the Random condition, as well.

Symbols:

- $\overline{X}_N < \overline{X}_B$
- $\overline{X}_N < \overline{X}_R$
- $\overline{X}_B < \overline{X}_R$

Notice that we are predicting that the Normal words will have a smaller reaction time, meaning that they will respond faster and the time to respond will be shorter.

What about the null hypothesis? What might that look like?

? Exercise 10.3.2

State the null hypothesis in words and symbols. .

Answer

The reaction time will be similar for the Normal condition, the Bigram condition, and the Random condition.

$\overline{X}_N = \overline{X}_B = \overline{X}_R$

Step 2: Find the Critical Values

Using the sample size information included in Table 1, you can now find the critical values from the Critical Values of F Table found on this page in the chapter that first discussed ANOVAs, or find a list of critical value tables at the end of this textbook on the Common Critical Value Tables page.

As shown on the bottom of the critical values page, the two Degrees of Freedom that you'll use is still from the numerator (Between Groups) and the denominator (Within Group or Error), but the denominator's df is calculated slightly differently.

✓ Example 10.3.2

What is the critical value for this scenario?

Solution

The df for the numerator is still k-1; 3-1 = 2.

The df for the denominator is (k-1)*(P-1), which means that we need to figure out P-1 first. Since P stands for the number of participants, that would be 38-1 = 37.

$$(k-1)*(P-1) = (3-1)*(38-1) = 2*37 = 74$$

The critical value of F for 2 and 74 in the 0.05 row is 3.15.

Step 3: Compute the Test Statistic

Using the Sum of Squares provide in the following ANOVA Summary Table (Table **10.3.2**) and the information from the scenario about the sample size and number of conditions, fill in the ANOVA Summary Table to determine the calculated F-value.

Table **10.3.2**- RM ANOVA Summary Table with Some Sums of Squares

Source	SS	df	MS	F
Between	1,424,914.00			
Participants	2,452,611.90			
Error				
Total	4,101,175.30			

Example 10.3.3

Complete the ANOVA Summary Table in Table **10.3.2** to determine the calculated F-score.

Solution

Table **10.3.3**- RM ANOVA Summary Table with Formulas

Source	SS	df	MS	F
Between	1,424,914.00	$k-1=3-1=2$	$\frac{SS_B}{df_B}=\frac{1424914}{2}=712457$	$\frac{MS_B}{MS_E}=\frac{712457}{3022.29}=235.73$
Participants	2,452,611.90	$P-1=38-1=37$	leave blank	leave blank
Error	$SS_{WG}=SS_{Total}-SS-BG-SS-Ps$ $=4101175.30-1424914-2452611.90$ $=223,649.40$	$(k-1)\times(P-1)=2$ 74	$\frac{SS_E}{df_E}=\frac{223649.40}{74}$ $=3022.29$	leave blank
Total	4,101,175.30	$N-1=113$ $(N=k\times P)$	leave blank	leave blank

So the ANOVA Summary Table should end up looking like Table **10.3.4**:

Table **10.3.4**- Completed RM ANOVA Summary Table

Source	SS	df	MS	F
Between	1,424,914.00	2	712,457.00	235.73
Participants	2,452,611.90	37		
Error	223,649.40	74	3,022.29	
Total	4,101,175.30	113		

Step 4: Make the Decision

We have the critical value (3.15) and the calculated value (235.73), so we can now make the decision just like we've been doing.

Table **10.3.5**- Rejecting or Retaining the Null Hypothesis

REJECT THE NULL HYPOTHESIS	RETAIN THE NULL HYPOTHESIS
Small p-values (p<.05)	Large p-values (p>.05)
A small p-value means a small probability that all of the means are similar. *Suggesting that at least one of the means is different from at least one other mean…*	A large p-value means a large probability that all of the means are similar.
We conclude that: At least one mean is different from one other mean. At least one group is not from the same population as the other groups.	We conclude that: The means for all of the groups are similar. All of the groups are from the same population.
The calculated F is further from zero (more extreme) than the critical F. In other words, the calculated F is bigger than the critical F. (Draw the standard normal curve and mark the calculated F and the critical F to help visualize this.)	The calculated F is closer to zero (less extreme) than the critical F. In other words, the calculated F is smaller than the critical F. (Draw the standard normal curve and mark the calculated F and the critical F to help visualize this.)
Reject the null hypothesis (which says that all of the means are similar).	Retain (or fail to reject) the null hypothesis (which says that the all of the means are similar).
Support the Research Hypothesis? MAYBE. Look at the actual means: • Support the Research Hypothesis if the means are in the directions that were hypothesized. ◦ The mean of the group that you said would be bigger, really is bigger; ◦ The mean of the group that you said would be smaller really is smaller; ◦ The means of the groups that you said would be similar are actually similar. • Partial support of the Research Hypothesis if some of the means are in the directions that were hypothesized, but some aren't. • Do not support the Research Hypothesis if none of the means are in the direction that were hypothesized.	Do not support the Research Hypothesis (because all of the means are similar).
Statistical sentence: F(df) = F-calc, p<.05 (fill in the df and the calculated F)	Statistical sentence: F(df) = F-calc, p>.05 (fill in the df and the calculated F)

Here's another way to show the info in Table **10.3.5**:

(Critical < Calculated) = Reject null = At least one mean is different from at least one other mean. = $p < .05$

(Critical > Calculated) = Retain null = All of the means are similar. = $p > .05$

? Exercise 10.3.3

Should we retain or reject the null hypothesis? Does this mean that we're saying that all of the means are similar, or that at least one mean different?

Answer

Because the calculated F-score of 235.73 is so much bigger than the critical value of 3.15, we reject the null hypothesis and say that at least one mean is different from at least one other mean.

Before we can write-up the results for this analysis, we need to determine if the mean differences are in the hypothesized direction. Behmer and Crump (2017) provided the following t-test results that tested whether pairs of means are different:

- Normal versus Bigram: t(37) = -10.61, p < 0.001
- Normal versus Random: t(37) = -15.78, p < 0.001
- Bigram versus Random: t(37) = 13.49, p < 0.001

? Exercise 10.3.4

What is one problem with using t-tests to check for multiple sets of mean differences in post-hoc analyses?

Answer

Alpha inflation; each t-test has a 5% of committing a Type I Error (rejecting the null hypothesis when there really is no difference between the sample's means in the population).

? Exercise 10.3.5

What did we learn about in the Between Groups ANOVA chapter to use instead to check for multiple sets of mean differences in post-hoc analyses?

Answer

A variety of post-hoc analyses that reduced the chance of a Type I Error in each pairwise comparison so that the total chance of a Type I Error was still 5%.

Write-up

Okay, we now have all we need to complete a conclusion that reports the results while including all of the required components:

- The statistical test is preceded by the descriptive statistics (means).
- The description tells you what the research hypothesis being tested is.
- A "statistical sentence" showing the results is included.
- The results are interpreted in relation to the research hypothesis.

? Exercise 10.3.6

What could the conclusion look like for this scenario?

Answer

The research hypothesis was that the Normal condition would have the fastest reaction time ($M = 779ms$) compared to the Bigram ($M = 869ms$)and the Random ($M = 1,037ms$) conditions, and the Bigram condition would also have a faster reaction time than the Random condition. This research hypothesis was fully supported (F(2,37)=235.73, p<0.05).

That's it! Let's try one more example; this time, we'll calculate the Sum of Squares and the pairwise comparison.

This page titled 10.3: Practice with RM ANOVA Summary Table is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **8.5: Real Data** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

10.3.1: Practice with Mindset

Okay, here's your time to practice the whole process of a Repeated Measures ANOVA, including each Sum of Squares formula!

Scenario

In another study by student researchers at a community college, mindset was measured with a Mindset Quiz at the beginning of the semester in a remedial English class, at the end the semester in the remedial English class, then at the beginning of the next semester in the next higher English class, and at the end of that semester in the next higher English class. Thus, the **same students** were measured four times to see if their mindset improved through journaling activities related to growth mindset. Mindset Quiz scores can range from 20 to 60, with higher scores showing that the student has more Growth Mindset (and less Fixed Mindset). At the end of the second semester, 12 students had completed the Mindset Quiz pre-test and post-test during each class (at the beginning and end of each semester), and had attended most classes throughout the semester to experience the journaling activities related to mindset (N = 48).

After reading the scenario, can you describe scenario?

Example 10.3.1.1

Answer the following questions to understand the variables and groups that we are working with.

1. Who is the sample?
2. Who do might be the population?
3. What is the IV (groups being compared)?
4. What is the DV (quantitative variable being measured)?

Solution

1. The sample is 48 community college students in English classes, who started in remedial English.
2. The population could be students taking remedial English at community colleges?
3. IV: Time- Beginning of Remedial English, End of Remedial English, Beginning of Next English, End of Next English
4. DV: Mindset Quiz

Step 1: State the Hypotheses

Using the descriptive statistics in Table **10.3.1.1**, what could be the research hypothesis?

Table **10.3.1.1**- Descriptive Statistics for each IV level and the Total

	N:	Mean:	SD:
Beginning of Remedial English (R1)	12	39.17	6.46
End of Remedial English (R2)	12	43.33	8.48
Beginning of Next English Class (N1)	12	42.25	5.24
End of Next English Class (N2)	12	45.42	6.35
Total	48	42.54	6.90

The notation in the parentheses is what we'll use as the subscript labels in the formulas.

Note

You are encourages to use the raw data found in Table **10.3.1.2** in Step 3 to practice calculating the means, standard deviations, and medians yourself!

Remember, the research hypothesis should predict how all levels of the IV related to all other levels of the IV. Because this scenario has four different IV levels, there's actually six combinations of pairs of means!

Example 10.3.1.2

What is a research hypothesis for this scenario in words and symbols?

Solution

Research Hypothesis: Students average Mindset Quiz scores will be higher in the beginning of the semesters than at the end, and higher in the next class than in the first (remedial) class.

Your research hypothesis might be slightly different than this one. Just make sure that at least one mean is predicted to be different than one other mean.

Symbols:

- $\overline{X}_{R1} < \overline{X}_{R2}$
- $\overline{X}_{R1} < \overline{X}_{N1}$
- $\overline{X}_{R1} < \overline{X}_{N2}$
- $\overline{X}_{R2} < \overline{X}_{N1}$
- $\overline{X}_{R2} < \overline{X}_{N2}$
- $\overline{X}_{N1} < \overline{X}_{N2}$

Despite having so many combinations, the null hypothesis for this scenario is still relatively simple, just like all null hypotheses.

Exercise 10.3.1.1

State the null hypothesis in words and symbols.

Answer

Null Hypothesis: Students will have a similar average scores on the Mindset Quiz in all four conditions: Beginning of Remedial English, End of Remedial English, Beginning of Next English Class, and End of Next English Class

Symbols: $\overline{X}_{R1} = \overline{X}_{R2} = \overline{X}_{N1} = \overline{X}_{N2}$

Step 2: Find the Critical Values

This step might be easier after you've completed the ANOVA Summary Table because you will have the Degrees of Freedom for both groups, but we'll keep following the steps as we've learned them.

Exercise 10.3.1.2

Using the Critical Values of F Table or finding the link in the Common Critical Values page at the end of the textbook, what is the critical value at $\alpha = 0.05$?

Answer

Critical F(3,33) = 2.92

The first Degree of Freedom (3), for the numerator (df_B) is found through: k – 1, with k being the number of groups.

The second Degree of Freedom (33), for the denominator (df_W) is found through $(k-1) \times (Ps-1)$, with Ps being the number of participants.

$$(k-1) \times (Ps-1) = (4-1) \times (12-1) = 3 * 11 = 33$$

Take a breath, because now we will start with the messy calculations!

Step 3: Compute the Test Statistic

If you will never have to calculate the Sums of Squares by hand, skip this part and just fill in the ANOVA Summary Table (Table **10.3.1.5**) at the end of this section. If you are practicing the Sums of Squares, each Sum of Square will have its own Example. Heads up, to do all of these can take about an hour!

Table **10.3.1.2**- Raw Mindset Quiz Scores for Four Conditions

Participant	**Pretest in Remedial English**	**Posttest in Remedial English**	**Pretest in Next English Level**	**Posttest in Next English Class**	***Total***
A	30	48	32	42	***152***
B	31	39	35	32	***137***
C	33	35	39	38	***145***
D	35	32	40	50	***157***
E	36	37	42	48	***163***
F	38	33	42	51	***164***
G	39	45	43	38	***165***
H	42	52	45	48	***187***
I	43	39	45	48	***175***
J	47	51	46	51	***195***
K	47	54	47	51	***199***
L	49	55	51	48	***203***
Sum:	***470***	***520***	***507***	***545***	***2042***
N:	***12***	***12***	***12***	***12***	***48***

Let's go!

Between Groups SS

This is the same formula that we learned in the Between Groups ANOVA.

Example 10.3.1.3

Calculate the Between Groups Sums of Squares.

Solution

$$SS_B = \sum_{EachGroup} \left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 \times (n_{group}) \right]$$

The $\sum_{EachGroup}$ means that you do everything following that for each intervention level, then add them all together. Let's start with what's inside the brackets for the Mindset Quiz at the beginning of the remedial English class (R1).

$$R1 = \left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 \times (n_{group}) \right]$$

$\overline{X}_{group}$ is asking for the mean of the group that we're looking at, and $\overline{X}_T$ is asking for the total mean, the mean for all 48 scores. Both of these means were provided in Table **10.3.1.1**. The number of scores in the group that we're looking at right now is what n_{group} is asking.

So let's plug those all in!

$$R1 = \left[(39.17–42.54)^2 \times (12) \right]$$

$$R1 = \left[(-3.37)^2 \times (12) \right]$$

The mean of the group minus the mean of the sample should be negative in this scenario, but that negative sign goes away when we square it:

$$R1 = [(11.36) \times (12)]$$

$$R1 = [136.28]$$

And let's do that process three more times, once for the End of the Remedial English class's semester, once for the Beginning of the Next English Class, and once for the End of the Next English class. Then we'll add those four numbers together to get the SS_B.

$$R2 = \left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 \times (n_{group}) \right]$$

$$R2 = \left[(43.33–42.54)^2 \times (12) \right]$$

$$R2 = \left[(0.79)^2 \times (12) \right]$$

$$R2 = [(0.62) \times (12)]$$

$$R2 = [7.44]$$

If you are doing all of your calculations in Excel or somehow saving all of the decimal points for each answer, this calculation will be a little higher (7.49). Don't worry, the rounding differences will mostly wash out by the end of the formula. We will be using the answers provided if you type the two numbers after the decimal point into a calculator.

Why don't you try to do it on your own for the Next English class conditions?

$$\left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 \times (n_{group}) \right]$$

$$N1 = [0.96]$$

$$N2 = [99.48]$$

Next step, add them all together! It's easy to forget this step, but the Sum of Squares ends up to be one number, so when you get lost or forget the next step, look back at the full formula:

$$\sum_{EachGroup} \left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 \times (n_{group}) \right]$$

$$\sum_{EachGroup} = [136.32 + 7.44 + 0.96 + 99.48]$$

$$\sum_{EachGroup} = 244.20$$

Again, if you had saved more than two decimals, you would end up with $SS_B = 244.31$. And if you used that number in the ANOVA Summary Table, the calculated F-value would be nearly identical to the same calculations but with only the two numbers after the decimal.

You did it! Only three more Sums of Squares to go!

Participant SS

This one is new!

Example 10.3.1.4

Calculate the Participant Sum of Squares.

Solution

$$SS_{Ps} = \left[\sum\left(\frac{(\sum X_{Ps})^2}{k}\right)\right] - \frac{((\sum X)^2)}{N}$$

This one is easiest to do in a table, as shown in Table **10.3.1.3**.

Table **10.3.1.3**- Raw Mindset Quiz Scores and Calculations for Participant Sum of Squares

Participant	**Pretest in Remedial English**	**Posttest in Remedial English**	**Pretest in Next English Level**	**Posttest in Next English Class**	**Sum of Ps**	**Squared**	**Divide by k**
A	30	48	32	42	152	23,104.00	5776.00
B	31	39	35	32	137	18,769.00	4692.25
C	33	35	39	38	145	21,025.00	5256.25
D	35	32	40	50	157	24,649.00	6162.25
E	36	37	42	48	163	26,569.00	6642.25
F	38	33	42	51	164	26,896.00	6724.00
G	39	45	43	38	165	27,225.00	6806.25
H	42	52	45	48	187	34,969.00	8742.25
I	43	39	45	48	175	30,625.00	7656.25
J	47	51	46	51	195	38,025.00	9506.25
K	47	54	47	51	199	39,601.00	9900.25
L	49	55	51	48	203	41,209.00	10302.25
Sum:	**$\sigma = 470$**	**$\sigma = 520$**	**$\sigma = 507$**	**$\sigma = 545$**	**N/A**	**N/A**	**$\sigma = 88,166.50$**
N:	**n=12**	**n=12**	**n=12**	**n=12**	**N/A**	**N/A**	**48 scores**

What's happening in Table **10.3.1.3**? The first column is labeling each participant. When you see a column like this, then you know that the scores are related, meaning that you should run a dependent t-test or a Repeated Measures ANOVA. The two rows on the bottom ("Sum:" and "N:") show the sum of each column, and the number of scores in that column. This makes is really easy to find the mean! The next four columns are the scores for the four conditions of the IV.

Then, we get some columns to help us calculate the Participant Sum of Squares. The Sum of Ps column just adds the four scores for each individual participant. For example,

$$\text{Sum of Participant A} = 30 + 48 + 32 = 42 = 152$$

The next column, Squared, is the individual participant's summed scores squared $\text{Participant A} = 152^2 = 23104.00$. That number is then divided by k, the number of groups; we have four groups in this scenario. So, the full process for each individual participant is:

$$\text{Participant A} = \sum X_{ParticipantA} = 152 = 152^2 = 23104.00 = \frac{23,104}{4} = 5776.00$$

Once that process is completed for each participant (so, 12 times in this scenario), we add all of the numbers in the last column (the column that divides by k). That provides the bracketed information in the formula:

$$SS_{Ps} = \left[\sum\left(\frac{(\sum X_{Ps})^2}{k}\right)\right] - \frac{((\sum X)^2)}{N}$$

The next set of calculations ($\frac{(\sum X)^2}{N}$) works with all of the scores. The sum of all of the scores was provided in Table **10.3.1.2** ($\sum X = 2042$). That score is squared, then divided by the total number of scores (not the total number of people); this is 48 for this scenario because we have 12 people in four conditions ($N = Ps \times k = 12 \times 4 = 48$).

$$\frac{(2042)^2}{48}$$

$$\frac{(4,169,764.00)}{48} = 86,870.08$$

The final step for the Participant Sum of Square is to subtract this squared average of the total (the stuff that we just calculated) *from* the individual participants' sum of squared averages (the stuff we calculated on the table) .

$$SS_{Ps} = \left[\sum\left(\frac{(\sum X_{Ps})^2}{k}\right)\right] - \frac{((\sum X)^2)}{N} = 88,166.50 - 86,870.08 = 1,296.42$$

Note that Sums of Squares should always be positive, so if your answer is not positive, then you did something wrong...

We still can't do the Within Groups (Error) Sum of Squares because we need the Total Sum of Squares to get it.

Total SS

This is also the same formula that we used in the Between Groups ANOVA.

Example 10.3.1.5

Calculate the Total Sum of Squares.

Solution

$$SS_T = \sum\left[\left(X - \overline{X}_T\right)^2\right]$$

This formula is also saying to subtract a mean from each score, but this time we should be subtracting the Total mean ($\overline{X}_T = 42.54$, found in Table **10.3.1.1**). This is again easiest to compute in a table.

Table **10.3.1.4** shows the Total mean subtracted from each score in the column to the right of the raw scores, then that is squared in the column to the next right. The squared values are then summed for *all* of the scores.

Table **10.3.1.4**- Total Mean Subtracted from Each Score, Squared, then Summed

IV Levels	Mindset Quiz Scores	minus Total mean	squared
R1	30	-12.54	157.25
R1	31	-11.54	133.17
R1	33	-9.54	91.01
R1	35	-7.54	56.85
R1	36	-6.54	42.77
R1	38	-4.54	20.61
R1	39	-3.54	12.53
R1	42	-0.54	0.29
R1	43	0.46	0.21
R1	47	4.46	19.89
R1	47	4.46	19.89
R1	49	6.46	41.73
R2	48	5.46	29.81
R2	39	-3.54	12.53
R2	35	-7.54	56.85
R2	32	-10.54	111.09
R2	37	-5.54	30.69
R2	33	-9.54	91.01
R2	45	2.46	6.05
R2	52	9.46	89.49
R2	39	-3.54	12.53
R2	51	8.46	71.57
R2	54	11.46	131.33
R2	55	12.46	155.25
N1	32	-10.54	111.09
N1	35	-7.54	56.85
N1	39	-3.54	12.53
N1	40	-2.54	6.45
N1	42	-0.54	0.29
N1	42	-0.54	0.29
N1	43	0.46	0.21
N1	45	2.46	6.05
N1	45	2.46	6.05

IV Levels	Mindset Quiz Scores	minus Total mean	squared
N1	46	3.46	11.97
N1	47	4.46	19.89
N1	51	8.46	71.57
N2	42	-0.54	0.29
N2	32	-10.54	111.09
N2	38	-4.54	20.61
N2	50	7.46	55.65
N2	48	5.46	29.81
N2	51	8.46	71.57
N2	38	-4.54	20.61
N2	48	5.46	29.81
N2	48	5.46	29.81
N2	51	8.46	71.57
N2	51	8.46	71.57
N2	48	5.46	29.81
Sum:	$\sigma = 2042$	**N/A**	$\sigma = 2240$

$$SS_T = \sum \left[\left(X - \overline{X}_T \right)^2 \right] = 2240$$

Okay, now we can do the Within Groups (Error) Sum of Squares! The easiest of all!

Within Groups (Error) SS

We get this Sum of Squares indirectly, through subtraction.

Example 10.3.1.6

Calculate the Within (Error) Sum of Squares.

Solution

$$SS_{WG} = SS_T - SS_{BG} - S_P$$

$$SS_{WG} = 2240 - 244.20 - 1296.42 = 699.40$$

If you saved more decimal points for all of the equations, you might end up with something like 699.19; that's fine!

Yay! You finished the Sum of Squares! Unfortunately, because the Within Group (Error) Sum of Squares is calculated indirectly, we cannot do a computation check for the Sum of Squares. :(

We're not done yet!

Now, plug those four Sums of Squares into the ANOVA Summary Table (Table **10.3.1.5**) so that you can fill out the rest of the table to calculate the final ANOVA F-value.

Table **10.3.1.5**- ANOVA Summary Table with SS only

Source	SS	df	MS	F
Between	144.72			
Participants	1296.40			
Error	699.40			
Total	2240.00			

? Exercise 10.3.1.3

Fill out the rest of the ANOVA Summary Table.

Answer

Table **10.3.1.6**- ANOVA Summary Table with Formulas

Source	SS	df	MS	F
Between	144.72	$k-1=4–1=3$	$\frac{SS_B}{df_B}=\frac{144.72}{3}=48.24$	$\frac{MS_B}{MS_E}=\frac{48.24}{21.19}=2.28$
Participants	1296.40	$P-1=12-1=11$	N/A	N/A
Error	699.40	$(k-1)*(p-1)3*11=33$	$\frac{SS_E}{df_E}=\frac{699.40}{33}=21.19$	N/A
Total	2240.00	$N-1=48–1=47$	N/A	N/A

*Your answer might also be 3.84 if you use a spreadsheet that keeps all of the decimals.

You can still do a computation check with the degrees of freedom:

$$df_T=df_B++df_{Ps}=df_E=3+11+33=47$$

Since we did that part correctly, we can move on to the next step of the process with this final ANOVA Summary Table (Table **10.3.1.7**):

Table **10.3.1.7**- Completed ANOVA Summary Table for Mindset by English Class

Source	SS	df	MS	F
Between	144.72	3	48.24	2.28*
Participants	1296.40	11		
Error	699.40	33	21.19	
Total	2240.00	47		

*Your answer might also be 3.84 if you use a spreadsheet that keeps all of the decimals.

Step 4: Make the Decision

Based on the completed ANOVA Summary Table (Table **10.3.1.7**), our calculated F-score is 3.84. If you remember all the way back to Step 2, we found our critical F-score to be 2.92.

(Critical < Calculated) = Reject null = At least one mean is different from at least one other mean. = $p < .05$

(Critical > Calculated) = Retain null = All of the means are similar. = $p > .05$

Exercise 10.3.1.4

Should we retain or reject the null hypothesis? Are we saying that the means are similar, or that at least one mean is different from at least one other mean?

Answer

We reject the null hypothesis because our critical value is smaller than our calculated value. This means that at least one time period's average Mindset Quiz score was different from at least one other time period's average Mindset Quiz score groups.

Example 10.3.1.6

Should we conduct pairwise comparisons? Why or why not?

Solution

Since we rejected the null hypothesis that all of the means are similar, we know that at least one mean is different from at least one other mean. But since we don't know yet which means are different from which other means, we need to conduct pairwise comparisons to find if the differences in the means match our research hypothesis.

To determine which time period's means are different from each other, we can either calculate the critical value (in this case, we'll use Tukey's HSD) or we can start by calculating the differences between each mean from each other. Since we have been finding the critical value (Step 2 in our process) before calculating the test statistic (Step 3 in our process), let's keep that pattern.

Example 10.3.1.7

Using Tukey's HSD, what is the critical value for differences between each pair of means?

Solution

The formula for Tukey's HSD

$$HSD = q * \sqrt{\frac{MSw}{n_{group}}}$$

requires information from the ANOVA Summary Table, as well as the q-value from a table. We've been using a table of critical q-values from RealStatistics.com: https://www.real-statistics.com/statistics-tables/studentized-range-q-table/ Just make sure to use the Alpha = 0.05 set of tables, and use k (the number of groups) not the Between Groups Degrees of Freedom.

Using that table of critical q-values, we find with k = 4 and the Degrees of Freedom from the Within Groups (Error) row (df_{WG} = 33) that the critical q-value is 3.825. Let's plug that into our formula:

$$HSD = \left(q \times \sqrt{\frac{MSw}{n_{group}}}\right)$$

$$HSD = \left(3.825 \times \sqrt{\frac{21.19}{12}}\right)$$

$$HSD = \left(3.825 \times \sqrt{1.77}\right)$$

$$HSD = (3.825 \times 1.33)$$

$$HSD = 5.08$$

Now that we have our critical value, let's calculate each pair of mean differences.

Example 10.3.1.8

What is the difference between each pair of means?

Solution

Using the means from Table 10.3.1.1, I subtract each mean from each other mean:

$$\bar{X}_{R1} - \bar{X}_{R2} = 39.17 - 43.33 = -4.16$$

$$\bar{X}_{R1} - \bar{X}_{N1} = 39.17 - 42.25 = -3.08$$

$$\bar{X}_{R1} - \bar{X}_{N2} = 39.17 - 45.42 = -6.25$$

$$\bar{X}_{R2} - \bar{X}_{N1} = 43.33 - 42.25 = 1.08$$

$$\bar{X}_{R2} - \bar{X}_{N2} = 43.33 - 45.42 = -2.09$$

$$\bar{X}_{N1} - \bar{X}_{N2} = 42.25 - 45.42 = -3.17$$

Comparing the absolute value of each mean difference to the critical mean difference of Tukey's HSF=5.08, we find that only one pair of means is significantly different from each other; the mean of the Beginning of Remedial English ($\bar{X}_{R1} = 39.17$) is statistically lower than the mean of the End of the Next English Class ($\bar{X}_{N2} = 45.42$).

Write-Up

Okay, here's the big finish!

Exercise 10.3.1.5

Write a conclusion to describe the results of the analysis. Don't forget to include the four components necessary in any report of results.

Answer

The researchers hypothesized that students' average Mindset Quiz scores will be higher in the beginning of the semesters than at the end, and higher in the next class than in the first (remedial) class. This research hypothesis was *partially* supported (F(3,33)=3.84, p<0.05) because the mean of the Beginning of Remedial English ($M_{R1} = 39.17$) is statistically lower than the mean of the End of the Next English Class ($M_{N2} = 45.42$). All other means were similar ($M_{R2} = 43.33, M_{N1} = 42.25$). This suggests that it takes a long time to change mindset, or that the journaling activity was not a strong intervention.

One addition to this write-up that isn't part of the four required components is that last sentence that's more of a conclusion about the whole study. The class that you are taking right now isn't just to teach you how to calculate statistics, or even to interpret them, but to think about what the results mean. If I was a college administrator, and the English professors asked for funding for journals or a paid journaling account to improve mindset, I might not fund them. I passionately want students to understand that success is possible if they keep trying and learning from their mistakes, but this data is not strong support that the journaling activity as a way to change students' minds.

You did it! Take a break and reward yourself!

Next up, we'll look at what to do if you don't think that your distribution is normally distributed when you have three or more related groups…

This page titled 10.3.1: Practice with Mindset is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

10.4: Non-Parametric RM ANOVA

Refresher on Non-Parametric Statistical Analyses

Remember when we talked about circumstances when we can't calculate these parametric analyses?

If not, non-parametric statistics are analyses that don't require the population data to be normally distributed. If the data are not normally distributed, then we can't compare means. Because there is no center!

Note

When might non-normally distributed data happen? If you're not sure, look back at the non-parametric analyses section of the chapter on independent t-tests…

If the data are not normally distributed, then our formulas don't provide a description of the center of the distribution. Instead, we analyze the ranks of the scores, not the scores themselves. This applies to all of the following analyses:

- Mean
- Standard deviation
- t-test
- *ANOVAs*
- Pearson's correlation

Alternatives

So far, we've talked about non-parametric analyses for t-tests and Between Groups ANOVAs.

Example 10.4.1

What were the non-parametric alternatives for:

- Independent t-test?
- Dependent t-test?
- BG ANOVA?

Solution

- Independent t-test: Mann-Whitney U
- Dependent t-test: Wilcoxon Matched-Pairs Signed-Rank
- BG ANOVA: Kruskal-Wallis H

What is the non-parametric alternative to a Repeated Measures ANOVA, you might ask in this page on non-parametric alternatives to RM ANOVAs? The answer is Friedman's test. The Friedman test is a non-parametric statistical test developed by Milton Friedman that used to detect differences in treatments across multiple test attempts. The procedure involves ranking each row (or block) together, then considering the values of ranks by columns.

Because you will probably never have to do this without statistical software, we aren't going to go over the formula here. But if you are interested, the Wikipedia page includes the formula and some other interesting information.

Contributors and Attributions

- Dr. MO (Taft College)
- Wikipedia: https://en.Wikipedia.org/wiki/Friedman_test

This page titled 10.4: Non-Parametric RM ANOVA is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

CHAPTER OVERVIEW

11: Factorial ANOVA (Two-Way)

- 11.1: Introduction to Factorial Designs
  - 11.1.1: Factorial Notations and Square Tables
- 11.2: Introduction to Main Effects and Interactions
  - 11.2.1: Example with Main Effects and Interactions
  - 11.2.2: Graphing Main Effects and Interactions
  - 11.2.3: Interpreting Main Effects and Interactions in Graphs
  - 11.2.4: Interpreting Interactions- Do Main Effects Matter?
  - 11.2.5: Interpreting Beyond 2x2 in Graphs
- 11.3: Two-Way ANOVA Summary Table
  - 11.3.1: Calculating Sum of Squares for the Factorial ANOVA Summary Table
- 11.4: When Should You Conduct Post-Hoc Pairwise Comparisons?
- 11.5: Practice with a 2x2 Factorial Design- Attention
  - 11.5.1: Practice 2x3 Factorial ANOVA on Mindset
- 11.6: Choosing the Correct Analysis- Mean Comparison Edition

---

This page titled 11: Factorial ANOVA (Two-Way) is shared under a not declared license and was authored, remixed, and/or curated by Michelle Oja.

11.1: Introduction to Factorial Designs

Research Designs

Phew, we've covered a lot in the past few chapters! Each of the inferential statistical analyses that we've covered is to compare means (quantitative variable) between different groups. The qualitative variable is the two different samples or an IV with two or more levels. With t-tests, with compared two different groups (which can be considered one IV with two levels). Sometimes, those groups were unrelated (independent t-test) and sometimes they were related (dependent t-test). With ANOVAs, we compared one IV with two or more levels. Sometimes those IV levels were unrelated (Between Groups ANOVA) and sometimes they were related (Repeated Measures ANOVA, sometimes called Within-Groups ANOVA).

ANOVAs are an amazing statistical tool because they also allow is to compare means for a *combination* of IVs (not just IV levels). This happens in a factorial design, when each level of each IV is combined so that a set of participants experiences the *combination* of levels of each IV.

Let's try an example. Let's say that we wanted to test to see if mindset (growth versus fixed mindset) affects how long students spend on their homework.

? Exercise 11.1.1

What would be the IV and IV levels? What would be the DV?

Answer

The IV would be Mindset, with the levels being growth mindset or fixed mindset.

The DV would be how long student spend on their homework.

As an educational researcher, I also know that outside obligations have a lot to do with how much time students can spend on their homework. To test for this, let's add another IV to our experiment called "Job" with the levels of: Full-Time, Part-Time, None.

Now we have a factorial design! We have two IVs, Mindset and Job. Mindset has two levels and Job has three levels. We have one DV, how long student spend on homework. This means that I'll measure how much time students spend on homework for:

- Students with growth mindset who work full-time,
- Students with growth mindset who work part-time,
- Students with growth mindset who don't work,
- Students with fixed mindset who work full-time,
- Students with fixed mindset who work part-time, and
- Students with fixed mindset who don't work.

This helps us understand how mindset affects homework time for students who work different hours and can give us a lot more information than a study with just one of those IVs. For example, folks with growth mindset know that success comes after putting in time and learning from their mistakes, so maybe whether you work full-time or doesn't matter for students with growth mindset, but how much you work does affect how long students with fixed mindset spend on homework. Can you see how combining the IVs gives us more information?

Cause & Effect

Researchers often use factorial designs to understand the causal influences behind the effects they are interested in improving. Effects are the change in a measure (DV) caused by a manipulation (IV levels). You get an effect any time one IV causes a change in a DV.

Distraction Scenario

Here is an example. We will stick with this one example for a while, so pay attention… In fact, the example is about paying attention! Let's say you wanted to measure something like paying attention. You could something like this:

1. Pick a task for people to do that you can measure. That will be the dependent variable. This is the effect.
2. Pick something that you think will cause differences in paying attention. For example, we know that people can get distracted easily when there are distracting things around. This is your IV. You could have two levels for your manipulation: No distraction versus distraction. This is what we think is the cause of changes in the DV.
3. Measure performance in the task (DV) under the conditions (IV levels or groups).
4. If your distraction manipulation changes how people perform the task, you may have successfully manipulated how well people can pay attention in your task.

1. Pick a Task (DV)

First, we pick a task. It's called Spot the Difference. You may have played this game before. You look at two pictures side-by-side, and then you locate as many differences as you can find. Figure **11.1.1** is an example:



Figure **11.1.1**: Two Images with Minor Differences (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

How many differences can you spot? When you look for the differences, it feels like you are doing something we would call "paying attention". If you pay attention to the clock tower, you will see that the hands on the clock are different. Yay! One difference spotted. We could give people 30 seconds to find as many differences as they can. Then we give them another set of pictures and do it again. Every time we will measure how many differences they can spot. So, our measure of performance, our dependent variable, could be the mean number of differences spotted.

That gives us an effect on our IV. But there are no groups here, no levels. Everyone gets the same pictures and does the same thing, so we're not actually testing anything.

2. Pick a Cause (IV)

Now, let's think about what might cause differences in how people pay attention. If people need to pay attention to spot differences, then presumably if we made it difficult to pay attention, people would spot less differences. What is a good way to distract people? I'm sure there are lots of ways to do this. How about we do the following:

1. No distraction condition: People do the task with no added distractions. They sit in front of a computer, in a quiet, distraction-free room, and find as many differences as they can for each pair of pictures. This is the comparison condition, and is sometimes called a "control" because the participants experience the "nothing" version of the IV levels.
2. Distraction condition: We blast super loud ambulance sounds and fire alarms and children's music while people attempt to spot differences. We also randomly turn the sounds on and off, and make them super-duper annoying and distracting. We make sure that the sounds aren't loud enough to do any physical damage to anybody's eardrums, but loud enough to be super distracting. If you don't like this, we could also tickle people with a feather, or whisper silly things into their ears, or surround them by clowns, or whatever we want, it just has to be super distracting.

3. & 4. Measure Performance to Find an Effect

If our distraction manipulation is super-distracting, then what should we expect to find when we compare spot-the-difference performance between the no-distraction and distraction conditions? We should find a difference! If our manipulation works, then we should find that people find more differences when they are not distracted, and less differences when they are distracted.

Imagine that the results show that people found an average of four differences when they were distracted, and an average of 10 differences when they were not distracted. The effect of distraction is a mean of six differences found in the picture. It's the difference between performance in the Distraction and No-Distraction conditions. In general, it is very common to use the word effect to refer to the differences caused by the IV. We manipulated distraction, it caused a difference, so we call this the "distraction effect".

How is this related to Factorial Designs?

You might be wondering how this scenario on distractions and Spotting the Difference is related to factorial designs and ANOVAs since a t-test would be great at analyzing this data. The answer is that science is cumulative.

We have done the hard work of finding an effect of interest, in this case the distraction effect. We think this distraction effect actually measures something about your ability to pay attention. But why stop there? If we think that we found an effect of distraction, maybe we could find a way to reduce distraction to help students focus on their homework? Or help employees focus on their jobs? Or help my partner focus on the intricate plot of TV show that we're watching? Or what if we found some individual differences, meaning that some people get distracted easier than other people? You were the kind of person who had a small distraction effect (maybe you find 10 differences when you are not distracted, and 9 differences when you are distracted), that could mean you are very good at ignoring distracting things while you are paying attention. On the other hand, you could be the kind of person who had a big distraction effect (maybe you found 10 differences under no distraction, and only one difference when you were distracted); this could mean you are not very good at ignoring distracting things while you are paying attention.

Overall now, we are thinking that our effect of distraction (IV) on spotting the difference (DV) by finding the difference in performance between the two conditions might be affected by more than just being distracted. This may lead us to want to know how to make people better at ignoring distracting things. Our first stab at science found that distraction affected paying attention, but we want to expand on that finding. Maybe now we want to know what makes people worse at ignoring things? In other words, we want to find out what other IVs might affect the size of the distraction effect (make it bigger or smaller, or even flip around!).

We are going to add a *second* IV to our task. Note that this is not a new level to our Distraction IV, but a whole new IV. The second IV will be reward. In one condition, people will get $5 for every difference they find (so they could leave the study with lots of money if they find lots of differences). In the other condition, people will get no money, but they will still have find differences (comparison control condition).

We could run this study on it's own, comparing receiving a reward to not receiving a reward. But this will be a factorial design, so everybody will have to find differences. A quarter of the sample will find differences when they are distracted but they will get a reward for each difference that they find. A quarter of the sample will find differences whey they are distracted, but won't know anything about a reward. A quarter of the sample will look for differences without being distracted, but they will be rewarded for each difference they found. And finally, a quarter of the sample will look for differences without being distracted and will also not earn a reward. Thus, the factorial design has the folloing combinations:

- Distracted with reward
- Distracted with no reward
- Not distracted with reward
- Not distracted with no reward

Note that the DV is the same for all of the IV conditions: how many differences were spotted.

Results

The question we are now asking is: Will the reward IV cause a change in the size of the distraction effect? We could predict that people receiving rewards will have a smaller distraction effect than people not receiving rewards.

- **No-Reward condition**: In the no-reward condition people played Spot the Difference when they were distracted and when they were not distracted. This is a replication of our first study. We should expect to find the same pattern of results, and that's what we found: People found six differences when they were distracted and 10 when they were not distracted. So, there was a distraction effect of four, same as we had last time.
- **Reward condition**: In the reward condition people played Spot the Difference when they were distracted and when they were not distracted. Except, they got $5 every time they spotted a difference. We predicted this would cause people to pay more attention and do a better job of ignoring distracting things. This is what happened. People found nine differences when they were distracted and 11 when they were not distracted. So, there was a distraction effect of two.

We conclude that reward can manipulate the distraction effect. When there was no reward, the size of the distraction effect was four. When there was reward, the size of the distraction effect was two. So, the reward manipulation changed the size of the distraction effect by two (4-2=2).

Why Factorial Designs?

Factorial designs are so useful because they allow researchers to find out what kinds of variables can cause changes in the effects they measure. We measured the distraction effect, then we found that reward causes changes in the distraction effect. If we were trying to understand how paying attention works, we would then need to explain how it is that reward levels could causally change how people pay attention (because science is cumulative). We would have some evidence that reward does cause change in paying attention, and we would have to come up with some explanations, and then run more experiments to test whether those explanations hold water.

That's the basic idea of factorial designs, but of course it gets more complex pretty quickly. To help researchers quickly identify what's going on in a factorial design, we use specific notation. It's a lot of new information for students, so pay attention!

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 11.1: Introduction to Factorial Designs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **9.2: Purpose of Factorial Designs** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

11.1.1: Factorial Notations and Square Tables

2x2 Designs

We've just started talking about a 2x2 Factorial design, which means that we have two IVs (the number of numbers indicates how many IVs we have) and each IV has two levels (the numbers represent the number of level for each IV). We said this means the IVs are crossed. To illustrate this, take a look at the following tables. Table **11.1.1.1** is a conceptual version. Although not exactly accurate, many call these types of tables a Punnett Square because it shows the combination of different levels of two categories.

Table **11.1.1.1**- Conceptual Example of a 2x2 Factorial Design

IV Levels	**IV1 Level I**	**IV1 Level 2**
IV2 Level 1	DV	DV
IV2 Level 2	DV	DV

Our study on distraction is a 2x2 design, so what would that look like in this type of table?

Example 11.1.1.1

Create a "Punnett's Square" for the IVs and DV of the Distraction scenario.

Solution

Table **11.1.1.2**- Factorial Design of Distraction Scenario

IV Levels	**IV1 (Distraction): Yes**	**IV1 (Distraction): No**
IV2 (Reward): Yes	DV = Number of differences spotted	DV = Number of differences spotted
IV2 (Reward): No	DV = Number of differences spotted	DV = Number of differences spotts

You could have just as easily made IV1 the reward and IV2 the Distraction, and the table would still be correct.

Let's talk about this crossing business. Here's what it means for the design. For the first level of Distraction (Yes), we measure the number of differences spotted performance for the people who were rewarded, as well as for the people who were not rewarded. So, for the people who were distracted we also manipulated whether or not they earned a reward. In the second level of the Distraction IV (No), we also manipulate reward, with some people earning a reward and some people not. We collect how many differences were spotted in all conditions.

We could say the same thing, but talk from the point of view of the second IV. For example, for participants who were rewarded, some are distracted and some are not. Similarly, for participants who were not rewarded, we distract some of the participants and don't distract some of them.

Each of the four squares representing a DV, is called a condition. So, we have 2 IVs, each with 2 levels, for a total of 4 conditions. This is why we call it a 2x2 design. 2x2 = 4. The notation tells us how to calculate the total number of conditions.

Factorial Notation

Anytime all of the levels of each IV in a design are fully crossed, so that they all occur for each level of every other IV, we can say the design is a fully factorial design. We use a notation system to refer to these designs. The rules for notation are as follows. Each IV get's it's own number. The number of levels in the IV is the number we use for the IV. Let's look at some examples:

2x2 = There are two IVS, the first IV has two levels, the second IV has 2 levels. There are a total of 4 conditions, 2x2 = 4.

2x3 = There are two IVs, the first IV has two levels, the second IV has three levels. There are a total of 6 conditions, 2x3 = 6

3x2 = There are two IVs, the first IV has three levels, the second IV has two levels. There are a total of 6 conditions, 3x2=6.

4x4 = There are two IVs, the first IV has 4 levels, the second IV has 4 levels. There are a total of 16 condition, 4x4=16

2x3x2 = There are a total of three IVs. The first IV has 2 levels. The second IV has 3 levels. The third IV has 2 levels. There are a total of 12 condition. 2x3x2 = 12.

Let's practice a little with this notation.

? Exercise 11.1.1.1

What is the factorial design notation for a study with two IVs, one has 2 levels and the other has 3 levels?

Answer

2x3

There are two IVs, so there are two numbers. Each number represents the number of levels for each IV.

? Exercise 11.1.1.2

What is the factorial design notation with a study with the following IVs:

2 (task presentation: computer or paper) by

2 (task difficulty: easy or hard) by

2 (student: high school or college)

Answer

2x2x2

There are three IVs, so there are three numbers. Each IV only has two levels, so there are three two's! Notice that there are no threes. The number of IVs is represented in the *number* of numbers.

Okay, let's try something a little more challenging.

Let's do a couple more to make sure that we have this notation business down.

? Exercise 11.1.1.5

For one of Dr. MO's dissertation studies, participants read about a character, then rated that character on several personality traits (DV). The race and gender of the character were varied systematically. Here are the IVs:

- Race of participant: White or Black
- Gender of participant: Woman or man
- Race of character: White or Black or none/neutral (the character's race was not mentioned)
- Gender of character: Woman or man or none/neutral (the character's gender was not mention)

What kind of factorial design was this study? 2x2? 2x3? Something else (what?)?

Answer

This was a 2x2x3x3 study because:

- Race of participant: White or Black = 2
- Gender of participant: Woman or man = 2
- Race of character: White or Black or none/neutral = 3
- Gender of character: Woman or man or none/neutral -= 3

Dr. MO wanted 30 participants in each cell, so she had to have 1,080 participants!

Last one! This one has a few more questions to better understand the scenario.

Exercise 11.1.1.6

Dr. MO has more Star Wars collectibles than can fit in her office, and she'd like to sell some. Her research question is whether she'd get a better price through CraigsList or eBay? Also, should she take the picture or use a stock photo?

1. What is the DV? (It is not explicitly labeled.)
2. For each IV, what are the levels?
3. Is this a 2x2 factorial design? If not, what kind of design is it?
4. List out each of the *combinations* of the levels of the IVs.

Answer

1. The DV is the price, or how much Dr. MO could earn for selling each collectible.
2. IV1's levels are CraigsList or eBay, so the IV name could be something like "website" or "platform". IVs' levels are personal photo or stock photo, so the IV name could be something like "Photo Type."
3. This is a 2x2 factorial design: 2 (Platform: CraigsList or eBay) by 2 (Photo Type: Personal or Stock)
4. List out each of the *combinations* of the levels of the IVs:
    - Posted on CraigsList with a personal photo.
    - Posted on Craigslists with a stock photo.
    - Posted on eBay with a personal photo.
    - Posted on eBay with a stock photo.

Just for fun, let's illustrate a 2x3 design using the same kinds of tables we looked at before for the 2x2 design.

Table **11.1.1.3**- Conceptual Example of a 2x3 Factorial Design

IV Levels	**IV1 Level I**	**IV1 Level 2**
IV2 Level 1	DV	DV
IV2 Level 2	DV	DV
IV2 Level 3	DV	DV

Our very first example of tiem spent studying is a 2x3 design, so what would that look like in this type of table?

Example 11.1.1.2

Create a "Punnett's Square" for the IV of Mindset (Growth or Fixed) and the IV of Job (Full-Time, Part-Time, or None) for time spend studying.

Solution

Table **11.1.1.4**- Factorial Design of Studying Scenario

IV Levels	**IV1 (Mindset): Growth**	**IV1 (Mindset): Fixed**
IV2 (Job): Full-Time	DV = Minutes spent studying	DV = Minutes spent studying
IV2 (Job): Part-Tiime	DV = Minutes spent studying	DV = Minutes spent studying
IV2 (Job): None	DV = Minutes spent studying	DV = Minutes spent studying

You could have just as easily made IV1 the Job and IV2 the Mindset, and the table would still be correct.

All we did was add another row for the second IV. It's a 2x3 design, so it should have 6 conditions. As you can see there are now 6 cells to measure the DV.

You might have noticed in the list of notation for different factorial designs that you can have three IVs (that's the 2x3x2 design). In fact, you can have as many IVs with as many levels as you'd like, but the Central Limit Theorem shows (through complicated math that we aren't going to go into) that each condtion (or cell) should have at least 30-50 participants, that can get expensive quickly! If a 2x2 has 4 conditions, and you want at least 30 participants in each condition, then you'd need 120 participants. If you have a 2x3, then you'd need at least 180 participants (2 * 3 * 30 = 180). So, for a 2x3x2, how many participants would you want?

$$\text{Participants} = 2 * 3 * 2 = 12$$
$$\text{Participants} = 12 * 30 = 360$$

Students always want to know how we would represent more than two IVs in a Punnett's Square, and the answer is that we don't. We create two Punnett Squares.

Let's say that we were looking at time spend studying for those with different mindsets and who have different jobs for different kinds of schools (community colleges or universities). That could look like:

Table **11.1.1.5**- Factorial Design of Studying Scenario FOR COMMUNITY COLLEGE STUDENTS

IV Levels	**IV1 (Mindset): Growth**	**IV1 (Mindset): Fixed**
IV2 (Job): Full-Time	DV = Minutes spent studying MATH	DV = Minutes spent studying MATH
IV2 (Job): Part-Tiime	DV = Minutes spent studying	DV = Minutes spent studying
IV2 (Job): None	DV = Minutes spent studying	DV = Minutes spent studying

AND:

Table **11.1.1.5**- Factorial Design of Studying Scenario FOR UNIVERSITY STUDENTS

IV Levels	**IV1 (Mindset): Growth**	**IV1 (Mindset): Fixed**
IV2 (Job): Full-Time	DV = Minutes spent studying MATH	DV = Minutes spent studying MATH
IV2 (Job): Part-Tiime	DV = Minutes spent studying	DV = Minutes spent studying
IV2 (Job): None	DV = Minutes spent studying	DV = Minutes spent studying

You could have just as easily made IV1 the Job and IV2 the Mindset, or even made a table for only students with Growth Mindset (and had IV1 be the type of school) and another table for only students with Fixed Mindset and the table would still be correct. It doesn't matter statistically which IV is placed where, it's more about interpreting and understanding what is besting tested.

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 11.1.1: Factorial Notations and Square Tables is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **9.1: Factorial Basics** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

11.2: Introduction to Main Effects and Interactions

When you conduct a design with more than one IV, you get more means to look at. As a result, there are more kinds of questions that you can ask of the data.

What kinds of questions are we talking about? Let's keep going with our Distraction scenario experiment. We have the first IV where we manipulated distraction. So, we could find the overall means in Spot the Difference activity for the distraction vs. no-distraction conditions (that's two means, one for when participants were distracted and one for the participants who were not distracted). The second IV was reward. We could find the overall means in spot-the-difference performance for the reward vs. no-reward conditions (that's two more means, one for participants who received a reward and one mean when reward was not mentioned). We could do what we already did, and look at the means for each combination, that is the mean for distraction/reward, distraction/no-reward, no-distraction/reward, and no-distraction/no-reward (that's four more means, if you're counting). There's even more. We could look at the mean distraction effect (the difference between distraction and no-distraction) for the reward condition, and the mean distraction effect for the no-reward condition (that's two more). I hope you see here that there are a lot of means to look. And they are all different means. Let's look at all of them together in one graph with four panels (Figure **11.2.1**). Remember, the bar graphs in Figure **11.2.1** show the mean for each group; they are not frequency charts.

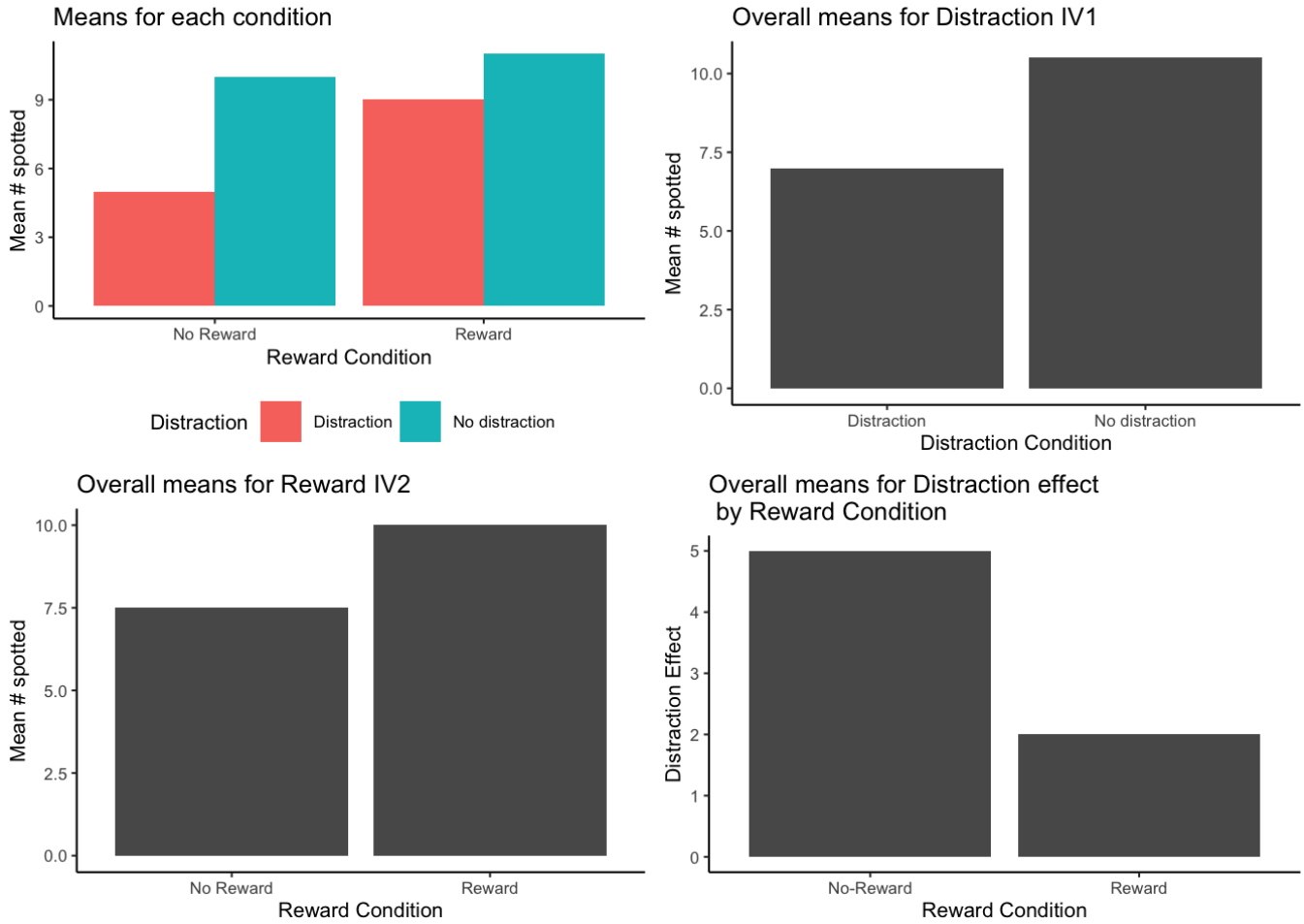


Figure **11.2.1**: Each panel shows the mean for different effects in the design. (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

The purpose of showing all of these means is to orient you to your problem. If you conduct a 2x2 design (and this is the most simple factorial that you can conduct), you will get all of these means. You need to know what you want to know from the means. That is, you need to be able to connect the research question to the specific means you are interested in analyzing.

For example, in our example, the research question was whether reward would change the size of the distraction effect. The top left panel gives us some info about this question. We can see all of the condition means, and we can visually see that the distraction effect was larger in the No-reward compared to the reward condition. But, to “see” this, we need to do some visual subtraction. You need to look at the difference between the red and aqua bars for each of the reward and no-reward conditions.

Does the top right panel tell us about whether reward changed the size of the distraction effect? NO, it just shows that there was an overall distraction effect (this is called the **main effect** of distraction). Main effects are any differences between the levels of one independent variable.

Does the bottom left panel tell us about whether reward changed the size of the distraction effect? NO! it just shows that there was an overall reward effect, called the main effect of reward. People who were rewarded spotted a few more differences than the people who weren’t, but this doesn’t tell us if they were any less distracted.

Finally, how about the bottom right panel. Does this tell us about whether the reward changed the size of the distraction effect? YES! Notice, the y-axis is different for this panel. The y-axis here is labeled “Distraction Effect”. You are looking at two *difference scores*. The distraction effect in the no-reward condition (10-5= 5), and the distraction effect in the Reward condition (11-9 = 2). These two bars are different as a function of reward. So, it looks like reward did produce a difference between the distraction effects! This was the whole point of the study. It is these means that were most important for answering the question of the study. As a very last point, this panel contains what we call an **interaction**. We explain this more in the next section, but basically, an interaction shows that the effect of on IV on the DV is changed by the effect of another IV on the DV. The IVs levels can combine *or* cancel each other out, but the point is that you wouldn't get the full picture of how the levels of either IV affect the DV if you only looked at one at a time. A factorial design lets you look at the combination of the IVs together.

> *Pro tip: Make sure you know what you want to know from your means before you run the study, otherwise you will just have way too many means, and you won’t know what they mean.*

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 11.2: Introduction to Main Effects and Interactions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **9.4: Knowing what you want to find out** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

11.2.1: Example with Main Effects and Interactions

You might realize by now the Dr. MO's student researchers did a couple studies on mindset. Remember growth mindset? The idea that mistakes help you learn, and that more time and effort leads to brain development? One group of student researchers collected data on improving growth mindset that should be analyzed with a factorial ANOVA. Here it is:

Scenario

The student researchers looked at the Difference scores of 106 students at Dr. MO's community college. The Difference scores were calculated by subtracting each students' Post-Test Mindset Quiz score (measured at the end of the semester) from their own Pre-Test Mindset Quiz score (measured at the beginning of the same semester); thus, positive scores mean that the student improved their mindset. In this study, we had two variables. One was the intervention, which had two levels: No Intervention (comparison control group) and Intervention (in which faculty tried different activities to improve mindset). We also collected information on what department these activities were happening in, and had enough data to analyze students in English classes and students in Psychology classes.

Can you identify the groups and variables in this scenario?

Example 11.2.1.1

Answer the following questions to understand the variables and groups that we are working with.

1. Who is the sample?
2. Who do might be the population?
3. What are the IVs and levels for each IV?
4. What is the DV (quantitative variable being measured)?
5. Is this a 2x2 factorial design? If not, what kind of design is it?
6. List out each of the *combinations* of the levels of the IVs:

Solution

1. The sample was 106 community college students taking an English class or a Psychology class.
2. The population could be all community colleges students, or maybe all community college students taking English or Psychology. Or maybe all community college students in a general education course?
3. One IV is Intervention, with the levels being Yes or No. The other IV is Department, with the levels being English or Psychology.
4. The DV is the Difference score from the Mindset Quiz pre-test to the Mindset Quiz post-test.
5. Yes, this is a 2x2 factorial design because there are two IVs (so there are two numbers __x__); the first IV (Intervention) has two levels and the second IV (Department) has two levels.
6. List out each of the *combinations* of the levels of the IVs:
    1. Intervention in English
    2. Intervention in Psychology
    3. No Intervention in English
    4. No Intervention in Psychology

Can you plug these IVs into a Punnett's Square grid?

Example 11.2.1.2

Complete a grid showing the factorial design IVs and DVS.

Solution

Table **11.2.1.1**- 2x2 Factorial Grid of Intervention by Department

IV Levels	**IV1- Yes Intervention**	**IV1- No Intervention**
IV2: English Department	2x2: Students who experienced an intervention in an English class.	2x2: Students who did not experience an intervention but who were in an English class.
IV2: Psychology Department	2x2: Students who experienced an intervention in a Psychology class.	2x2: Students who did not experience an intervention but who were in a Psychology class.

You could have put Department as IV1 (columns) and Intervention as IV2 (rows). It's really about what makes the most sense to you; Dr. MO wanted what she thinks is the main influence of mindset on the top.

Participants in each "cell" of this design have a unique combination of IV conditions.

Three Effects

With a 2x2 factorial design, you have three effects to look at. Remember, "effects" are the results of the DV, what was measured. Here are the three effects that you need to look at:

1. The main effect of the one IV: How does one IV affect the DV (independent of the other IV)
2. The main effect of the other IV: How does the other IV affect the DV (independent of the first IV)
3. The interaction of the two IVs -- how they **jointly** affect the DV

Example 11.2.1.3

In our mindset scenario, what are these three effects?

Solution

1. The main effect of the intervention; did the intervention improve Mindset Quiz scores?
2. The main effect of the department; did which class you were in affect Mindset Quiz scores?
3. The interaction of the intervention by department; is one department more likely to improve Mindset Quiz scores than another department when there's an intervention? Another way to think about these variables is to ask whether the departments started out with different average Mindset Quiz scores, so that even in the No Intervention condition the Difference score would be statistically significantly different between the two departments.

Let's look at these main effects in Table **11.2.1.2**, in which the marginal means were included. Marginal means are, you guessed, it the means on the margins of the table. These means on the margin show the means for each level of each IV, which are the *main effects*. The marginal means do not show the combination of the IVs' levels, so they do not show an interaction.

Table **11.2.1.2**- 2x2 Factorial Grid of Intervention by Department with Marginal Means

IV Levels	**IV1- Yes Intervention**	**IV1- No Intervention**	Marginal Means of IV2
IV2: English Department	2x2: Students who experienced an intervention in an English class.	2x2: Students who did not experience an intervention but who were in an English class.	**DV = 2.49**
IV2: Psychology Department	2x2: Students who experienced an intervention in a Psychology class.	2x2: Students who did not experience an intervention but who were in a Psychology class.	**DV = 1.34**
Marginal Means of IV1	**DV = 2.08**	**DV = 1.76**	

Again, you could have put Department as IV1 (columns) and Intervention as IV2 (rows). And again, participants in each "cell" of this design have a unique combination of IV conditions.

Main Effects

Let's go through the marginal means for Table **11.2.1.2**.

Example 11.2.1.1

What are the marginal means for the Intervention?

Solution

Intervention Yes = 2.08

Intervention No =1.76

We'll look at statistical significant later, but just based on these means, it looks like the students who experienced an Intervention had a larger Difference score than students who did not have an intervention. This is what was expected; students who experienced an intervention had a higher Mindset Quiz score at the end of the semester (post-test), after the interventions, than at the beginning of the semester (pretest). This is your *main effect*of Intervention. When we do the statistical analyses, we'll follow the same process for null hypothesis significance testing:

$$\textbf{Critical} < |\textit{Calculated}| = \textbf{Reject null} = \textbf{means are different} = \textbf{main effects} = p < .05$$

$$\textbf{Critical} > |\textit{Calculated}| = \textbf{Retain null} = \textbf{means are similar} = \textbf{no main effects} = p > .05$$

Let's move on to the other independent variable.

Exercise 11.2.1.1

Based on Table **11.2.1.2**, what are the marginal means for department? Which department seemed to have a higher Difference score?

Answer

Marginal Means:

- English=2.49
- Psychology=1.34

There seems to be a main effect of department such that students in English had a higher Difference score than students in Psychology classes.

We can't be say much more than that without looking at actual statistical results. Instead, we will look at the individual cells of our grid to see if there was an *interaction* between Department and Intervention on the Difference scores.

Interaction

Table **11.2.1.3** has the means for each combination of each IV level in the individual cells. Looking at Table **11.2.1.3**, how do the inside cells seem relate to each other?

Table **11.2.1.3**- 2x2 Factorial Grid of Intervention by Department with Marginal Means & Cell Means

IV Levels	**IV1- Yes Intervention**	**IV1- No Intervention**	Marginal Means of IV2
IV2: English Department	**DV=3.70**	**DV=1.29**	**DV=2.49**
IV2: Psychology Department	DV=0.46	DV=2.23	**DV=1.34**
Marginal Means of IV1	**DV=2.08**	**DV=1.76**	

Again, you could have put Department as IV1 (columns) and Intervention as IV2 (rows). And again, participants in each "cell" of this design have a unique combination of IV conditions.

Example 11.2.1.5

Answer the following questions related the cell means in Table **11.2.1.3**,

1. Was one cell substantially higher than the others?
2. Was one cell substantially lower than the others?

Solution

1. It looks like students in English class who experienced an intervention had higher Difference scores.
2. It looks like students in a Psychology class who experienced an intervention had a really low difference score.

This pattern of cell means is your *interaction* of Intervention and Department: Experiencing the intervention *interacts* with the department to affect Differences in Mindset Quiz scores such that the intervention seems to improve mindset for those in English but not in Psychology (the Psychology students who experienced an intervention were essentially unchanged). What's strange is that Psychology students in the control condition (No Intervention) might have actually improved their Mindset Quiz scores more than both English students with no intervention and Psychology student with the intervention! You might be wondering why these strange effects for the Psychology students, and Dr. MO has no answer for you. This is real data, and there's no good explanation for this with the variables that we have. ::shrug::

As you can see, the main effects of each IV can relate to the interaction in several different ways. Let's look at that next.

This page titled 11.2.1: Example with Main Effects and Interactions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

11.2.2: Graphing Main Effects and Interactions

Let's back to our example on distraction and use the following graphs to help interpret the main effects of reward, the main effects of distraction, and to see how they interact.

Distraction Effect: What Do You See?

Results from 2x2 designs are also often plotted with line graphs. Figure **11.2.2.1** shows four different graphs, using bars and lines to plot *the same means*. Dr. Crump wanted to show you this so that you realize that how you graph your data matters, and it makes it more or less easy for people to understand the results. Also, how the data is plotted matters for what you need to look at to interpret the results.

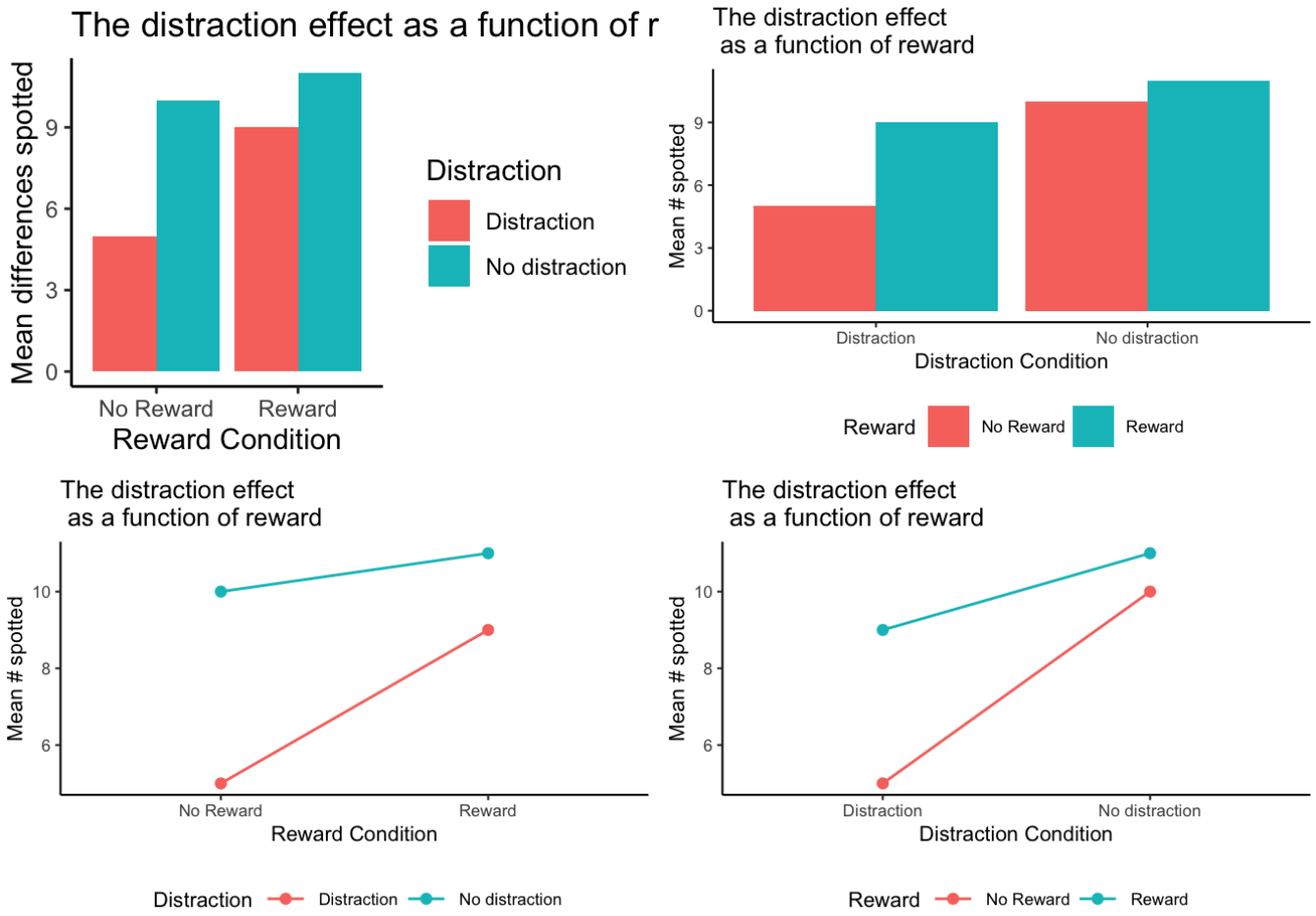


Figure **11.2.2.1**: The same example means plotted using bar graphs or line graphs, and with Distraction or Reward on the x-axis. (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

What do you see in both of the graphs on the left? Dr. MO sees that distraction matters, but distraction is particularly effective in reducing the number of differences spotted when there is no reward. In both graphs on the left of Figure **11.2.2.1**, the one really low bar or dot is when there is no reward and the participant was distracted.

What do you see in the graphs on the right in Figure **11.2.2.1**? The first thing that Dr. MO saw was that the line chart makes it clear that there's a high level of spotting differences in the No Distraction condition. Then she noticed that the dot for participants who were Distracted and Not Rewarded was way low compared to all of the other dots. The bar chart shows that as well; the left most bar is clearly shorter than all of the others.

What you are seeing in these charts are the main effects and interactions. The main effect of IV1 shows that distraction matters; those participants who were distracted spotted fewer distractions. The main effect of IV2 shows that reward matters; those

participants who were rewarded spotted more distractions. And finally, these charts show an interaction. Although both variables mattered, their individual effects seemed to be magnified when combined. Participants who were Distracted and Not Rewarded were pretty horrible at spotting differences, more horrible than either IV's main effect would suggest on its own. That's the power of interactions.

On the Playground: What Do You See?

Let's try another one. Check out Figure **11.2.2.2** and determine the factorial notations. Figure **11.2.2.2** is a bar chart that shows the means of the different conditions (levels of the IV). This is different from the frequency charts that we first learned about. When not listed, it can be helpful to list the IVs, so we'll start with that. The DV is number of times a student was pushed on the playground.

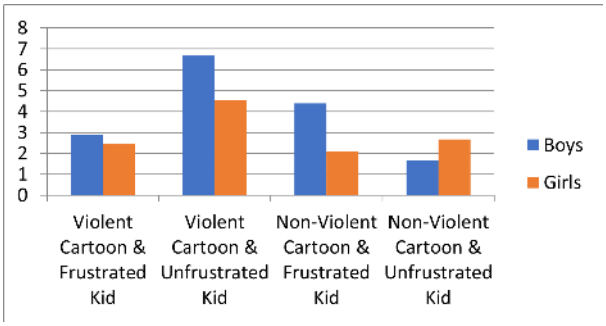


Figure **11.2.2.2**: Example of Factorial Design in a Bar Chart. (CC-BY-SA; graph created by Michelle Oja)

Let's answer the questions that we talked about when identifying factorial notations about Figure **11.2.2.2** to see if we can find the main effects and interactions.

? Exercise 11.2.2.3

For Figure **11.2.2.2**:

1. To determine one main effect, what are the levels of one of the IVs? What could you name that IV?
2. To determine the other main effect, what are the levels of another of the IVs? What could you name that IV?
3. There's one more main effect! What are the levels of the last IV? What could you name that IV?

Answer

It doesn't matter what order you found these in:

1. One IV's level is Boys or Girls, so that could be called the main effect of "Gender".
2. Another IV's levels looks like violent or non-violent cartoons. That main effect could be called "Cartoon" or "Cartoon Type" or "Cartoon Violence."
3. A final IV's levels is frustrated kid or unfrustrated kid. We could call that main effect "Mood".

Now that we see the main effects, let's determine what kind of factorial design is shown in Figure **11.2.2.1**.

? Exercise 11.2.2.4

What kind of factorial designs is shown in Figure **11.2.2.2**? 2x2? 2x3? 2x2x2? Other?

Answer

Figure **11.2.2.2** shows a 2x2x2

2 (Gender: Boy or Girl) by 2 (Cartoon: Violent or Not) by 2 (Mood: Frustrated or Not)

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 11.2.2: Graphing Main Effects and Interactions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **9.3: Graphing the means** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

11.2.3: Interpreting Main Effects and Interactions in Graphs

Designs with multiple factors are very common. When you read a research article (and you will!), you will often see tables and graphs that show the results from designs with multiple factors. It would be good for you if you were comfortable interpreting the meaning of those results. You should look at the main effects, but understand that any statistically significant interaction is the real result because the interaction tells you more about the levels of each IV than any main effect can.

2x2 designs

Let's take the case of 2x2 designs. There will always be the possibility of two main effects and one interaction. You will always be able to compare the means for each main effect and interaction. If the two means from one variable are different, then there is a main effect. If the two means from the other variable are different, then there is a main effect. And then you look at the cell means (the means in the four combinations of the two IV's two levels), and those could have all sorts of relationships. There a bunch of ways all of this can turn out. Check out the ways, there are eight of them:

1. no IV1 main effect, no IV2 main effect, no interaction (This is the null hypothesis!)
2. IV1 main effect, no IV2 main effect, no interaction
3. IV1 main effect, no IV2 main effect, yes interaction
4. IV1 main effect, IV2 main effect, no interaction
5. IV1 main effect, IV2 main effect, yes interaction
6. no IV1 main effect, IV2 main effect, no interaction
7. no IV1 main effect, IV2 main effect, yes interaction
8. no IV1 main effect, no IV2 main effect, yes interaction

OK, so if you run a 2x2, any of these eight general patterns could occur in your data. That's a lot to keep track of. As you develop your skills in examining graphs that plot means, you should be able to look at the graph and visually guesstimate if there is, or is not, a main effect or interaction. You will need you inferential statistics to tell you for sure, but it is worth knowing how to know see the patterns. Look at the eight patterns in Figure 11.2.3.1.

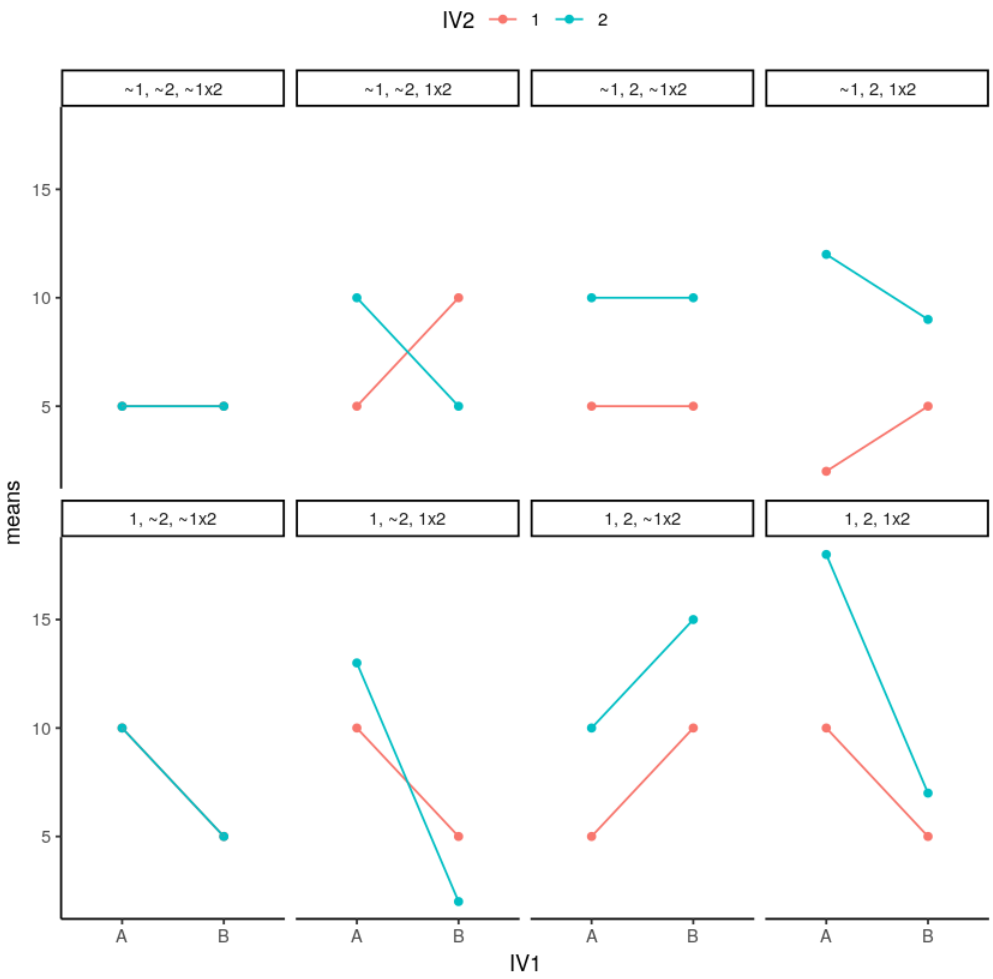


Figure **11.2.3.1**: Line graphs showing 8 possible general outcomes for a 2x2 design. (CC-BY-SA Matthew J. C. Crump from 10.1 of Answering Questions with Data)

Whenever the lines cross, or would cross if they kept going, you have a possibility of an interaction. Whenever the lines are parallel, there *can't* be an interaction. When *both* of the points on the A side are higher *or* lower than *both* of the points on the B side, then you have a main effect for IV1 (A vs B). Whenever the green line is above *or* below the red line, then you have a main effect for IV2 (1 vs. 2).

We know this is complicated. You should see what all the possibilities look like when we start adding more levels or more IVs! It gets nuts. Because of this nuttiness, it is often good practice to make your research designs simple (as few IVs and levels as possible to test your question). That way it will be easier to interpret your data. Whenever you see that someone ran a 4x3x7x2 design, your head should spin. It's just too complicated.

Do you remember Stephen Chew's video series on how to learn? Well, his second video (What Students Should Know about How People Learn) has a factorial design (with an extra control group).

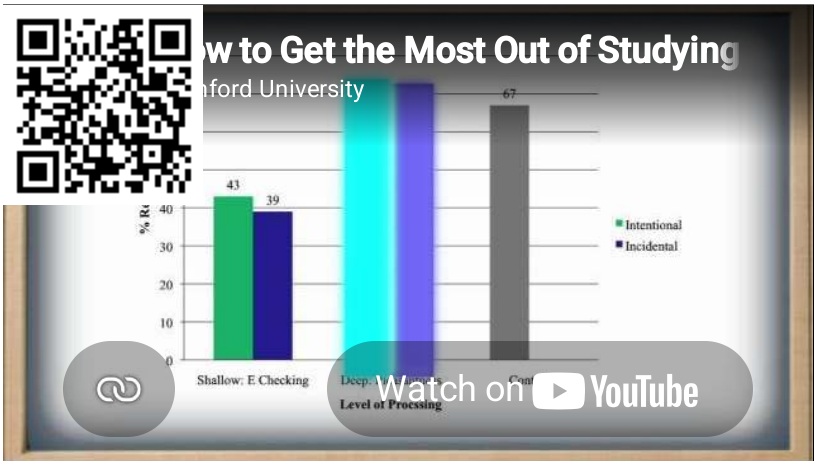


Figure 11.2.3.2: What Students Should Know about How People Learn. (Stephen Chew via YouTube)

Watching this video can help you understand factorial designs, including main effects and interactions), as well as learn about learning!

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)
- Dr. Stephen Chew's "What Students Should Know About How People Learn" (Part 2 of "How to Get the Most Out of Studying" series)

This page titled 11.2.3: Interpreting Main Effects and Interactions in Graphs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **10.1: Looking at main effects and interactions** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

11.2.4: Interpreting Interactions- Do Main Effects Matter?

The interpretation of main effects and interactions can get tricky. Consider the concept of a main effect. This is the idea that a particular IV has a consistent effect. For example, drinking five cups of coffee makes you more awake compared to not drinking five cups of coffee. The main effect of drinking five cups of coffee versus not drinking coffee will generally be true across the levels of other IVs in our life. For example, let's say you conducted an experiment testing whether the effect of drinking five cups of coffee versus not, changes depending on whether you are in your house or in a car. Perhaps the situation matters? No, probably not so much. You will probably still be more awake in your house, or your car, after having five cups of coffee, compared to if you hadn't.

The coffee example is a reasonably good example of a consistent main effect. Another silly kind of example might be the main effect of shoes on your height. For example, if your IV was wearing shoes or not, and your DV was height, then we could expect to find a main effect of wearing shoes on your measurement of height. When you wear shoes, you will become taller compared to when you don't wear shoes. Wearing shoes adds to your total height. In fact, it's hard to imagine how the effect of wearing shoes on your total height would ever interact with other kinds of variables. You will be always be that extra bit taller wearing shoes. Indeed, if there was another manipulation that could cause an interaction that would truly be strange. For example, imagine if the effect of being inside a bodega or outside a bodega interacted with the effect of wearing shoes on your height. That could mean that shoes make you taller when you are outside a bodega, but when you step inside, your shoes make you shorter…but, obviously this is just totally ridiculous. That's correct, it is often ridiculous to expect that one IV will have an influence on the effect of another, especially when there is no good reason.

The summary here is that it is convenient to think of main effects as a consistent influence of one manipulation. However, when an *interaction* is observed, this messes up the consistency of the main effect. That is the very definition of an interaction. It means that some main effect is *not* behaving consistently across different situations. Indeed, whenever we find an interaction, sometimes we can question whether or not there really is a general consistent effect of some manipulation, or instead whether that effect only happens in specific situations.

For this reason, you will often see that researchers report their findings this way:

"We found a main effect of X, BUT, this main effect was qualified by an interaction between X and Y".

Notice the big *BUT*. Why is it there? The sentence points out that before they talk about the main effect, they need to first talk about the interaction, which is making the main effect behave inconsistently. In other words, the interpretation of the main effect *depends* on the interaction, the two things have to be thought of together to make sense of them.

Here are some examples to help you make sense of these issues:

Consistent

A consistent main effect and an interaction means that the results of the interaction follows the results of the main effect. What's going on here? As you can see, the top line (Level 2 of IV2) goes down. The bottom line (Level 1 of IV2) goes up. IV1 is represented by the endpoints of the lines, with Level A of IV1 on the left and Level B of IV1 on the right.

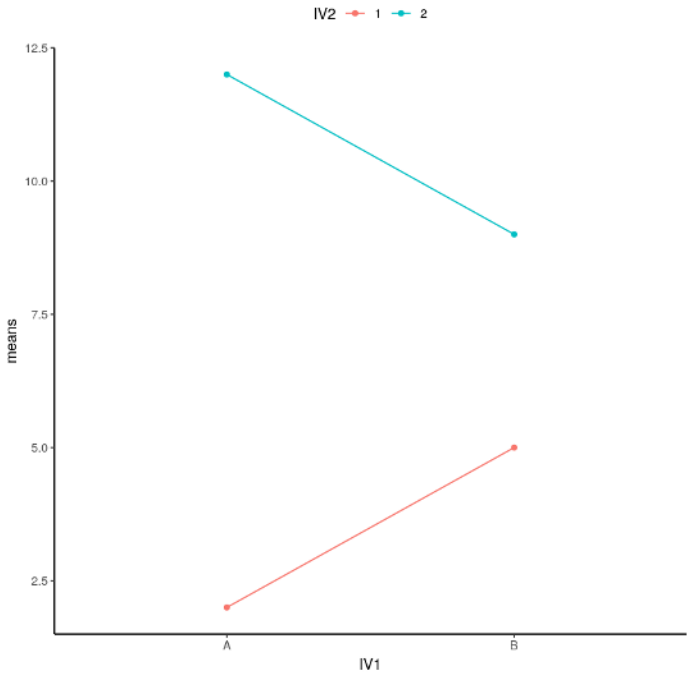


Figure **11.2.4.1**: Example means showing a generally consistent main effect along with an interaction. (CC-BY-SA Matthew J. C. Crump from 10.2 of Answering Questions with Data)

There is a main effect of IV2: the level 1 means (red line on the bottom) are both lower than the level 2 means (aqua line on the top). There is also an interaction. The size of the difference between the red and aqua points in the A condition (left) is bigger than the size of the difference in the B condition. This means that you couldn't just look at the main effect of IV2, you have to also take into account how IV1 affects IV2.

How would we interpret this?

We could say there WAS a main effect of IV2, BUT it was qualified by an IV1 x IV2 interaction.

What's the qualification?

The size of the IV2 effect changed as a function of the levels of IV1. It was big for level A, and small for level B of IV1.

What does the qualification mean for the main effect?

Well, first it means the main effect can be changed by the other IV. That's important to know. Does it also mean that the main effect is not a real main effect because there was an interaction? Not really, there is a generally consistent effect of IV2. The green points are above the red points in all cases. Whatever IV2 is doing, it seems to work in at least a couple situations, even if the other IV also causes some change to the influence.

Let's try an example to show that a consistent interaction, when the interaction is similar to one or both sets of main effects.

Table **11.2.4.1**- Consistent 2x2

IV Levels	**IV1 (Task Presentation): Paper**	**IV1 Task Presentation): Computer**
IV2 (Task Difficulty): Easy	80	90
IV2 (Task Difficulty): Hard	40	70
Marginal Means of IV1:	60	80

How would we interpret this?

As shown in Table **11.2.4.1**, there is a main effect for Task Presentation; overall, performance was better using computer than using paper. (There is also a main effect for Task Difficulty; overall, performance was better on the easier task, but the means are shown to make these examples easier to follow.).

What's the qualification?

The interaction shows that both Task Presentation and Task Difficulty affect the DV.

What does the qualification mean for the main effect?

The pattern of the main effect of Task Presentation is consistent for *both* easy tasks *and* for hard tasks.

Consistency amongst interactions and mains effect is the simplest relationship between a main effect and an interaction. Let's look at one that makes interpreting the main effects more difficult.

Inconsistent

An inconsistent main effect and an interaction is shown in Figure **11.2.4.2**. What's going on here? As you can see, the top line (Level 2 of IV2) goes down sharply. The bottom line (Level 1 of IV2) stays flat. IV1 is represented by the endpoints of the lines, with Level A of IV1 on the left and Level B of IV1 on the right.

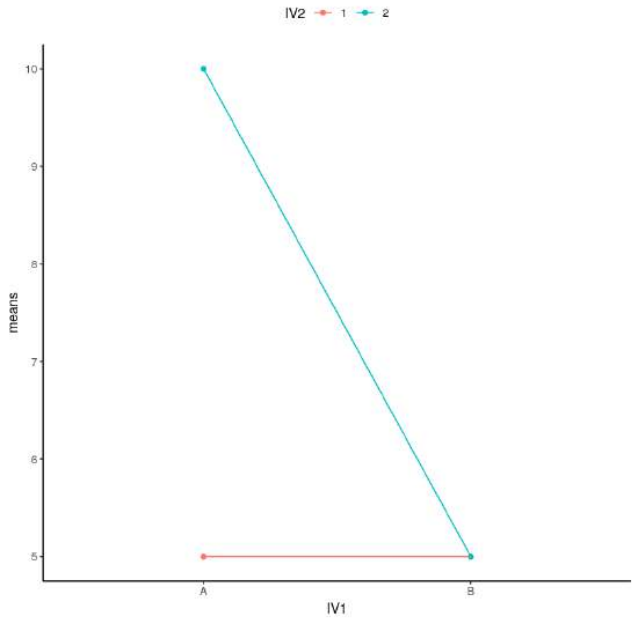


Figure **11.2.4.2**: Example data showing how an interaction exists, and a main effect does not, even though the means for the main effect may show a difference. (CC-BY-SA Matthew J. C. Crump from 10.2 of Answering Questions with Data)

You should see an interaction here straight away. The difference between the aqua and red points in condition A (left two dots) is huge, and there is no difference between them in condition B. Is there an interaction? Yes!

Are there any main effects here? With data like this, sometimes an ANOVA will suggest that you do have significant main effects. For example, what is the mean difference between level 1 and 2 of IV2? That is the average of the green points ((10+5)/2 = 15/2= 7.5) compared to the average of the red points (5). There will be a difference of 2.5 for the main effect (7.5 vs. 5).

Starting to see the issue here? From the perspective of the main effect (which collapses over everything and ignores the interaction), there is an overall effect of 2.5. In other words, level 2 adds 2.5 in general compared to level 1. However, we can see from the graph that IV2 does not do anything in general. It does not add 2.5s everywhere. It adds 5 in condition A, and nothing in condition B. It only does one thing in one condition.

What is happening here is that a “main effect” is produced by the process of averaging over a clear interaction.

How would we interpret this?

We might have to say there was a main effect of IV2, BUT we would definetely say it was qualified by an IV1 x IV2 interaction.

What's the qualification?

The size of the IV2 effect completely changes as a function of the levels of IV1. It was big for level A, and nonexistent for level B of IV1.

What does the qualification mean for the main effect?

In this case, we might doubt whether there is a main effect of IV2 at all. It could turn out that IV2 does not have a general influence over the DV all of the time, it may only do something in very specific circumstances, in combination with the presence of other factors.

Let's try that with some examples.

Table **11.2.4.2**- Inconsistent 2x2

IV Levels	**IV1 (Task Presentation): Paper**	**IV1 Task Presentation): Computer**
IV2 (Task Difficulty): Easy	90	90
IV2 (Task Difficulty): Hard	40	70
Marginal Means of IV1:	65	80

How would we interpret this?

Sometimes the interaction shows effects that the main effects miss or are opposite of the main effects. In Table **11.2.4.2**, there is a main effect for Task Presentation, overall performance was better using computer presentation than using paper presentation. (There is also a main effect for Task Difficulty; overall, performance was better on the easier task, but the means are shown to make these examples easier to follow.).

What's the qualification?

Example 11.2.4.1

In Table **11.2.4.2**, what did the main effect of Task Presentation miss?

Solution

Performance is better on the computer but *only for hard tasks* (not for easy tasks). Performance the same for Paper and Computer for Easy tasks.

What does the qualification mean for the main effect?

This qualification means that the interaction is inconsistent with the main effects; performance is affected by *both* Task Difficulty and Task Presentation.

Another form of inconsistency is when the main effects do not show a difference, but the IV levels do affect the DV.

Table **11.2.4.3**- Inconsistent 2x2 Showing Null Main Effect

IV Levels	**IV1 Task Presentation): Computer**	**IV1 (Task Presentation): Paper**
IV2 (Task Difficulty): Easy	90	70
IV2 (Task Difficulty): Hard	40	60
Marginal Means of IV1:	65	65

How would we interpret this?

There is no main effect for Task Presentation in Figure 3, overall performance was equivalent using computer presentation and using paper presentation.

What's the qualification?

This pattern is inconsistent with the interaction.

What does the qualification mean for the main effect?

In this example, Task Presentation does matter for easy tasks.

This is why interactions are so important to look at. Sometimes the main effects tell us most of what we need to know, but sometimes main effects hide what is really happening.

Try it Yourself

Let's look at the Mindset Quiz pretest scores when faculty in English and Psychology tried to improve their students' mindset. This is different from the Difference scores that we looked at earlier. This time, we are only looking at the posttests, the scores on the Mindset Quiz at the end of the semester. Review Table **11.2.4.4** and Figure **11.2.4.3**, and decide if you think that the main effects are consistent or inconsistent with the interaction.

Table **11.2.4.4** shows the marginal means and the cell means for our two IVs (Intervention: Yes or No; Department: English or Psychology).

Table **11.2.4.4**- Mindset Pretest Scores

IV Levels	**IV1: Yes Intervention**	**IV1: No Intervention**	Marginal Means for IV2:
IV2: English Department	41.15	41.43	41.29
IV2: Psychology Department	46.46	41.60	44.03
Margin Means for IV1:	43.8	41.51	(You could put the total mean here...)

How would we interpret this?

Let's first look at the main effects in Table **11.2.4.4**. Without doing a null hypothesis significance test, we might guess that there was a significant different between those who were going to receive the intervention and those who were not going to receive the intervention based on the means; remember, these are pretest scores so no one had received the intervention yet! So right off, that's an odd finding. Why would students already having a higher score on the Mindset Quiz? Do you think that the professor was subtly supporting growth mindset in only the classes that were going to have the intervention?

Now for the department. Without doing an null hypothesis significance testing, we might guess that there was a significant difference with the Psychology students having higher scores on the Mindset Quiz (slightly more growth mindset) than the English students. Again, this is surprising for a pretest. Why would Psychology students have a growth mindset more than English students?

On the other hand, we didn't conduct any null hypothesis significance testing, so maybe these means are actually close enough to say that they are similar. At this point, we just don't know.

But let's move on to the interaction. Looking at the four cells in the middle of the table, what do you see?

Example 11.2.4.2

Does one cell in the middle of Table **11.2.4.4** seem substantially higher? Or substantially lower?

Solution

It looks like the combination of IV levels strengthened the effect of the IVs alone. When discussing the main effects we said that it looks like the students who were going to get the intervention (IV1 = Yes) already scored higher on the Mindset Quiz, *and* the Psychology students scored higher on the pretest than English students. When looking at these in combination, it looks like the only the Psychology classes that were going to receive the intervention scored higher, not everyone who was going to

receive the intervention. This high average score on the Mindset Quiz pretest for Psychology students who were going to receive the intervention seems to have increased the marginal mean for both the Intervention IV and the Department IV.

When you look at Figure **11.2.4.3**, do you see these means? The top line represents those who experienced an intervention (IV1, Level: Yes), and rises sharply. The dashed bottom line represents those who did not experience an intervention (IV1, Level: No). This dashed line most stays flat. IV2 is represented by the endpoints of the lines, with the English Department on the left and the Psychology Department on the right. This means that the one group that looks really different are those in Psychology who experienced the intervention.

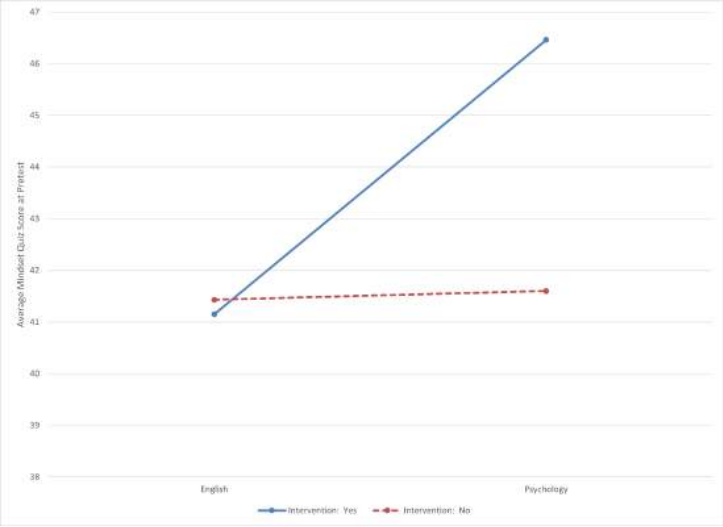


Figure **11.2.4.3**: Mindset Quiz Pretest Means. (CC-BY-SA; Michelle Oja)

Figure **11.2.4.3** makes it very clear that there is an interaction (the lines are not parallel). It also makes clear that the scores for English students weren't affected by the other IV (dots on the left), and that those who were not going to experience an intervention (dashed line) did not differ between the two departments. But, woah nelly, look at the blue dot representing students who were going to receive an intervention in a Psychology class! Based on the scaling of this chart, it looks like their mean was wildly higher than the other combinations.

What's the qualification?

The size of the effect of the Department (IV2) completely changes as a function of the levels of Intervention (IV1).

What does the qualification mean for the main effect?

It might mean that the main effect of Department doesn't really matter because the change in Mindset Quiz pretest scores only changes when in a Psychology class that will receive an intervention.

Consistent or Inconsistent?

Example 11.2.4.3

Are the main effects are consistent or inconsistent with the interaction?

Solution

Both main effects are consistent with the interaction.

Do main effects mean anything?

Example 11.2.4.4

When should you pay attention to mean differences in main effects?

Solution

When there are no interactions.

When there are no interactions, the main effects for each IV are like separate ANOVAs. An ANOVA with one IV. This is what makes interpreting main effects risky when there's an interaction. The interaction may be all that is happening, but based on the means of only one of the IVs, it looks like that IV is influencing the DV on its own.

Contributors and Attributions

Matthew J. C. Crump (Brooklyn College of CUNY)

Dr. MO (Taft College)

This page titled 11.2.4: Interpreting Interactions- Do Main Effects Matter? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **10.2: Interpreting main effects and interactions** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

11.2.5: Interpreting Beyond 2x2 in Graphs

Our graphs so far have focused on the simplest case for factorial designs, the 2x2 design, with two IVs, each with 2 levels. It is worth spending some time looking at a few more complicated designs and how to interpret them. We talked about more complicated designs in the Factorial Notations and Square Tables section, but here's a more focused approach to interpreting the graphs of these advanced designs.

2x3 design

In a 2x3 design there are two IVs. IV1 has two levels, and IV2 has three levels. Typically, there would be one DV. Let's talk about the main effects and interaction for this design.

First, let's make the design concrete. Let's imagine we are running a memory experiment. We give people some words to remember, and then test them to see how many they can correctly remember. Our DV is the proportion (percentage) that participants remembered correctly out of all tries. We know that people forget things over time. Our first IV will be time of test, immediate versus one week later. The time of test IV will produce a forgetting effect. Generally, people will have a higher proportion correct on an immediate test of their memory for things they just saw, compared to testing a week later.

We might be interested in manipulations that reduce the amount of forgetting that happens over the week. Let's make the second IV the number of time people got to study the items before the memory test, once, twice or three times. We call IV2 the repetition manipulation.

We might expect data that looks like Figure 11.2.5.1. The top line shows the means when there is no delay (Immediate) for the three levels of repetition. Notice that the proportion correct (y-axis) increases for the Immediate group with each repetition. The bottom ling shows the One Week Delay group over the three levels of repetition. Again, more repetition seems to increase the proportion correct. Let's talk about this graph in terms of main effects and interaction.

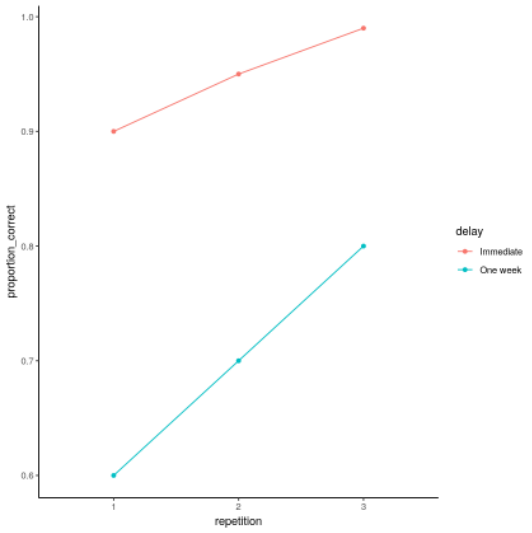


Figure 11.2.5.1: Example means for a 2x3 factorial design. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

First, the main effect of delay (time of test) is shown by in each differently-colored line, and seems obvious; the red line is on the top, way above the aqua line. This tells us that the proportion correct on the memory test is always higher when the memory test is taken immediately compared to after one week.

Second, the main effect of repetition is presented on the x-axis, and seems to be clearly present. The more times people saw the items in the memory test (once, twice, or three times), the more they remembered, as measured by increasingly higher proportion correct as a function of number of repetitions.

Is there an interaction? Yes, there is. Remember, an interaction occurs when the effect of one IV depends on the levels of an another. The delay IV measures the forgetting effect. Does the size of the forgetting effet change across the levels of the repetition variable? Yes it does. With one repetition the forgetting effect is 0.9 - 0.6 = 0.4. With two repetitions, the forgetting effect is a little

bit smaller, and with three, the repetition is even smaller still. So, the size of the forgetting effect changes as a function of the levels of the repetition IV. There is evidence in the means for an interaction. You would have to conduct an inferential test on the interaction term to see if these differences were likely or unlikely to be due to sampling error.

If there was no interaction, and say, no main effect of repetition, we would see something like Figure **11.2.5.2**.

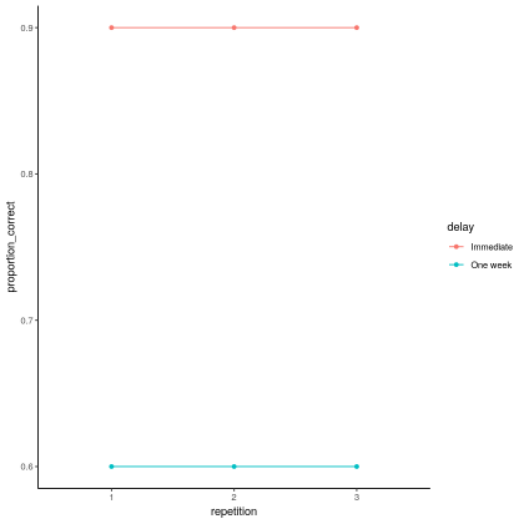


Figure **11.2.5.2**: Example means for a 2x3 design when there is only one main effect. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

What would you say about the interaction if you saw something like Figure **11.2.5.3**? The Immediate group is high, but repetition doesn't seem to matter. The One Week Delay group is flat until the third repetition, then increases the proportion correct.

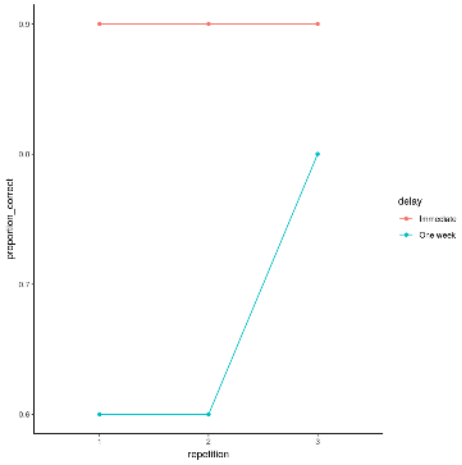


Figure **11.2.5.3**: Example means for a 2x3 design showing another pattern that produces an interaction. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

There is evidence in the means for an interaction. Remember, we are measuring the forgetting effect (effect of delay) three times. The forgetting effect is the same for repetition condition 1 and 2, but it is much smaller for repetition condition 3. The size of the forgetting effect depends on the levels of the repetition IV, so here again there is an interaction.

2x2x2 designs

Let's take it up a notch and look at a 2x2x2 design. Here, there are three IVs with 2 levels each. There are three main effects, three two-way (2x2) interactions, and one 3-way (2x2x2) interaction.

We will use the same example as before but add an additional manipualtion of the kind of material that is to be remembered. For example, we could present words during an encoding phase either visually or spoken (auditory) over headphones.

Figure **11.2.5.4** shows two pairs of lines, one side (the panel on the left) is for the auditory information to be remembered, and the panel on the right is when the information was presented visually. The lines still show the delay, and the y-axis still shows the

number of repetitions. In this version of the study, the was only two repetitions levels: once or twice. The two lines on the left show auditory IV levels and the two lines on the right show visual information. The top lines show when there's no delay, and the diagonal lines show when there is a week delay. There are only two levels of repetition, so there are only two dots representing this IV (1 repetition on the right and 2 repetitions on the left for both auditory and visual information).

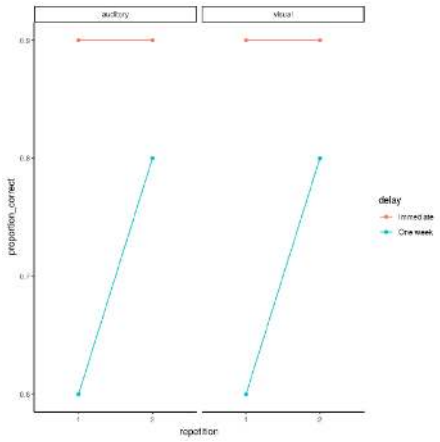


Figure **11.2.5.4**: Example means from a 2x2x2 design with no three-way interaction. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

You can think of the 2x2x2, as two 2x2s, one for auditory and one for visual. What's the take home from this example data? We can see that the graphs for auditory and visual are the same. They both show a 2x2 interaction between delay and repetition. People forgot more things across the week when they studied the material once, compared to when they studied the material twice. There is a main effect of delay, there is a main effect of repetition, but there is no main effect of modality (no difference between auditory or visual information), and there is not a three-way interaction.

What is a three-way interaction anyway? That would occur if there was a difference between the 2x2 interactions. For example, consider the next pattern of results (Figure **11.2.5.5**).

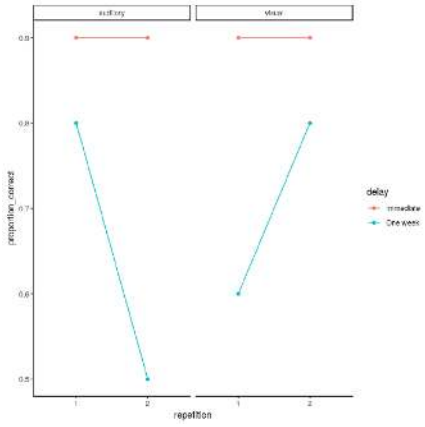


Figure **11.2.5.5**: Example means from a 2x2x2 design with a three-way interaction. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

We are looking at a 3-way interaction between modality, repetition and delay in Figure **11.2.5.5**. What is going on here? These results would be very strange, but here is an interpretation.

> For auditory stimuli, we see that there is a small forgetting effect when people studied things once, but the forgetting effect gets bigger if they studies things twice. A pattern like this would generally be very strange, usually people would do better if they got to review the material twice.
>
> The visual stimuli show a *different pattern*. This different pattern is where we get the three-way interaction. Here, the forgetting effect is large when studying visual things once, and it get's smaller when studying visual things twice.
>
> We see that there is an interaction between delay (the forgetting effect) and repetition for the auditory stimuli; BUT, this interaction effect is *different* from the interaction effect we see for the visual stimuli. The 2x2 interaction for the auditory

stimuli is *different* from the 2x2 interaction for the visual stimuli. In other words, *there is an interaction between the two interactions*, as a result there is a three-way interaction, called a 2x2x2 interaction.

We will note a general pattern here. Imagine you had a 2x2x2x2 design. That would have a 4-way interaction. What would that mean? It would mean that the pattern of the 2x2x2 interaction changes across the levels of the 4th IV. If two three-way interactions are different, then there is a four-way interaction.

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 11.2.5: Interpreting Beyond 2x2 in Graphs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **10.4: More complicated designs** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

11.3: Two-Way ANOVA Summary Table

Factorial designs are designs, not a statistical analysis. Research designs are about how data are collected, instead of how data are analyzed. However, factorial designs are often analyzed with ANOVAs, so we will walk through ANOVA Summary Tables for factorial designs with at least two IVs with at least two levels each ("Two-way" means that there are two variables, each with two or more levels.). The main point of factorials designs is how the different IV's and their levels are combined, which is a research design issue.

ANOVA Summary Table Refresher

Remember the ANOVA Summary Table for Between Groups designs (when the groups were independent) shown in Table **11.3.1**?

Table **11.3.1**: ANOVA Table

Source	SS	df	MS	F
Between Groups	SS_B	$k-1$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_W}$
Within Groups (Error)	SS_W	$N-k$	$\frac{SS_W}{df_W}$	N/A
Total	SS_T	$N-1$	N/A	N/A

Once you had the Sum of Squares (SS) and the Degrees of Freedom (df), you could easily calculate the Mean Square (MS), and then the final F.

What was added for the Repeated Measures ANOVA Summary Table (example shown in Table **11.3.2**)?

Table **11.3.2**- RM ANOVA Summary Table with Formulas for Degrees of Freedom

Source	SS	df	MS	F
Between Groups	formula elsewhere	$k-1$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_W}$
Participants	formula elsewhere	$P-1$	N/A	N/A
Within Groups (Error)	$SS_{WG}=SS_T-SS_{BG}-SS_P$	$(k-1)\times(P-1)$	$\frac{SS_W}{df_W}$	N/A
Total	formula elsewhere	$N-1$	N/A	N/A

Right, we added one more row to account for the variation within each participant. But just like the BG ANOVA Summary Table, once you had all of the Sums of Squares (SS) and the Degrees of Freedom (df), you could easily calculate the Mean Square (MS), and then the final F.

The good news is that the Two-Way ANOVA Summary Table, the ANOVA Summary Table used when your study has two IVs, has similar properties. The bad news is that, just like the BG and RM ANOVA Summary Tables, the SS and df have some new formulas.

In sum, whatever the type of ANOVA you have, your general process will be:

1. Calculate the degrees of freedom.
2. Calculate the Sum of Squares.
3. Use SS to calculate the Mean Square (SS/df).
4. Calculate the F as a ratio of the MS_{BG} divided by the MS_{WG} (or MS_{Error}, they are the same thing).
    1. The only difference with a two-way ANOVA is that you have two Between Group MS's, one for each variable, and a row for the interaction.
5. Locate a critical value.
6. Determine whether you retain or reject the null hypothesis.

What would that kind of ANOVA Summary Table look like? Let's look at one in Table **11.3.3**:

Table **11.3.3**- Two-Way ANOVA Summary Table

Source	SS	df	MS	F
IV 1				
IV 2				
Interaction				
Within Groups (Error)				
Total				

Sum of Squares and Degrees of Freedom

With independent sample Between Groups ANOVAs, we had three Sums of Squares and three degrees of freedom.

? Exercise 11.3.1

What were the names of the three rows called? In other words, what were the names of the three Sums of Squares?

Answer

1. Between Groups
2. Within-Groups (or Error)
3. Total

With dependent sample Repeated Measures ANOVAs, we needed four Sums of Squares and four degrees of freedom.

? Exercise 11.3.2

What were the names of the four rows called? In other words, what were the names of the four Sums of Squares?

Answer

1. Between Groups
2. Participants
3. Within-Groups (or Error)
4. Total

With two-way ANOVAs you need six Sums of Squares and six degrees of freedom! If you look at Table 3, though, you'll only see five rows. What's the deal? There's an extra calculation called "cells" that allows you to fill in the ANOVA Summary. So, the Sums of Squares and Degrees of Freedom that are need are:

1. Cells: This is the SS between all of the groups (This is not listed on the ANOVA Summary Table, but is used to calculate some of the others.)
2. Between group for one variable (IV_1)
3. Between group for the other variable (IV_2)
4. Interaction
5. Within group
6. Total

The next major section shows how to calculate the Sums of Squares using tables, but it is exceedingly unlikely that you would every calculate a factorial ANOVA by hand. This page is unmodified from the original source (Dr. Crump's Answering Questions with Data), and uses a statistical analysis software call R. It's so unlikely that you'll every do this by hand that the page is not updated or formatted for this textbook. So let's look at how to calculate each Degree of Freedom.

Formulas for Degrees of Freedom

The following are the formulas for each Degree of Freedom.

1. Cells: $(k_1 \times k_2)–1$
    1. Remembering that "k" is the number of groups, k_1 is the number of levels of the first IV and , k_2 is the number of levels of the other IV.
2. Between group for one variable ($IV1_1$): $k_1–1$
3. Between group for the other variable (IV_2): $k_2)–1$
4. Interaction: $df_1 \times df_2$
5. Within group: $df_{Total}–df_{Cells}$
6. Total: $N-1$
    1. With N being the number of scores.

Dr. MO has been skating around the issue, but a two-factor another can be Between Groups (with all IV levels being independent) or Repeated Measures (with all IV levels being repeated or related) or a *combination* (some IVs are independent and some IVs are repeated). To make things easier in our calculations, we'll just use the Two-Way ANOVA Summary Table as if all of the IVs are independent, even though it's actually pretty common to have a combination of IVs. In any case, use the N of the number of scores, which is usually the number of participants.

When we have an example, we'll replace the subscript numbers with subscript letters that represent the name of the IVs.

Mean Square

In two-way ANOVAs, you need to four Mean Squares.

Formulas for Mean Square

Regardless of which Mean Square you are calculating, it will be the same as all of the Mean Squares that we've covered in the last couple chapters:

$$\text{MS} = \frac{SS}{df}$$

You use the Sum of Squares of the row that you are working with, and divide if by the Degrees of Freedom of the row that you are working with to get the Mean Square for that row.

Calculated F

In two-way ANOVAs, you end up with three calculated F's, one for each variable and one for the interaction. For factorial designs with more IVs, you would end with more calculated F-scores.

Formulas for Calculated F-Scores

Each F-score is calculated the same was as the F in a Between Groups ANOVA Summary Table or a Repeated Measures ANOVA Summary Table: MS_{IV} / $MS_{\text{WG or Error}}$

For a two-way ANOVA, that looks like:

? Exercise 11.3.3

What is similar to past F-score calculations? What is similar amongst all three of these calculations?

Answer

As with prior F calculations, the ratio is the Mean Square of the variable that we're looking and the Mean Square of the within-groups error term. That is what is similar amongst all three of these calculations, too.

Each calculated F will be compared to a different critical value based on the Degrees of Freedom of the numerator and the denominator.

Two-Way ANOVA Summary Table with Formulas

Here's the ANOVA Summary Table for a two-way factorial design with all of the formulas included, other than the Sum of Squares. The formulas for the Sum of Squares aren't provided because they are beyond the scope of this introduction to statistics (and you'd never do them by hand, anyway). Note that there will again be three blank cells, labeled "N/A" in Table **11.3.4**.

Table **11.3.4**- Two-Way ANOVA Summary Table

Source	SS	df	MS	F
IV 1	formula elsewhere	$k_1 - 1$	$\frac{SS_1}{df_1}$	$\frac{MS_1}{MS_{WG}}$
IV 2	formula elsewhere	$k_2 - 1$	$\frac{SS_2}{df_2}$	$\frac{MS_2}{MS_{WG}}$
Interaction	formula elsewhere	$df_1 * df_2$	$\frac{SS_{INT}}{df_{INT}}$	$\frac{MS_{Interaction}}{MS_{WG}}$
Within Groups (Error)	formula elsewhere	$df_{Total} - df_{cells}$	$\frac{SS_{WG}}{df_{WG}}$	N/A
Total	formula elsewhere	$N - 1$	N/A	N/A

To complete the table, you need to need to calculate the degrees of freedom of cells ($df_{cells} = (k_1 \times k_2)$–1) although there's no place in the table to include this information.

Normally, we put the variable with fewer levels before (above) the variable with more levels in the notation and in the ANOVA Summary Table.

Practice with a Two-Way ANOVA Summary Table

Example 11.3.1

Example 1 Complete the following ANOVA Summary Table with the information provided. The Sums of Squares are already in Table **11.3.5**.

Participants (N = 48) joined a weight-loss program designed to increase the time people exercised. The program lasted either one month, two months, or three months. At the beginning and end of their program, the participants measured their weight to see how many pounds they lost. Because hormones affect weight loss, the gender of each participant was recorded and used as a variable in this 2x3 factorial design.

8 men & 8 women who exercised 1 month

8 men & 8 women who exercised 2 months

8 men & 8 women who exercised 3 months

Solution

To complete the calculations you'll find:

1. The Mean Square for one IV
2. The Mean Square for the other IV
3. The Mean Square for the interaction of the two groups.
4. The Mean Square for the Error (the denominator of the calculated F-score), which is sometimes called Mean Square for Within-Groups.
5. The F-ratio for one group
6. The F-ratio for the other group
7. The F-ratio for the interaction of the two groups.
8. F for one group: $= \frac{MS_1}{MS_{WG}}$
9. F for other group: $= \frac{MS_2}{MS_{WG}}$

10. F for the interaction: $= \frac{MS_{Interaction}}{MS_{WG}}$

Table **11.3.5**- Two-Way ANOVA Summary Table

Source	SS	df	MS	F
IV1 (Gender)	330.75	$k_1 - 1$ = 2 - 1 = 1	$= \frac{SS_1}{df_1} = \frac{330.75}{1} = 330.75$	$= \frac{MS_1}{MS_{WG}} = \frac{330.75}{14.79} = 22.36$
IV2 (Length of Program: 1, 2, or 3 months)	1065.50	$k_2 - 1$ = 3 - 1 = 2	$= \frac{SS_2}{df_2} = \frac{1065.50}{2} = 532.75$	$= \frac{MS_2}{MS_{WG}} = \frac{532.75}{14.79} = 36.02$
Interaction	350.00	$df_1 * df_2$ = 1 * 2 = 2	$= \frac{SS_{INT}}{df_{INT}} = \frac{350.00}{2} = 175.00$	$= \frac{MS_{Interaction}}{MS_{WG}} = \frac{175.00}{14.79} = 11.84$
Within Groups (Error)	621.00	$df_{Total} - df_{cells}$ = 47 - 5 = 42	$= \frac{SS_{WG}}{df_{WG}} = \frac{621.00}{42} = 14.79$	leave blank
Total	2367.25	$N - 1 = 48 - 1 = 47$	leave blank	leave blank

$(df_{cells} = (k_1 \times k_2) – 1) = (2 \times 3) – 1 = 6 – 1 = 5$

As before, we can do a Calculation Check to see if we did all of the Degrees of Freedom correctly because they all should add up to the total again!

The rest of the process is the same as we've been doing, we just have to do it for each of the calculated F-scores: compare each of the three calculated F-scores to three critical values from the Critical Values of F Table to decide if we retain or reject the null hypothesis.

Like in the Between Groups ANOVA and the Repeated Measures ANOVA, the null hypothesis says that *all* of the means are similar. For factorial designs, this includes the interaction; the null hypothesis is saying that we will find no main effects, and that the means for each combination of IV levels will also be similar. Just for fun, let's look at what the combination of IVs would look like in a Punnett's Square.

Example 11.3.1

Complete a grid showing the factorial design IVs and DVS for the scenario in Example **11.3.1**.

Solution

Add text here.

Table **11.3.6**- Factorial Design Grid

IV Levels	**IV1 (Gender): Men**	**IV1 (Gender: Women**
IV2 (Time): 1 Month	**2x3: Men who exercised for one month**	**2x3: Women who exercised for one month**
IV2 (Time): 2 Months	2x3: Men who exercised for two months	2x3: Women who exercised for two months
IV2 (Time): 3 Months	2x3: Men who exercised for three months	2x3: Women who exercised for three months

What would we do next?

For each calculated F, we would determine if the null hypothesis was retained or not. Let's do that!

Example 11.3.1

Find the critical values for each of the calculated F-scores in the Table of Critical F-Scores (which can be found in the Common Critical Value Tables page at the end of the book), then determine if you should retain or reject the null hypothesis.

Solution

For the first IV of gender, the calculated F-score was 22.36. The critical value from the Critical Values of F Table the using the Degrees of Freedom of Gender and the Within-Groups (Error) (Gender df = 1, Within-Groups df = 42) is 4.08 (using the df for the denominator of 40 because that is the closest to 42 or 42 rounding down on the table). Because the following is still true, we reject the null hypothesis and say that the average pounds lost differed between men and women.

$$\textbf{(Critical} < \textbf{Calculated)} = \textbf{Reject null} = \textbf{At least one mean is different from at least one other mean.} = p < .05$$

$$\textbf{(Critical} > \textbf{Calculated)} = \textbf{Retain null} = \textbf{All of the means are similar.} = p > .05$$

For the second IV of length of the exercise program, the calculated F-score was 36.02. The critical value from the Critical Values of F Table the using the Degrees of Freedom of Gender and the Within-Groups (Error) (Exercise Length df = 2, Within-Groups df = 42) is 3.23 (using the df for the denominator of 40 because that is the closest to 42 or 42 rounding down on the table). Because the critical value is smaller than the calculated value, we reject the null hypothesis and say that the average pounds lost differed between at least two of the three groups (1 month, 2 months or 3 months.

For the interaction between gender and length of exercise program, the calculated F-score was 11.83. The critical value from the Critical Values of F Table the using the Degrees of Freedom of Gender and the Within-Groups (Error) (Interaction df = 2, Within-Groups df = 42) is 3.23 (using the df for the denominator of 40 because that is the closest to 42 or 42 rounding down on the table). Because the critical value is smaller than the calculated value, we reject the null hypothesis and say that the average pounds lost differed between at least two of the combinations of gender and exercise length.

What about Pairwise Comparisons?

Is your next step pairwise comparisons? Maybe!

Had we retained any of the null hypotheses, we would not have conducted pairwise comparisons for that variable. Just like in the other ANOVAs, if we don't find a differences between the means, we don't look for where the difference might be. In this case, though, we could have retained one variable (say, the Interaction), but still rejected the null hypotheses for each IV. In that case, we would have needed to conduct some pairwise comparisons. But maybe not as many as you'd think!

If the IV only has two levels and the F-score results in rejecting the null hypothesis, then pairwise comparisons are still not needed. The ANOVA is enough. The ANOVA says that at least two means are different from each other, and when we only have two means, we have all we need to know. So, with our Gender IV, we would not need to conduct pairwise comparisons even though we rejected the null hypothesis.

If the IV has three or more levels and the F-score results in rejecting the null hypothesis, then we would need to conduct pairwise comparisons to find which means are different from each other. This statistically significant main effect shows that at least two means are different from each other, but we need to figure out which specific means.

If the F-score results in rejecting the null hypothesis for the Interaction, then we would need to conduct pairwise comparisons to find which means are different from each other. This statistically significant interaction shows that at least two means are different from each other, but we'll have at minimum four means (a 2x2 factorial design results in four combinations) so we need to figure out which means differ from each other.

Phew, that was a lot! Next up is looking at calculating Sums of Squares for a two-way ANOVA. If you are never going to do that by hand, skip that section and start with section on Post-Hoc Pairwise Comparisons.

This page titled 11.3: Two-Way ANOVA Summary Table is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

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

11.3.1: Calculating Sum of Squares for the Factorial ANOVA Summary Table

You must be wondering how to calculate a 2x2 ANOVA. We haven't discussed this yet. We've only shown you that you don't have to do it when the design is a 2x2 repeated measures design (note this is a special case).

We are now going to work through some examples of calculating the ANOVA table for 2x2 designs. We will start with the between-subjects ANOVA for 2x2 designs. We do essentially the same thing that we did before (in the other ANOVAs), and the only new thing is to show how to compute the interaction effect.

Remember the logic of the ANOVA is to partition the variance into different parts. The SS formula for the between-subjects 2x2 ANOVA looks like this:

$$SS_{\text{Total}} = SS_{\text{Effect IV1}} + SS_{\text{Effect IV2}} + SS_{\text{Effect IV1xIV2}} + SS_{\text{Error}}$$

In the following sections we use tables to show the calculation of each SS. We use the same example as before with the exception that **we are turning this into a between-subjects design**. There are now 5 different subjects in each condition, for a total of 20 subjects. As a result, we remove the subjects column.

SS Total

We calculate the grand mean (mean of all of the score). Then, we calculate the differences between each score and the grand mean. We square the difference scores, and sum them up. That is SS_{Total}, reported in the bottom yellow row.

	All Conditions				Difference from Grand Mean				Squared Differences			
	No Reward		Reward		No Reward		Reward		No Reward		Reward	
	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction
	A	B	C	D	A-GrandM	B-GrandM	C-GrandM	D-GrandM	(A-GrandM)^2	(B-GrandM)^2	(C-GrandM)^2	(D-GrandM)^2
	10	5	12	9	1.05	-3.95	3.05	0.05	1.1025	15.6025	9.3025	0.0025
	8	4	13	8	-0.95	-4.95	4.05	-0.95	0.9025	24.5025	16.4025	0.9025
	11	3	14	10	2.05	-5.95	5.05	1.05	4.2025	35.4025	25.5025	1.1025
	9	4	11	11	0.05	-4.95	2.05	2.05	0.0025	24.5025	4.2025	4.2025
	10	2	13	12	1.05	-6.95	4.05	3.05	1.1025	48.3025	16.4025	9.3025
Means	9.6	3.6	12.6	10								
Grand Mean	8.95											
sums								Sums	7.3125	148.3125	71.8125	15.5125
SS Total								SS Total	242.95			

SS Distraction

We need to compute the SS for the main effect for distraction. We calculate the grand mean (mean of all of the scores). Then, we calculate the means for the two distraction conditions. Then we treat each score as if it was the mean for it's respective distraction condition. We find the differences between each distraction condition mean and the grand mean. Then we square the differences and sum them up. That is $SS_{\text{Distraction}}$, reported in the bottom yellow row.

	All Conditions				Distraction Mean - GM				Squared Differences			
	No Reward		Reward		No Reward		Reward		No Reward		Reward	
	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction
	A	B	C	D	NDM-GM A	DM-GM B	NDM-GM C	DM-GM D	(NDM-GM)^2 A	(DM-GM)^2 B	(NDM-GM)^2 C	(DM-GM)^2 D
	10	5	12	9	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
	8	4	13	8	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
	11	3	14	10	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
	9	4	11	11	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
	10	2	13	12	2.15	-2.15	2.15	-2.15	4.6225	4.6225	4.6225	4.6225
Means	9.6	3.6	12.6	10								
Grand Mean	8.95	No Distraction	11.1	Distraction	6.8							
sums								Sums	23.1125	23.1125	23.1125	23.1125
SS Distraction								SS Distraction	92.45			

These tables are a lot to look at! Notice here, that we first found the grand mean (8.95). Then we found the mean for all the scores in the no-distraction condition (columns A and C), that was 11.1. All of the difference scores for the no-distraction condition are 11.1-8.95 = 2.15. We also found the mean for the scores in the distraction condition (columns B and D), that was 6.8. So, all of the difference scores are 6.8-8.95 = -2.15. Remember, means are the balancing point in the data, this is why the difference scores are +2.15 and -2.15. The grand mean 8.95 is in between the two condition means (11.1 and 6.8), by a difference of 2.15.

SS Reward

We need to compute the SS for the main effect for reward. We calculate the grand mean (mean of all of the scores). Then, we calculate the means for the two reward conditions. Then we treat each score as if it was the mean for it's respective reward condition. We find the differences between each reward condition mean and the grand mean. Then we square the differences and sum them up. That is SS_{Reward}, reported in the bottom yellow row.

	All Conditions				Reward Mean - GM				Squared Differences			
	No Reward		Reward		No Reward		Reward		No Reward		Reward	
	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction
	A	B	C	D	NRM-GM A	NRM-GM B	RM-GM C	RM-GM D	(NRM-GM)^2 A	(NRM-GM)^2 B	(RM-GM)^2 C	(RM-GM)^2 D
	10	5	12	9	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
	8	4	13	8	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
	11	3	14	10	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
	9	4	11	11	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
	10	2	13	12	-2.35	-2.35	2.35	2.35	5.5225	5.5225	5.5225	5.5225
Means	9.6	3.6	12.6	10								
Grand Mean	8.95	No Reward	6.6	Reward	11.3							
sums								Sums	27.6125	27.6125	27.6125	27.6125
SS Reward								SS Reward	110.45			

Now we treat each no-reward score as the mean for the no-reward condition (6.6), and subtract it from the grand mean (8.95), to get -2.35. Then, we treat each reward score as the mean for the reward condition (11.3), and subtract it from the grand mean (8.95), to get +2.35. Then we square the differences and sum them up.

SS Distraction by Reward

We need to compute the SS for the interaction effect between distraction and reward. This is the new thing that we do in an ANOVA with more than one IV. How do we calculate the variation explained by the interaction?

The heart of the question is something like this. Do the individual means for each of the four conditions do something a little bit different than the group means for both of the independent variables.

For example, consider the overall mean for all of the scores in the no reward group, we found that to be 6.6 Now, was the mean for each no-reward group in the whole design a 6.6? For example, in the no-distraction group, was the mean for column A (the no-reward condition in that group) also 6.6? The answer is no, it was 9.6. How about the distraction group? Was the mean for the reward condition in the distraction group (column B) 6.6? No, it was 3.6. The mean of 9.6 and 3.6 is 6.6. If there was no hint of an interaction, we would expect that the means for the reward condition in both levels of the distraction group would be the same, they would both be 6.6. However, when there is an interaction, the means for the reward group will depend on the levels of the group from another IV. In this case, it looks like there is an interaction because the means are different from 6.6, they are 9.6 and 3.6 for the no-distraction and distraction conditions. This is extra-variance that is not explained by the mean for the reward condition. We want to capture this extra variance and sum it up. Then we will have measure of the portion of the variance that is due to the interaction between the reward and distraction conditions.

What we will do is this. We will find the four condition means. Then we will see how much additional variation they explain beyond the group means for reward and distraction. To do this we treat each score as the condition mean for that score. Then we subtract the mean for the distraction group, and the mean for the reward group, and then we add the grand mean. This gives us the unique variation that is due to the interaction. We could also say that we are subtracting each condition mean from the grand mean, and then adding back in the distraction mean and the reward mean, that would amount to the same thing, and perhaps make more sense.

Here is a formula to describe the process for each score:

$$\bar{X}_{\text{condition}} - \bar{X}_{\text{IV1}} - \bar{X}_{\text{IV2}} + \bar{X}_{\text{Grand Mean}}$$

Or we could write it this way:

$$\bar{X}_{\text{condition}} - \bar{X}_{\text{Grand Mean}} + \bar{X}_{\text{IV1}} + \bar{X}_{\text{IV2}}$$

When you look at the following table, we apply this formula to the calculation of each of the differences scores. We then square the difference scores, and sum them up to get $SS_{\text{Interaction}}$, which is reported in the bottom yellow row.

	All Conditions				Interaction Differences				Squared Differences			
	No Reward		Reward		No Reward		Reward		No Reward		Reward	
	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction	No Distraction	Distraction
	A	B	C	D	A-ND-NR+GM	B-D-NR+GM	C-ND-R+GM	D-D-R+GM	(A-ND-NR+GM)^2 A	(B-D-NR+GM)^2 B	(C-ND-R+GM)^2 C	(D-D-R+GM)^2 D
	10	5	12	9	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
	8	4	13	8	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
	11	3	14	10	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
	9	4	11	11	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
	10	2	13	12	0.85	-0.85	-0.85	0.85	0.7225	0.7225	0.7225	0.7225
Means	9.6	3.6	12.6	10								
Grand Mean	8.95											
sums								Sums	3.6125	3.6125	3.6125	3.6125
SS Interaction								SS Interaction	14.45			

SS Error

The last thing we need to find is the SS Error. We can solve for that because we found everything else in this formula:

$$SS_{\text{Total}} = SS_{\text{Effect IV1}} + SS_{\text{Effect IV2}} + SS_{\text{Effect IV1xIV2}} + SS_{\text{Error}}$$

Even though this textbook meant to explain things in a step by step way, we guess you are tired from watching us work out the 2x2 ANOVA by hand. You and me both, making these tables was a lot of work. We have already shown you how to compute the SS for error before, so we will not do the full example here. Instead, we solve for SS Error using the numbers we have already obtained.

$$\begin{aligned} SS_{\text{Error}} &= SS_{\text{Total}} - SS_{\text{Effect IV1}} - SS_{\text{Effect IV2}} - SS_{\text{Effect IV1xIV2}} \\ &= 242.95 - 92.45 - 110.45 - 14.45 \\ &= 25.6 \end{aligned}$$

Check your work

We are going to skip the part where we divide the SSes by their dfs to find the MSEs so that we can compute the three F-values. Instead, if we have done the calculations of the SSes correctly, they should be same as what we would get if we used R to calculate the SSes. Let's make R do the work, and then compare to check our work.

```
library(xtable)
A <- c(10,8,11,9,10)  #nD_nR
B  <- c(5,4,3,4,2)  #D_nR
C <- c(12,13,14,11,13)  #nD_R
D  <- c(9,8,10,11,12)  #D_R
```

```
Number_spotted <- c(A, B, C, D)
Distraction    <- rep(rep(c("No Distraction", "Distraction"), each=5),2)
Reward         <- rep(c("No Reward","Reward"),each=10)
Distraction <- as.factor(Distraction)
Reward      <- as.factor(Reward)
all_df <- data.frame(Distraction, Reward, Number_spotted)
aov_summary <- summary(aov(Number_spotted~Distraction*Reward, all_df))
knitr::kable(xtable(aov_summary))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Distraction	1	92.45	92.45	57.78125	F)" style="vertical-align:middle;">0.000011
Reward	1	110.45	110.45	69.03125	F)" style="vertical-align:middle;">0.000003
Distraction:Reward	1	14.45	14.45	9.03125	F)" style="vertical-align:middle;">0.008379
Residuals	16	25.60	1.60	NA	F)" style="vertical-align:middle;">NA

A quick look through the column `Sum Sq` shows that we did our work by hand correctly. Congratulations to us! Note, this is not the same results as we had before with the repeated measures ANOVA. We conducted a between-subjects design, so we did not get to further partition the SS error into a part due to subject variation and a left-over part. We also gained degrees of freedom in the error term. It turns out with this specific set of data, we find p-values of less than 0.05 for all effects (main effects and the interaction, which was not less than 0.05 using the same data, but treating it as a repeated-measures design)

This page titled 11.3.1: Calculating Sum of Squares for the Factorial ANOVA Summary Table is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **9.6: 2x2 Between-subjects ANOVA** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

11.4: When Should You Conduct Post-Hoc Pairwise Comparisons?

The pairwise comparison calculations for a factorial design are the same as any pairwise comparison after any significant ANOVA. Instead of reviewing them here (because you can review them in the prior two chapters), we are going to discuss when (and why) you would or would not conduct pairwise comparisons in a factorial design.

Long Answer

We'll start by refreshing our member on what we've done before. Let's start with t-test, from oh-so long ago!

? Exercise 11.4.1

Did we conduct pairwise comparisons when we *retained* the null hypothesis when comparing two groups with a t-test? Why or why not?

Answer

We did not conduct pairwise comparisons when the null hypothesis was retained with a t-test. We didn't need to find which means were difference because the null hypothesis (which we retained) says that all of the means are similar.

? Exercise 11.4.2

Did we conduct pairwise comparisons when we *rejected* the null hypothesis when comparing two groups with a t-test? Why or why not?

Answer

I know that it was a long time ago, but no, we did not conduct pairwise comparisons with t-tests. Even when we rejected the null hypothesis (which said that the means were similar, so we are saying that they are probably different), we only had two means. The t-test was our "pairwise" comparison. In other words, because there were only two means, so we knew that if the means were statistically different from each other that the bigger one was statistically significantly bigger.

What about an ANOVA that compared three groups? To answer these questions, it doesn't matter if the ANOVA was BG or RM, just that there one was IV with three (or more) groups.

? Exercise 11.4.3

Did we conduct pairwise comparisons when we *retained* the null hypothesis when comparing three groups with an ANOVA? Why or why not?

Answer

No, we did not conduct pairwise comparisons with ANOVAS with three groups if we retained the null hypothesis. With any retained null hypothesis, we are agreeing that the means are similar, so we wouldn't spend time looking for any pairs of means that are different.

? Exercise 11.4.4

Did we conduct pairwise comparisons when we *rejected* the null hypothesis when comparing three groups with an ANOVA? Why or why not?

Answer

Yes, this is when we would conduct pairwise comparisons. The null hypothesis says that all of the means are similar, but when we reject that we are only saying that at least one mean is different from one other mean (one pair of means differs).

When we have three or more groups, we need to figure out which means differ from which other means. In other words, a significant ANOVA shows us that at least one of the means is different from at least one other mean, but we don't know which means are different from which other means. We have to do pairwise mean comparisons to see which means are significantly different from which other means.

Finally, on to factorial designs! If you've been answering the Exercises as you go, these should be pretty easy.

? Exercise 11.4.5

Do we conduct pairwise comparisons when we *retain* the null hypothesis for main effects in a factorial design? Why or why not?

Answer

No. When we retain a null hypothesis, we are saying that all of the means are similar. Let's not waste time looking for a difference when we just said that there wasn't one.

Okay, this one might be a little challenging, so we'll walk through it together.

✓ Example 11.4.1

Do we conduct pairwise comparisons when we *reject* the null hypothesis for main effects in a factorial design?

Solution

It depends!

If we only have two means, we don't have to conduct pairwise comparisons because (just like with a t-test) rejecting the null hypothesis for the main effect means that we know that the bigger mean is statistically significantly bigger.

But if our IV has more than two groups, then we would need to conduct pairwise comparisons (just like an in ANOVA) to find which means are different from which other means.

Back to an easier one on null hypotheses and post-hoc tests.

? Exercise 11.4.6

Do we conduct pairwise comparisons when we *retain* the null hypothesis for an interaction in a factorial design? Why or why not?

Answer

No. The null hypothesis says that all of the means are similar. If we retain the null hypothesis, then we are saying that all of the means are probably similar. Why would we look for a difference between pairs of means that we think are similar?

This one should be clear if you understand the reasoning for when we do and do not conduct post-hoc pairwise comparisons.

✓ Example 11.4.2

Do we conduct pairwise comparisons when we *reject* the null hypothesis for an interaction in a factorial design? Why or why not?

Solution

Yes! The smallest factorial design is a 2x2, which means that we have our means representing the combination of the two IVs. Rejecting the null hypothesis for the interaction says that at least one of those means is different from at least one other mean.

> We should use pairwise comparisons to find which combination of IV levels has a different mean from which other combination.

Short Answer

Table **11.4.1**- Short Answer for When to Conduct Post-Hoc Pairwise Comparisons

	Only Two Groups	Three or More Groups or Two or More IVs
Retain the Null Hypothesis	No- means are similar	No- means are similar
Reject the Null Hypothesis	No- The bigger group is statistically bigger	Yes- Find which mean is different from each other mean by comparing each pair of means.

Time to practice next!

This page titled 11.4: When Should You Conduct Post-Hoc Pairwise Comparisons? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

11.5: Practice with a 2x2 Factorial Design- Attention

Let's go through the process of looking at a 2x2 factorial design in the wild.

Scenario: Stand at Attention

Do you pay more attention when you are sitting or standing? This was the kind of research question the researchers were asking in the study we will look at. The research taht we are basing our calculations on is from a paper called "Stand by your Stroop: Standing up enhances selective attention and cognitive control" (Rosenbaum, Mama, & Algom, 2017), although Dr. MO tweaked the experiment to simplify ANOVA Summary Table calculations a tiny bit. This paper asked whether sitting versus standing would influence a measure of selective attention, the ability to ignore distracting information.

They used a classic test of selective attention, called the Stroop effect. In a typical Stroop experiment, participants name the color of words as fast as they can. The trick is that sometimes the color of the word is the same as the name of the word, and sometimes it is not. Some examples are in Figure **11.5.1**. Try it yourself! It's really difficult to name the color of the word in the Incongruent condition; your brain wants to read the word, not look at the color of the letters!

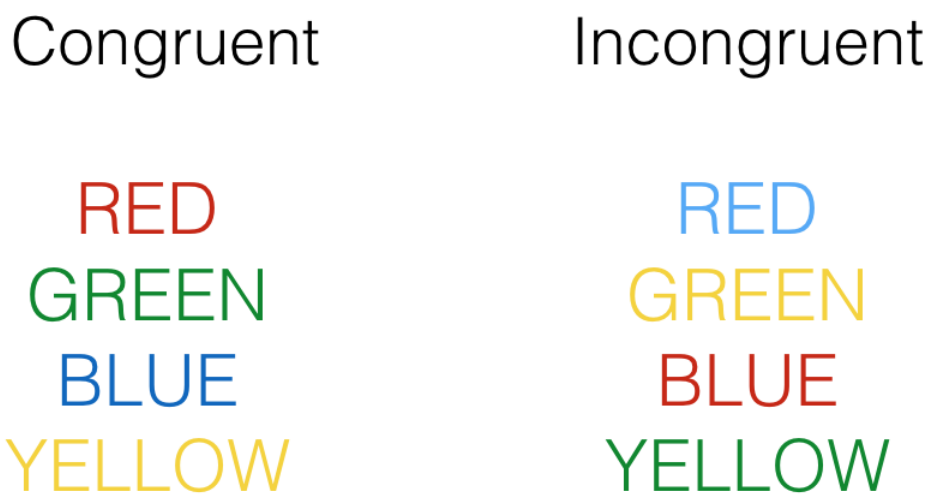


Figure **11.5.1**: Examples of congruent and incongruent Stroop stimuli. The task is to name the color, not the word. (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

The task is to name the color of the letters. Congruent trials occur when the color and word match. So, the correct answers for each of the congruent stimuli shown would be to say, red, green, blue and yellow. Incongruent trials occur when the color and word mismatch. The correct answers for each of the incongruent stimuli would be: blue, yellow, red, green.

The Stroop effect is an example of a well-known phenomena. What happens is that people are faster to name the color of the congruent items compared to the color of the incongruent items. This difference (incongruent reaction time - congruent reaction time) is called the Stroop effect.

Many researchers argue that the Stroop effect measures something about selective attention, the ability to ignore distracting information. If this theory is true, Dr. MO is very bad at ignoring distracting information! In the Stroop task, the target information that you need to pay attention to is the color of the word, not the word itself. For each item, the word itself is distracting, it is not the information that you are supposed to respond to. However, it seems that most people can't help but notice the word, and their performance in the color-naming task is subsequently influenced by the presence of the distracting word.

People who are good at ignoring the distracting words should have small Stroop effects. They will ignore the word, and it won't influence them very much for either congruent or incongruent trials. As a result, the difference in performance (the Stroop effect) should be fairly small if you have "good" selective attention in this task. People who are bad at ignoring the distracting words

should have big Stroop effects. They will not ignore the words, causing them to be relatively fast when the word names the color of the letters, and relatively slow when the word mismatches. As a result, they will show a difference in performance between the incongruent and congruent conditions.

If we take the size of the Stroop effect as a measure of selective attention, we can then start wondering what sorts of things improve selective attention (e.g., that make the Stroop effect smaller), and what kinds of things impair selective attention (e.g., make the Stroop effect bigger).

The research question of this study was to ask whether standing up improves selective attention compared to sitting down. They predicted smaller Stroop effects when people were standing up and doing the task, compared to when they were sitting down and doing the task.

Let's start by seeing if you understand the scenario.

? Exercise 11.5.1

Answer the following questions to understand the variables and groups that we are working with.

1. What are the IVs and their levels?
2. What is the DV (quantitative variable being measured)?
3. Is this a 2x2 factorial design? If not, what kind of design is it?
4. List out each of the *combinations* of the levels of the IVs.

Answer

1. One IV is whether the color of the letters and the word are congruent or incongruent; let's call that one Congruency. The other IV is whether the participants were standing or sitting, and we can call that Posture.
2. The DV is how long it took participants to name the color. This was probably measured in milliseconds.
3. Yes, this is a 2x2 factorial design because there are two IVs (two numbers) and each IV has two levels (each number is a "2").
4. Because 2x2 = 4, we will have four combinations:
    1. Congruent color while standing
    2. Congruent color while sitting
    3. Incongruent color while standing
    4. Incomgruent color while sitting

Step 1. State the Hypothesis

They had participants perform many individual trials responding to single Stroop stimuli, both congruent and incongruent. And they had participants stand up sometimes and do it, and sit-down sometimes and do it. Figure **11.5.2** is a bar graph of the means.

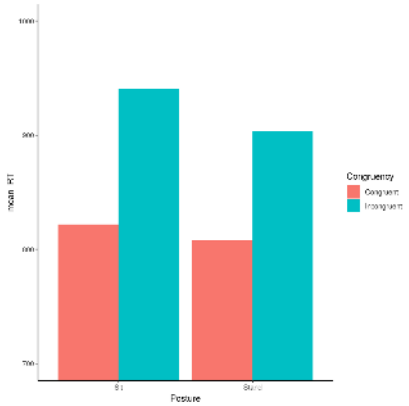


Figure **11.5.2**: Means from Rosenbaum et al (2017). (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

We can see in Figure **11.5.2** that Stroop effects were observed in both the sitting position and the standing position because it took participants longer to respond when the word was incongruent with the color of the letters. In the sitting position, mean congruent reaction times (RT) were shorter than mean incongruent RTs (the red bar on the left of each pair is lower than the aqua bar on the right of each pair). The same general pattern is observed for the standing position.

Based on these means, what do you predict for your research hypotheses? Will there be a main effect for Congruency? A main effect for posture? An interaction of Congruency and Posture?

Example 11.5.1

Use the information in Figure **11.5.2** to describe any main effects or interaction that you predict (in words only). Make sure that you predict the direction of effects by naming which group will have shorter reaction times.

Solution

Looking at the set of two differently colored bars, Dr. MO predicts that there will be a main effect of Congruency such that participants will react faster (smaller reaction time) when the words and colors match (Congruent) than when they do not match (Incongruent).

Trying to look at only Posture (sort "averaging" the pairs of bars in my head, Dr. MO predicts that there may be a main effect of Posture, too, such that participants react faster while standing rather than sitting. [Note to readers: This is not at all clear in the data or in the way that the data is presented, but that's why it's a hypothesis, an educated guess!]

Is there an interaction? Dr. MO would say yes, reaction time seems to be even faster for congruent words when standing than sitting, with participants with incongruent words while sitting being particularly slow to react to.

Research hypothesis are pretty complex when ther are more than two IVs. Researchers have prior reserach and theory that often suggests one specific interaction to expect. Being new to statistics and new to this field of selective attention, we are sorta going in blind.

What about null hypotheses? We still have those in factorial designs!

Exercise 11.5.2

What are the null hypotheses for this scenario (in words)?

Answer

You could simply state that all means will be similar.

If it makes more sense to you to mirror the research hypotheses, you can also list out the main effects and interaction:

- The null hypothesis for Congruency is that the mean reaction time for the congruent condition will be similar as the reaction time for the incongruent condition.
- The null hypothesis for Posture is that the mean reaction time while sitting will be similar as the reaction time while standing.
- The null hypothesis for the interactoin is that the average reaction for Congruency and Posture combined will be similar. [You could list out all of the combinations here again, too.]

Step 2. Find the Critical Values

You are welcome to complete this step after you've completed the ANOVA Summary Table so that we have all of the necessary Degrees of Freedom, but we'll use the formulas for Degrees of Freedom to do that now.

Formulas for Degrees of Freedom

1. Cells: $(k_1 \times k_2)-1$
    1. Remembering that “k” is the number of groups, k_1 is the number of levels of the first IV and , k_2 is the number of levels of the other IV.

2. Between group for one variable ($IV1_1$): k_1–1
3. Between group for the other variable (IV_2): k_2–1
4. Interaction: $df_1 \times df_2$
5. Within group: df_{Total}–df_{Cells}
6. Total: $N - 1$
    1. With N being the number of scores.

? Exercise 11.5.3

What are the Degrees of Freedom for each of the six sources if N = 50?

Answer

1. Cells: $(k_1 \times k_2)$–$1 = (2 \times 2) - 1 = 4 - 1 = 3$
2. Between group for one variable ($IV1_1$): k_1–$1 = 2 - 1 = 1$
3. Between group for the other variable (IV_2) k_2)–$1 = 2 - 1 = 1$
4. Interaction: $df_1 \times df_2 = 1 \times 1 = 1$
    1. Notice that we are multiplying the Degrees of Freedom of each, not k! It's so easy to get confused...
5. Within group: df_{Total}–$df_{Cells} = 49 - 3 = 46$
    1. You have to find the Degrees of Freedom for the Total (#6 here) before you can find this one.
6. Total: $N - 1 = 50 - 1 = 49$

Well, that was a lot!

Now, let's use these to find the critical values.

? Exercise 11.5.4

Using the Table of Critical F-Scores, what are the critical values of F for p=0.05 for:

1. Main Effect of IV 1 (Congruency)?
2. Main Effect of IV 2 (Posture)?
3. Interaction?

Answer

1. Main Effect of IV 1 (Congruency) = Using the Degrees of Freedom of the numerator (IV 1) and the denominator (WG), $F_{Crit}(1, 46) = 4.08$ (if you round down) or 4.03 (if you find the nearest df of the denominator).
2. Main Effect of IV 2 (Posture) = Using the Degrees of Freedom of the numerator (IV 2) and the denominator (WG), $F_{Crit}(1, 46) = 4.08$ (if you round down) or 4.03 (if you find the nearest df of the denominator).
3. Interaction = Using the Degrees of Freedom of the numerator (Interaction) and the denominator (WG), $F_{Crit}(1, 46) = 4.08$ (if you round down) or 4.03 (if you find the nearest df of the denominator).

Step 3: Compute the Test Statistic

We'll complete an ANOVA Summary to whether the differences in the means are likely or unlikely to be due to chance. The ANOVA Summary Table will give us main effects for Congruency and Posture (the two IVs), as well as one interaction effect to evaluate (Congruency X Posture). The interaction effect tells us whether the congruency effect changes across the levels of the posture manipulation.

Table **11.5.1**- Two-Way ANOVA Summary Table

Source	SS	df	MS	F
IV 1	576,821.64			
IV 2	32,303.45			
Interaction	6,560.34			
Within Groups (Error)	35,859.07			
Total	651,544.50			

Example 11.5.2

Use the Sum of Squares provided in Table **11.5.1** to fill in the rest of the ANOVA Summary Table in Table **11.5.2**.

Solution

Table **11.5.2**- Two-Way ANOVA Summary Table with Calculations

Source	SS	df	MS	F
IV 1	576,821.64	$k_1 - 1$ = 2 - 1 1	$\frac{SS_1}{df_1} = \frac{576821.64}{1} = 576821.64$	$\frac{MS_1}{MS_{WG}} = \frac{576821.64}{779.55} = 739.94$
IV 2	32,303.45	$k_2 - 1$ = 2 - 1 = 1	$\frac{SS_2}{df_2} = \frac{35303.45}{1} = 32303.45$	$\frac{MS_2}{MS_{WG}} = \frac{32303.45}{779.55} = 41.44$
Interaction	6,560.34	$df_1 \times df_2$ = 1 * 1 = 1	$\frac{SS_{INT}}{df_{INT}} = \frac{6560.34}{1} = 6560.34$	$\frac{MS_{Interaction}}{MS_{WG}} = \frac{6560.34}{779.55} = 8.42$
Within Groups (Error)	35,859.07	$df_{Total} - df_{cells}$ = 49 - 3 = 46	$\frac{SS_{WG}}{df_{WG}} = \frac{35859.07}{46} = 779.55$	N/A
Total	651,544.50	$N - 1 = 50 - 1 = 49$	N/A	N/A

To complete the table, you need to need to calculate the degrees of freedom of cells \((df_{cells} = (k_1 \times k_2) - 1)\) although there's no place in the table to include this information:

$$df_{cells} = (k_1 * k_2)–1$$
$$df_{cells} = (2 * 2) - 1$$
$$df_{cells} = 4 - 1$$
$$df_{cells} = 3$$

That wasn't so hard! Now, let's figure out what this all means.

Step 4. Make the Decision

To make the decision, we need to compare each calculated F-score to the critical values that we found in Step 2.

Exercise 11.5.5

For each calculated F (main effect for IV 1, main effect for IV 2, interaction), decide if the null hypothesis should be retained or rejected.

(Critical < Calculated) = Reject null = At least one mean is different from at least one other mean. = $p < .05$

(Critical > Calculated) = Retain null = All of the means are similar. = $p > .05$

Answer

1. For the Main Effect of IV 1 (Congruency): $F_{Crit}(1, 46) = 4.08$ (if you round down) or 4.03 (if you find the nearest df of the denominator). With an $F - Calc = 739.94$, we reject the null hypothesis.
2. For the Main Effect of IV 2 (Posture): $F_{Crit}(1, 46) = 4.08$ (if you round down) or 4.03 (if you find the nearest df of the denominator). With an $F_{Calc} = 41.44$, we reject the null hypothesis.
3. For the Interaction: $F_{Crit}(1, 46) = 4.08$ (if you round down) or 4.03 (if you find the nearest df of the denominator). With an $F_{Calc} = 8.42$, we reject the null hypothesis.

That wasn't so hard, but now what?

Example \(\PageIndex{3}\)

What would the statistical sentences look like for each effect?

Solution

1. Main Effect of IV 1 (Congruency): F(1, 46)=739.94,p<.05
2. Main Effect of IV 2 (Posture): F(1, 46)=41.44,p<.05
3. Interaction: F(1, 46)=8.42,p<.05

Before we can write the conclusion, we need to know the means and see if they are in the direction that we hypothesized way back in Step 1. The means are shown in the Punnett's Square in Table **11.5.3**.

Table **11.5.3**-Means for 2x2 (Congruency by Posture) Factorial ANOVA

IV Levels	IV1: Congruent	IV1: Incongruent	Marginal Means for IV2
IV2 (Posture): Sitting	821.92	940.79	881.35
IV2 (Posture): Standing	807.96	903.91	855.94
Marginal Means for IV1	814.94	922.35	(You could put the Total Mean here)

Oh, snap, what about post-hoc pairwise comparisons?

The good news is that pairwise comparisons aren't necessary to calculate with either statistically significant main effect because both of our IVs only have two levels. When we rejected the null hypothesis for the main effects, we determined that the bigger of the two means was statistically bigger.

But what about the interaction? We rejected that null hypothesis, too! And as we discussed in the prior section, pairwise comparisons are necessary when the null hypothesis for an interaction is rejected. However, just like the Sums of Squares are more of a hassle to do by hand than is worth it for factorial designs, the post-hoc analyses are also confusing to calculate with the combinations of cells. Instead, let's find the pairwise mean differences, and compare them to a critical value of 14.86; any mean differences that have an absolute value larger than 14.86 are statistically significantly different.

Example 11.5.4

Calculate the mean differences for each pair of means in the interaction, and determine which are statistically significantly different when the critical value is 14.86.

Solution

Table **11.5.4**- Mean Differences

Pairs Being Compared	Mean Difference	Statistically Significant?
Congruent while Sitting minus Congruent while Standing	13.96	No (13.96 is smaller than the critical value of 14.86)
Congruent while Sitting minus Incongruent while Sitting	-118.87	Yes
Congruent while Sitting minus Incongruent while Standing	-81.99	Yes
Congruent while Standing minus Incongruent while Sitting	-132.83	Yes
Congruent while Standing minus Incongruent while Standing	-95.95	Yes
Incongruent while Sitting minus Incongruent while Standing	36.88	Yes

Based on these calculations, all of the pairwise comparison are statistically significant (the mean differences are larger than the critical value of 14.86) except for the combination of Congruent while Sitting compared with Congruent while Standing.

We are ready to do the write-up! To remind ourselves of the research hypotheses, let's list them here and identify if they were supported or not.

> For IV 1, Congruency, Dr. MO predicted a main effect such that participants will react faster (smaller reaction time) when the words and colors match (Congruent) than when they do not match (Incongruent). This prediction was supported because there was a significant main effect ($F(1, 46) = 739.94$, $p<.05$). Looking at the means in Table **11.5.3**, the Congruent group ($M = 814.94$ ms) did have a faster (smaller) reaction time than the Incongruent group ($M = 922.35$ ms).
>
> For IV 2, Posture, Dr. MO predicted a main effect such that participants react faster while standing rather than sitting. This prediction was supported because there was a significant main effect $F(1, 46) = 41.44$, $p<.05$). Looking at the means in Table **11.5.3**, the Standing group ($M = 855.94$ ms) did have a faster (smaller) reaction time than the Sitting group ($M = 881.35$ ms).
>
> Dr. MO predicted an interaction such that reaction time seems to be even faster for congruent words when standing than sitting, with participants with incongruent words while sitting being particularly slow to react to. There was a significant interaction $F(1, 46) = 8.42$, $p<.05$). Looking at the means in Table **11.5.3** and the mean differences in Table **11.5.4**, this prediction was partially supported. It is true that participants were particularly slow to react to incongruent words while sitting ($M = 940.79$), but reaction times did not differ for congruent words while standing ($M = 807.96$) compared to sitting ($M = 821.92$).

Oops, we accidently just did the whole write-up!

Example 11.5.5

Are all of the four required components when Reporting Results included?

Solution

- The statistical test is preceded by the descriptive statistics (means)- We included almost all of the means in the actual write-up, and directed readers to the tables with all of the means.
- The description tells you what the research hypothesis being tested is- Yes, each paragrph re-stated what the prediction was.
- A "statistical sentence" showing the results is included- Yes, each prediction had a "statistical sentence."
- The results are interpreted in relation to the research hypothesis- Yes, it was stated whether the research hypotheses (predictions) were supported, or partially supported.

Summary of Stand At Attention

Another way to think about this is to look at the mean differences for the Stroop effect (which was the main effect of IV 1, the difference between the congruent condition and the incongruent condition).

- In the sitting condition the Stroop effect was roughly 941-822=119 ms.
- In the standing condition the Stroop effect was roughly 904-808=96 ms.

So, the Stroop effect was 119-96=23 ms smaller when people were standing. This is a pretty small effect in terms of the amount of time reduced, but even though it is small, a difference even this big was *not* very likely to be due to chance.

Based on this research there appears to be some support for the following logic chain. First, the researchers can say that standing up reduces the size of a person's Stroop effect. Fine, what could that mean? Well, if the Stroop effect is an index of selective attention, then it could mean that standing up is one way to improve your ability to selectively focus and ignore distracting information. The actual size of the benefit is fairly small, so the real-world implications are not that clear. Nevertheless, maybe the next time you lose your keys, you should stand up and look for them, rather than sitting down and not look for them.

Let's try one more practice, this time with mindset data!

Contributors and Attributions

- Matthew J. C. Crump (Brooklyn College of CUNY)
- Dr. MO (Taft College)

This page titled 11.5: Practice with a 2x2 Factorial Design- Attention is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **9.8: Real Data** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

11.5.1: Practice 2x3 Factorial ANOVA on Mindset

One more time to practice with a factorial ANOVA Summary Table and interpretation!

Scenario

We are going back to one of the student research projects to see if faculty can increase understanding and belief in growth mindset. To see if students' scores on the Mindset Quiz improved so that more student held stronger growth mindset view, the outcome that we are trying to improve is the Difference between their Mindset Quiz score at the beginning of the semester (pretest), and their Mindset Quiz score at the end of the semester (posttest). A positive score means that students' mindset improved, while a negative score means that their mindset actually became more fixed. We are only going to look at students who did not start holding strong growth mindset beliefs to see if any improvement was made. In this version of the study, we had three types of interventions:

1. A comparison control group who did not complete any activities related to learning about growth mindset. They only took the Mindset Quiz at the beginning of the semester and again at the end of the semester.
2. A minimal intervention group who had some activities and discussion about growth mindset, but not every week.
3. A super-intervention group who had weekly activities and discussion about growth mindset. The concept of mindset was embedded in their curriculum.

At conference discussing prior studies in this series, the student researchers had a discussion with researcher from another school and realized that all of the professors who participated in the study were women (of varying racial identities). We then went back to see if the gender of the student influenced their changes in growth mindset. We had four students whose gender was unknown or non-binary, but they were all in the minimal intervention group, so we excluded them from the analyses and only compared men versus women. Because this study took place in real classrooms, the size of each group varied, but there was a total of 46 students who completed both the pretest and posttest of the Mindset Quiz and indicated their gender.

Let's start by seeing if you understand what's going on in this scenario.

? Exercise 11.5.1.1

Answer the following questions to understand the variables and groups that we are working with.

1. What are the IVs and their levels?
2. What is the DV (quantitative variable being measured)?
3. Is this a 2x2 factorial design? If not, what kind of design is it?
4. List out each of the *combinations* of the levels of the IVs.

Answer

1. One IV was the Mindset Intervention, with the three levels being Control, Minimal, or Super. The other IV was gender, with the levels being men or women.
2. The DV is the Difference between each students score on the first Mindset Quiz (pretest) subtracted from their score on the later Mindset Quiz (posttest).
3. This is a 2x3 factorial design because there are two IVs (two numbers) and the IV with fewer levels has 2 levels, and the IV with more levels had 3 levels.
4. Because 2x3 = 6, we will have six combinations:
    1. Women in the Control group
    2. Women in the Minimal group
    3. Women in the Super group
    4. Men in the Control group
    5. Men in the Minimal group
    6. Men in the Super group

Step 1. State the Hypothesis

Knowing nothing else, what do you predict for your research hypotheses? Will there be a main effect for gender? A main effect for the mindset intervention? An interaction of gender by mindset intervention?

? Exercise 11.5.1.2

Describe any main effects or interaction that you predict (in words only). Make sure that you predict the direction of effects by naming which group will have larger difference scores.

Answer

Dr. MO predicts that there will not be a main effect for the average Differences score from the pretest to the posttest based on gender.

Dr. MO predicts that there will be a main effect for the intervention such that the Control group will have the smallest Difference scores, the Minimal intervention group will be in the middle, and the Super intervention group will have the largest Difference score. All means will be statistically significantly different from each other.

Dr. MO predicts that there will be an interaction such that women will have a larger Difference score in in the Minimal and Super intervention scores than men, but gender won't matter in the Control group. She's basing this on the fact that all of the professors were women.

Your research hypothesis may differ! No means were provided, so we're really just guessing based on the scenario and what we know about psychology and growth mindset…

What about the null hypotheses?

? Exercise 11.5.1.3

What are the null hypotheses for this scenario (in words)?

Answer

All means will be similar.

If it makes more sense to you to mirror the research hypotheses, you can also list out the main effects and interaction:

- The null hypothesis for gender is that the mean Difference score for the men condition will be similar to the Difference score for women.
- The null hypothesis for the Mindset Intervention is that the mean Difference will be similar between the Control group, Minimal group, and Super group.
- The null hypothesis for the interaction is that the average Difference score for Gender and Mindset Intervention combined will be similar. [You could list out all of the combinations here again, too.]

Step 2. Find the Critical Values

You are welcome to complete this step after you've completed the ANOVA Summary Table so that we have all of the necessary Degrees of Freedom. Instead, we will find the Degrees of Freedom now.

Formulas for Degrees of Freedom

1. Cells: $(k_1 \times k_2)-1$
2. Between group for one variable ($IV1_1$): k_1-1
3. Between group for the other variable (IV_2): k_2-1
4. Interaction: $df_1 \times df_2$
5. Within group: $df_{Total}-df_{Cells}$
6. Total: N-1

? Exercise 11.5.1.4

What are the Degrees of Freedom for each of the six sources if N = 46?

Answer

1. Cells: $(k_1 \times k_2)–1 = (2 \times 3) – 1 = 6 – 1 = 5$
2. Between group for one variable ($IV1_1$): $k_1–1 = 2 – 1 = 1$
3. Between group for the other variable (IV_2) $k_2–1 = 3 – 1 = 2$
4. Interaction: $df_1 \times df_2 = 1 \times 2 = 2$
    1. Notice that we are multiplying the Degrees of Freedom of each, not k! It's so easy to get confused...
5. Within group (You have to find the Degrees of Freedom for the Total (#6 here) before you can find this one.): $df_{Total}–df_{Cells} = 45 – 5 = 40$
6. Total: $N – 1 = 46 – 1 = 45$

Now that we have all of the Degrees of Freedom, we can use them to find the critical values.

? Exercise 11.5.1.5

Using the Table of Critical F-Scores, what are the critical values of F for p=0.05 for:

1. Main Effect of IV 1 (Gender)?
2. Main Effect of IV 2 (Mindset Intervention)?
3. Interaction?

Answer

1. Main Effect of IV 1 (Gender) = Using the Degrees of Freedom of the numerator (IV 1) and the denominator (WG), $F_{Crit}(1, 40) = 4.08$.
2. Main Effect of IV 2 (Mindset Intervention) = Using the Degrees of Freedom of the numerator (IV 2) and the denominator (WG), $F_{Crit}(2, 40) = 3.23$.
3. Interaction = Using the Degrees of Freedom of the numerator (Interaction) and the denominator (WG), $F_{Crit}(2, 40) = 3.23$.

Step 3. Compute the Test Statistic

We'll skip calculating the Sums of Squares; you can find them already in the ANOVA Summary Table (Table 11.5.1.1).

Table 11.5.1.1- Factorial ANOVA Summary Table

Source	*SS*	*df*	*MS*	*F*
IV 1	141.69			
IV 2	31.82			
Interaction	53.01			
Within Groups (Error)	1,072.08			
Total	1,298.60			

✓ Example 11.5.1.1

Use the Sum of Squares provided in Table 11.5.1.1 to fill in the rest of the ANOVA Summary Table in Table 11.5.1.2.

Solution

To complete the table, you need to need to calculate the degrees of freedom of cells ($df_{cells} = (k_1 \times k_2)$–1) although there's no place in the table to include this information:

$$df_{cells} = (k_1 \times k_2)–1$$
$$df_{cells} = (2 \times 3) - 1$$
$$df_{cells} = 6 - 1$$
$$df_{cells} = 5$$

Table **11.5.1.2**-Factorial ANOVA Summary Table

Source	SS	df	MS	F
IV 1	141.69	$k_1 - 1 = 2 - 1 = 1$	$\frac{SS_1}{df_1} = \frac{141.69}{1} = 141.69$	$\frac{MS_1}{MS_{WG}} = \frac{141.69}{26.80} = 5.29$
IV 2	31.82	$k_2 - 1 = 3 - 1 = 2$	$\frac{SS_2}{df_2} = \frac{31.82}{2} = 15.91$	$= \frac{MS_2}{MS_{WG}} = \frac{15.91}{26.80} = 0.59$
Interaction	53.01	$df_1 \times df_2 = 1 \times 2 = 2$	$= \frac{SS_{INT}}{df_{INT}} = \frac{53.01}{2} = 26.51$	$\frac{MS_{Interaction}}{MS_{WG}} = \frac{26.51}{26.80} = 0.99$
Within Groups (Error)	1,072.08	$df_{Total} - df_{cells} = 45 - 5 = 40$	$= \frac{SS_{WG}}{df_{WG}} = \frac{1072.08}{40} = 26.80$	N/A
Total	1,298.60	$N - 1 = 46 - 1 = 45$	N/A	N/A

We did it! The completed ANOVA Summary Table for this scenario looks like Table **11.5.1.3**

Table **11.5.1.3**-Factorial ANOVA Summary Table

Source	SS	df	MS	F
IV 1	141.69	1	141.69	5.29
IV 2	31.82	2	15.91	0.59
Interaction	53.01	2	26.51	0.99
Within Groups (Error)	1,072.08	40	26.80	leave blank
Total	1,298.60	45	leave blank	leave blank

ANOVA Summary Tables with more than a 2x2 design are less repetitive than ANOVA Summary Tables for a 2x2, huh?

Step 4. Make the Decision

To make the decision, we need to compare each calculated F-score to the critical values that we found in Step 2.

Note

Remember:

(Critical < Calculated) = Reject null = At least one mean is different from at least one other mean. = $p < .05$

(Critical > Calculated) = Retain null = All of the means are similar. = $p > .05$

Exercise 11.5.1.6

For each calculated F (main effect for IV 1, main effect for IV 2, interaction), decide if the null hypothesis should be retained or rejected.

Answer

1. For the Main Effect of IV 1 (Gender) = $F_{Crit}(1, 40) = 4.08$, so we reject the null hypothesis. The higher mean is higher (because there are only two means).
2. Main Effect of IV 2 (Mindset Intervention) = $F_{Crit}(2, 40) = 3.23$, so we retain the null hypothesis. The three means for the Control group, Minimal group, and Super group were similar.
3. Interaction = $F_{Crit}(2, 40 = 3.23$, so we retain the null hypothesis. Gender and the Mindset Intervention did not interact.

Factorial designs are more complex, but it's the same basic process that we've been working through this whole time.

But before we move on to the full reporting of results, let's just make sure that we know how to write the "statistical sentences".

? Exercise 11.5.1.1

What would the statistical sentences look like for each effect?

Answer

1. Main Effect of IV 1 (Gender): F(1, 40)=5.29, p<.05
2. Main Effect of IV 2 (Mindset Intervention): F(2,40)=0.59, p>.05
3. Interaction: F(2, 40)=0.99, p>.05

Pay attention to the greater than/lesser than sign after the p! If p<.05, it means that there's a very small probability that the means are similar. If p>.05, there's a large probability that the means are similar (that the null hypothesis is correct).

Before we can write the conclusion, we need to know the means to see if they are in the direction that we hypothesized way back in Step 1. The means are shown in in Table **11.5.1.4**. We do not need to conduct any pairwise comparisons because the only means that were statistically significantly different were for gender, which only had two levels, so we don't have to do any extra calculations to know that the higher mean is significantly higher.

Table **11.5.1.4**-Mean Differences for 2x3 (Gender by Mindset Intervention) Factorial ANOVA

IV Levels	**IV1 (Gender): Men**	**IV1 (Gender): Women**	Marginal Means for IV2
IV2 (Mindset Intervention): Control	-2.33	1.56	0.50
IV2 (Mindset Intervention): Minimal	-3.50	3.16	1.34
IV2 (Mindset Intervention): Super	1.00	2.33	1.62
Marginal Means for IV1	-1.31	2.57	1.22

We are ready to do the write-up! Remember, these are means of the Difference score between each participants' pretest and posttest, and negative scores mean that the student's posttest was actually lower than their pretest; another way to say that is that the student was closer to holder growth mindset beliefs at the beginning of the semester (before the intervention) than at the end of the semester.

It can be easier to go through each research hypothesis, one at a time.

✓ Example 11.5.1.2

Evaluate the research hypothesis for the main effect of gender by writing up the conclusion.

Solution

It was predicted that there will not be a main effect for the average Differences score from the pretest to the posttest based on gender. She was incorrect; there were gender differences in the Difference score (F(1, 40)=5.29, p<.05) such that women scored higher (M=2.57) than men (M=-1.31).

Now for the second main effect:

Example 11.5.1.3

Evaluate the research hypothesis for the main effect of the intervention by writing up the conclusion.

Solution

It was predicted that there will be a main effect for the intervention such that the Control group will have the smallest Difference scores, the Minimal intervention group will be in the middle, and the Super intervention group will have the largest Difference score. This was not supported (F(2, 40)=0.59, p>.05). Instead, the three means were similar ($M_{Control} = 0.50; M_{Minimal} = 1.34, M_{Super} = 1.62$).

And then the interaction:

Example 11.5.1.4

Evaluate the research hypothesis for the interaction by writing up the conclusion.

Solution

It was predicted predicted that there will be an interaction such that women will have a larger Difference score in in the Minimal and Super intervention scores than men, but gender won't matter in the Control group. This was not supported (F(2,40)=0.99, p>.05). The means by gender and mindset intervention did not statistically significantly interact. The means for each combination can be seen in Table **11.5.1.4**.

Before you move on, look at back at your own write-ups to makes sure that each of them includes all of the four required components when Reporting Results.

Summary of the Current Mindset Study

Now that we've interpreted all of the statistical analyses, let's step back a bit and think about this study. It was predicted that gender wouldn't matter on its own, but that the mindset intervention would. But the results actually are the opposite! What happened?

Well, we don't know for sure, and would have to do several more studies to try to figure out exactly what was going on. What we can see is that these interventions designed to increase Mindset Quiz scores actually made men have a more fixed mindset. We're not really sure why that is, and future research should have more than professors who are women provide the mindset interventions. We could also look at the interventions themselves, and see if those may have inadvertently encouraged women but discouraged men. But this is the fun of research! We accidently found a puzzle, and can do more and more research to find all of the puzzle pieces to understand what's going on!

Or, we could be experiencing a Type I Error and/or a Type II Error!

Example 11.5.1.5

What is a Type I Error, and what is a Type II Error (in your own words)? How could they relate to this scenario?

Solution

Type I Errors are when you reject null hypothesis when it is true. In this scenario, we could be saying that there is a difference in changes in mindset between the genders when there wouldn't be any change in a larger sample or in the population.

Type II Errors are when we retain the null hypothesis when it is false. In this scenario, we said that the intervention didn't work. If we committed a Type II Error, we could be missing how influential these interventions are because we had a wonky sample.

The good news is that conducting more research studies solves those problems, too! Win-win! If we conduct more research, we can figure out the puzzle, and rule out either of these errors.

Okay, we're almost done with this chapter, and with learning how to compare means between groups. The only next step is to decide which statistical analysis to run!

This page titled 11.5.1: Practice 2x3 Factorial ANOVA on Mindset is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

11.6: Choosing the Correct Analysis- Mean Comparison Edition

It's time to step back for a second and reflect on all that you've learned!

Even though this is a statistics course, you've actually learned a lot about research design. You've learned that participants can be different in the different groups or levels of the IV (independent or between groups designs), or they can be the same people or somehow linked together (dependent or repeated measures designs). You've learned that we can compare the means for one IV with two or more groups, or even statistically compare the means of more than one IV. It was sorta hidden, but we also talked about factorial designs that could have one IV that was independent and a second IV that is dependent; this is called a mixed designed.

We've almost exclusively focused on DVs that are means (quantitative variables), but sometimes mentioned ranked DVs (ordinal variables) when discussing non-parametric analyses.

Knowing the design of the IVs and DV determines what kind of statistical analysis that you use. So far, we've covered several ways to compare the mean of groups:

- One-sample t-test
- Independent t-test
- Dependent t-test
- Between Groups ANOVA
- Repeated Measures ANOVA

And these compare the ranks of groups:

- Mann-Whitney U
- Wilcoxon Match-Pair Signed-Rank test
- Kruskal-Wallis H
- Friedman's test

So when would you use which one? It depends! It depends on:

- The type of DV (qualitative or quantitative or ranked/ordinal)
- The number of groups
- Whether the groups are independent or dependent or the population

This Decision Tree handout is a great resource for making that decision. I encourage you to save this document for your future classes. You can use it now to go through the analyses that we've learned so far to see why we use them when we use them. Pay attention to the names, they tell you a lot about why they are used when they are used.

Example 11.6.1

Identify the type of DV, number of groups, and type of groups for the analyses that we've learned so far:

- One-sample t-test
- Independent t-test
- Dependent t-test
- Between Groups ANOVA
- Repeated Measures ANOVA
- Mann-Whitney U
- Wilcoxon Match-Pair Signed-Rank test
- Kruskal-Wallis H
- Friedman's test

Solution

- One-Sample t-test
  - The type of DV: Compares means, so quantitative
  - The number of groups: Compares a sample to the population

    - Whether the groups are independent or dependent or the population: The population
- Two-sample independent t-test
    - The type of DV: Compares means, so quantitative
    - The number of groups: Compares two samples
    - Whether the groups are independent or dependent or the population: Independent (unrelated)
- Two-sample dependent t-test
    - The type of DV: Compares means, so quantitative
    - The number of groups: Compares two samples
    - Whether the groups are independent or dependent or the population: Dependent (related)
- Between Groups ANOVA
    - The type of DV: Compares means, so quantitative
    - The number of groups: Compares two or more levels of an IV
    - Whether the groups are independent or dependent or the population: Independent (unrelated)
- Repeated Measures ANOVA (also called Within Groups ANOVA)
    - The type of DV: Compares means, so quantitative
    - The number of groups: Compares two or more levels of an IV
    - Whether the groups are independent or dependent or the population: Dependent (related)
- Mann-Whitney U
    - The type of DV: Ranked data, or we assume that the distribution is NOT normally distributed.
    - The number of groups: Compares two groups
    - Whether the groups are independent or dependent or the population: Independent (unrelated)
- Wilcoxon Match-Pair Signed-Rank test
    - The type of DV: Ranked data, or we assume that the distribution is NOT normally distributed.
    - The number of groups: Compares two groups
    - Whether the groups are independent or dependent or the population: Dependent (related)
- Kruskal-Wallis One-Way ANOVA
    - The type of DV: Ranked data, or we assume that the distribution is NOT normally distributed.
    - The number of groups: Compares two or more groups
    - Whether the groups are independent or dependent or the population: Independent (unrelated)
- Friedman's test
    - The type of DV: Ranked data, or we assume that the distribution is NOT normally distributed.
    - The number of groups: Compares two or more groups
    - Whether the groups are independent or dependent or the population: Dependent (related)

You might notice that all these have quantitative (or ranked) DVs with two or more groups. That's what this unit is about! In the next set of chapters, we'll learn about the appropriate analyses when the DVs aren't means.

Because this is an introductory textbook, we leave out a full discussion other analyses of means, and particularly of mixed designs. In this chapter, in particular, we are leaving out are the formulas to construct the Sum of Squares for each effect. There are many good more advanced textbooks that discuss these issues in much more depth. And, these things can all be found online.

Here's an interactive website to help you practice when to use which kind of statistical analysis. To start, click on the kinds of analyses that you are familiar with, then hit Submit.

Contributors and Attributions

This page was extensively adapted by Michelle Oja (Taft College) from work by Matthew J. C. Crump (Brooklyn College of CUNY)

This page titled 11.6: Choosing the Correct Analysis- Mean Comparison Edition is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **10.3: Mixed Designs** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

CHAPTER OVERVIEW

12: Correlation (and a little about Regression)

Chapter by Matthew Crump

> *Correlation does not equal causation*
>
> *— Every Statistics and Research Methods Instructor Ever*

In the last chapter we had some data. It was too much too look at and it didn't make sense. So, we talked about how to look at the data visually using plots and histograms, and we talked about how to summarize lots of numbers so we could determine their central tendencies (sameness) and variability (differentness). And, all was well with the world.

Let's not forget the big reason why we learned about descriptive statistics. The big reason is that we are interested in getting answers to questions using data.

If you are looking for a big theme to think about while you take this course, the theme is: how do we ask and answer questions using data?

For every section in this book, you should be connecting your inner monologue to this question, and asking yourself: How does what I am learning about help me answer questions with data? Advance warning: we know it is easy to forget this stuff when we dive into the details, and we will try to throw you a rope to help you out along the way…remember, we're trying to answer questions with data.

We started Chapter two with some fake data on human happiness, remember? We imagined that we asked a bunch of people to tell us how happy they were, then we looked at the numbers they gave us. Let's continue with this imaginary thought experiment.

What do you get when you ask people to use a number to describe how happy they are? A bunch of numbers. What kind of questions can you ask about those numbers? Well, you can look at the numbers and estimate their general properties as we already did. We would expect those numbers tell us some things we already know. There are different people, and different people are different amounts of happy. You've probably met some of those of really happy people, and really unhappy people, and you yourself probably have some amount of happiness. "Great, thanks Captain Obvious".

Before moving on, you should also be skeptical of what the numbers might mean. For example, if you force people to give a number between 0-100 to rate their happiness, does this number truly reflect how happy that person is? Can a person know how happy they are? Does the question format bias how they give their answer? Is happiness even a real thing? These are all good questions about the **validity** of the construct (happiness itself) and the measure (numbers) you are using to quantify it. For now, though, we will side-step those very important questions, and assume that, happiness is a thing, and our measure of happiness measures something about how happy people are.

OK then, after we have measured some happiness, I bet you can think of some more pressing questions. For example, what causes happiness to go up or down. If you knew the causes of happiness what could you do? How about increase your own happiness; or, help people who are unhappy; or, better appreciate why Eeyore from Winnie the Pooh is unhappy; or, present valid scientific arguments that argue against incorrect claims about what causes happiness. A causal theory and understanding of happiness could be used for all of those things. How can we get there?

Imagine you were an alien observer. You arrived on earth and heard about this thing called happiness that people have. You want to know what causes happiness. You also discover that planet earth has lots of other things. Which of those things, you wonder, cause happiness? How would your alien-self get started on this big question.

As a person who has happiness, you might already have some hunches about what causes changes in happiness. For example things like: weather, friends, music, money, education, drugs, books, movies, beliefs, personality, color of your shoes, eyebrow length, number of cat's you see per day, frequency of subway delay, a lifetime supply of chocolate, etcetera etcetera (as Willy Wonka would say), might all contribute to happiness in someway. There could be many different causes of happiness.

12.1: If something caused something else to change, what would that look like?
12.2: Pearson’s r
12.3: Turning the numbers into a measure of co-variance
12.4: Examples with Data
12.5: Regression — A mini intro
12.6: Interpreting Correlations
12.7: Summary
12.8: References

This page titled 12: Correlation (and a little about Regression) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew J. C. Crump via source content that was edited to the style and standards of the LibreTexts platform.

12.1: If something caused something else to change, what would that look like?

Before we go around determining the causes of happiness, we should prepare ourselves with some analytical tools so that we could identify what causation looks like. If we don't prepare ourselves for what we might find, then we won't know how to interpret our own data. Instead, we need to anticipate what the data could look like. Specifically, we need to know what data would look like when one thing does not cause another thing, and what data would look like when one thing does cause another thing. This chapter does some of this preparation. Fair warning: we will find out some tricky things. For example, we can find patterns that look like one thing is causing another, even when that one thing DOES NOT CAUSE the other thing. Hang in there.

Charlie and the Chocolate factory

Let's imagine that a person's supply of chocolate has a causal influence on their level of happiness. Let's further imagine that, like Charlie, the more chocolate you have the more happy you will be, and the less chocolate you have, the less happy you will be. Finally, because we suspect happiness is caused by lots of other things in a person's life, we anticipate that the relationship between chocolate supply and happiness won't be perfect. What do these assumptions mean for how the data should look?

Our first step is to collect some imaginary data from 100 people. We walk around and ask the first 100 people we meet to answer two questions:

1. how much chocolate do you have, and
2. how happy are you.

For convenience, both the scales will go from 0 to 100. For the chocolate scale, 0 means no chocolate, 100 means lifetime supply of chocolate. Any other number is somewhere in between. For the happiness scale, 0 means no happiness, 100 means all of the happiness, and in between means some amount in between.

Here is some sample data from the first 10 imaginary subjects.

subject	chocolate	happiness
1	1	1
2	1	1
3	2	2
4	2	4
5	4	5
6	4	5
7	7	5
8	8	5
9	8	6
10	9	6

We asked each subject two questions so there are two scores for each subject, one for their chocolate supply, and one for their level of happiness. You might already notice some relationships between amount of chocolate and level of happiness in the table. To make those relationships even more clear, let's plot all of the data in a graph.

Scatter plots

When you have two measurements worth of data, you can always turn them into dots and plot them in a scatter plot. A scatter plot has a horizontal x-axis, and a vertical y-axis. You get to choose which measurement goes on which axis. Let's put chocolate supply on the x-axis, and happiness level on the y-axis. The plot below shows 100 dots for each subject.

```
library(ggplot2)
subject<-1:100
chocolate<-round(1:100*runif(100,.5,1))
happiness<-round(1:100*runif(100,.5,1))
the_df_CC<-data.frame(subject,chocolate,happiness)
ggplot(the_df_CC,aes(x=chocolate,y=happiness))+
  geom_point()+
  theme_classic()
```

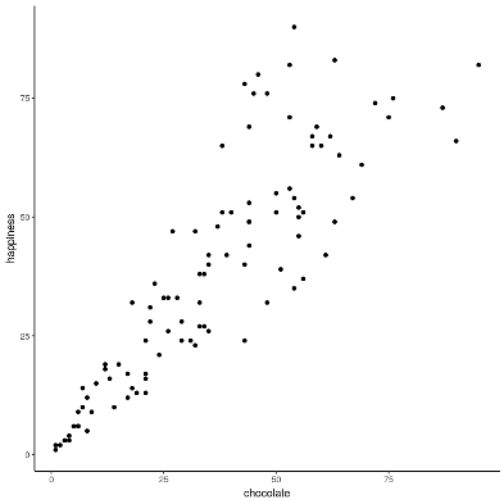


Figure 12.1.1: Imaginary data showing a positive correlation between amount of chocolate and amount happiness.

You might be wondering, why are there only 100 dots for the data. Didn't we collect 100 measures for chocolate, and 100 measures for happiness, shouldn't there be 200 dots? Nope. Each dot is for one subject, there are 100 subjects, so there are 100 dots.

What do the dots mean? Each dot has two coordinates, an x-coordinate for chocolate, and a y-coordinate for happiness. The first dot, all the way on the bottom left is the first subject in the table, who had close to 0 chocolate and close to zero happiness. You can look at any dot, then draw a straight line down to the x-axis: that will tell you how much chocolate that subject has. You can draw a straight line left to the y-axis: that will tell you how much happiness the subject has.

Now that we are looking at the scatter plot, we can see many things. The dots are scattered around a bit aren't they, hence **scatter plot**. Even when the dot's don't scatter, they're still called scatter plots, perhaps because those pesky dots in real life have so much scatter all the time. More important, the dots show a relationship between chocolate supply and happiness. Happiness is lower for people with smaller supplies of chocolate, and higher for people with larger supplies of chocolate. It looks like the more chocolate you have the happier you will be, and vice-versa. This kind of relationship is called a **positive correlation**.

Positive, Negative, and No-Correlation

Seeing as we are in the business of imagining data, let's imagine some more. We've already imagined what data would look like if larger chocolate supplies increase happiness. We'll show that again in a bit. What do you imagine the scatter plot would look like if the relationship was reversed, and larger chocolate supplies decreased happiness. Or, what do you imagine the scatter plot would look like if there was no relationship, and the amount of chocolate that you have doesn't do anything to your happiness. We invite your imagination to look at these graphs:

```
library(ggplot2)
subject_x<-1:100
chocolate_x<-round(1:100*runif(100,.5,1))
happiness_x<-round(1:100*runif(100,.5,1))
```

```
df_positive<-data.frame(subject_x,chocolate_x,happiness_x)
subject_x<-1:100
chocolate_x<-round(1:100*runif(100,.5,1))
happiness_x<-round(100:1*runif(100,.5,1))
df_negative<-data.frame(subject_x,chocolate_x,happiness_x)
subject_x<-1:100
chocolate_x<-round(runif(100,0,100))
happiness_x<-round(runif(100,0,100))
df_random<-data.frame(subject_x,chocolate_x,happiness_x)
all_data<-rbind(df_positive,df_negative,df_random)
all_data<-cbind(all_data,correlation=rep(c("positive","negative","random"),each=100))
ggplot(all_data,aes(x=chocolate_x,y=happiness_x))+
  geom_point()+
  theme_classic()+
  facet_wrap(~correlation)+
  xlab("chocolate supply")+
  ylab("happiness")
```

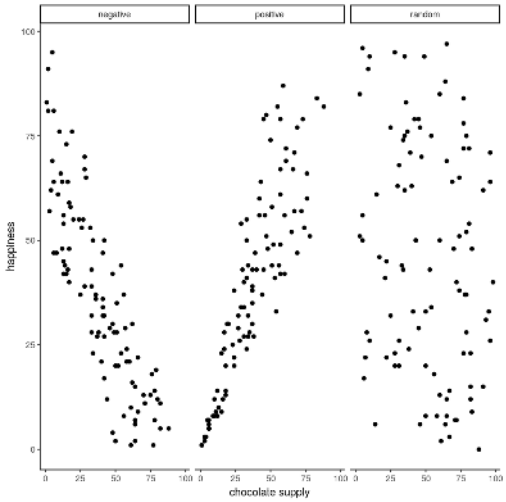


Figure 12.1.2: Three scatterplots showing negative, positive, and zero correlation.

The first panel shows a **negative correlation**. Happiness goes down as chocolate supply increases. Negative correlation occurs when one thing goes up and the other thing goes down; or, when more of X is less of Y, and vice-versa. The second panel shows a **positive correlation**. Happiness goes up as chocolate as chocolate supply increases. Positive correlation occurs when both things go up together, and go down together: more of X is more of Y, and vice-versa. The third panel shows **no correlation**. Here, there doesn't appear to be any obvious relationship between chocolate supply and happiness. The dots are scattered all over the place, the truest of the scatter plots.

Note

We are wading into the idea that measures of two things can be related, or correlated with one another. It is possible for the relationships to be more complicated than just going up, or going down. For example, we could have a relationship that where the dots go up for the first half of X, and then go down for the second half.

Zero correlation occurs when one thing is not related in any way to another things: changes in X do not relate to any changes in Y, and vice-versa.

This page titled 12.1: If something caused something else to change, what would that look like? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew J. C. Crump via source content that was edited to the style and standards of the LibreTexts platform.

- **3.1: If something caused something else to change, what would that look like?** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

12.2: Pearson's r

If Beyoncé was a statistician, she might look at these scatter plots and want to "put a number on it". We think this is a good idea too. We've already learned how to create descriptive statistics for a single measure, like chocolate, or happiness (i.e., means, variances, etc.). Is it possible to create a descriptive statistic that summarized the relationship between two measures, all in one number? Can it be done? Karl Pearson to the rescue.

> **Note**
>
> The stories about the invention of various statistics are very interesting, you can read more about them in the book, "The Lady Tasting Tea" (Salsburg 2001)

There's a statistic for that, and Karl Pearson invented it. Everyone now calls it, "Pearson's r". We will find out later that Karl Pearson was a big-wig editor at Biometrika in the 1930s. He took a hating to another big-wig statistician, Sir Ronald Fisher (who we learn about later), and they had some stats fights…why can't we all just get along in statistics.

How does Pearson's r work? Let's look again at the first 10 subjects in our fake experiment:

subject	chocolate	happiness
1	1	1
2	2	2
3	2	3
4	3	3
5	3	3
6	5	5
7	4	6
8	5	5
9	9	5
10	6	9
Sums	40	42
Means	4	4.2

What could we do to these numbers to produce a single summary value that represents the relationship between the chocolate supply and happiness?

The idea of co-variance

"Oh please no, don't use the word variance again". Yes, we're doing it, we're going to use the word variance again, and again, until it starts making sense. Remember what variance means about some numbers. It means the numbers have some change in them, they are not all the same, some of them are big, some are small. We can see that there is variance in chocolate supply across the 10 subjects. We can see that there is variance in happiness across the 10 subjects. We also saw in the scatter plot, that happiness increases as chocolate supply increases; which is a positive relationship, a positive correlation. What does this have to do with variance? Well, it means there is a relationship between the variance in chocolate supply, and the variance in happiness levels. The two measures vary together don't they? When we have two measures that vary together, they are like a happy couple who share their variance. This is what co-variance refers to, the idea that the pattern of varying numbers in one measure is shared by the pattern of varying numbers in another measure.

Co-variance is **very, very, very ,very** important. We suspect that the word co-variance is initially confusing, especially if you are not yet fully comfortable with the meaning of variance for a single measure. Nevertheless, we must proceed and use the idea of co-variance over and over again to firmly implant it into your statistical mind (we already said, but redundancy works, it's a thing).

> *Pro tip: Three-legged race is a metaphor for co-variance. Two people tie one leg to each other, then try to walk. It works when they co-vary their legs together (positive relationship). They can also co-vary in an unhelpful way, when one person tries to move forward exactly when the other person tries to move backward. This is still co-variance (negative relationship). Funny random walking happens when there is no co-variance. This means one person does whatever they want, and so does the other person. There is a lot of variance, but the variance is shared randomly, so it's just a bunch of legs moving around accomplishing nothing.*

> *Pro tip #2: Succesfully playing paddycake occurs when two people coordinate their actions so they have postively shared co-variance.*

This page titled 12.2: Pearson's r is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew J. C. Crump via source content that was edited to the style and standards of the LibreTexts platform.

- **3.2: Pearson's r** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

12.3: Turning the numbers into a measure of co-variance

"OK, so if you are saying that co-variance is just another word for correlation or relationship between two measures, I'm good with that. I suppose we would need some way to measure that." Correct, back to our table…notice anything new?

subject	chocolate	happiness	Chocolate_X_Happiness
1	1	1	1
2	2	2	4
3	2	3	6
4	3	3	9
5	3	3	9
6	5	5	25
7	4	6	24
8	5	5	25
9	9	5	45
10	6	9	54
Sums	40	42	202
Means	4	4.2	20.2

We've added a new column called "Chocolate_X_Happiness", which translates to Chocolate scores multiplied by Happiness scores. Each row in the new column, is the product, or multiplication of the chocolate and happiness score for that row. Yes, but why would we do this?

Last chapter we took you back to Elementary school and had you think about division. Now it's time to do the same thing with multiplication. We assume you know how that works. One number times another, means taking the first number, and adding it as many times as the second says to do,

$2 * 2 = 2 + 2 = 4$

$2 * 6 = 2 + 2 + 2 + 2 + 2 + 2 = 12$, or $6 + 6 = 12$, same thing.

Yes, you know all that. But, can you bend multiplication to your will, and make it do your bidding when need to solve a problem like summarizing co-variance? Multiplication is the droid you are looking for.

We know how to multiple numbers, and all we have to next is think about the consequences of multiplying sets of numbers together. For example, what happens when you multiply two small numbers together, compared to multiplying two big numbers together? The first product should be smaller than the second product right? How about things like multiplying a small number by a big number? Those products should be in between right?.

Then next step is to think about how the products of two measures sum together, depending on how they line up. Let's look at another table:

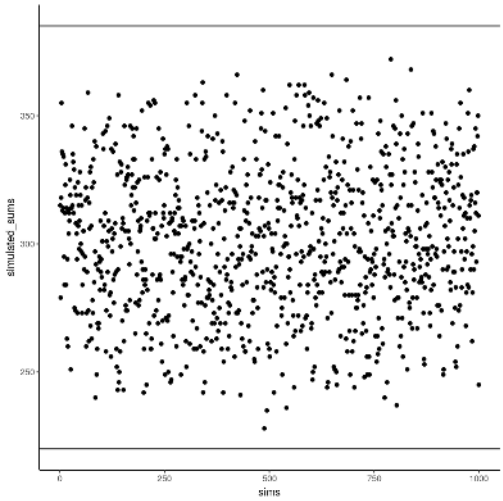
scores	X	Y	A	B	XY	AB
1	1	1	1	10	1	10
2	2	2	2	9	4	18
3	3	3	3	8	9	24
4	4	4	4	7	16	28
5	5	5	5	6	25	30
6	6	6	6	5	36	30
7	7	7	7	4	49	28
8	8	8	8	3	64	24
9	9	9	9	2	81	18
10	10	10	10	1	100	10
Sums	55	55	55	55	385	220
Means	5.5	5.5	5.5	5.5	38.5	22

Look at the X and Y column. The scores for X and Y perfectly co-vary. When X is 1, Y is 1; when X is 2, Y is 2, etc. They are perfectly aligned. The scores for A and B also perfectly co-vary, just in the opposite manner. When A is 1, B is 10; when A is 2, B is 9, etc. B is a reversed copy of A.

Now, look at the column ***XY***. These are the products we get when we multiply the values of X across with the values of Y. Also, look at the column ***AB***. These are the products we get when we multiply the values of A across with the values of B. So far so good.

Now, look at the `Sums` for the XY and AB columns. Not the same. The sum of the XY products is 385, and the sum of the AB products is 220. For this specific set of data, the numbers 385 and 220 are very important. They represent the biggest possible sum of products (385), and the smallest possible sum of products (220). There is no way of re-ordering the numbers 1 to 10, say for X, and the numbers 1 to 10 for Y, that would ever produce larger or smaller numbers. Don't believe me? Check this out:

```
library(ggplot2)
simulated_sums<-length(0)
for(sim in 1:1000){
        X<-sample(1:10)
        Y<-sample(1:10)
        simulated_sums[sim]<-sum(X*Y)
}
sim_df<-data.frame(sims=1:1000,simulated_sums)
ggplot(sim_df,aes(x=sims,y=simulated_sums))+
  geom_point()+
  theme_classic()+
  geom_hline(yintercept = 385)+
  geom_hline(yintercept = 220)
```



The above graph shows 1000 computer simulations. I convinced my computer to randomly order the numbers 1 to 10 for X, and randomly order the numbers 1 to 10 for Y. Then, I multiplied X and Y, and added the products together. I did this 1000 times. The dots show the sum of the products for each simulation. The two black lines show the maximum possible sum (385), and the minimum possible sum (220), for this set of numbers. Notice, how all of the dots are in between the maximum and minimum possible values. Told you so.

"OK fine, you told me so…So what, who cares?". We've been looking for a way to summarize the co-variance between two measures right? Well, for these numbers, we have found one, haven't we. It's the sum of the products. We know that when the sum of the products is 385, we have found a perfect, positive correlation. We know, that when the sum of the products is 220, we have found a perfect negative correlation. What about the numbers in between. What could we conclude about the correlation if we found the sum of the products to be 350. Well, it's going to be positive, because it's close to 385, and that's perfectly positive. If the sum of the products was 240, that's going to be negative, because it's close to the perfectly negatively correlating 220. What about no correlation? Well, that's going to be in the middle between 220 and 385 right.

We have just come up with a data-specific summary measure for the correlation between the numbers 1 to 10 in X, and the numbers 1 to 10 in Y, it's the sum of the products. We know the maximum (385) and minimum values (220), so we can now interpret any product sum for this kind of data with respect to that scale.

> *Pro tip: When the correlation between two measures increases in the positive direction, the sum of their products increases to its maximum possible value. This is because the bigger numbers in X will tend to line up with the bigger numbers in Y, creating the biggest possible sum of products. When the correlation between two measures increases in the negative direction, the sum of their products decreases to its minimum possible value. This is because the bigger numbers in X will tend to line up with the smaller numbers in Y, creating the smallest possible sum of products. When there is no correlation, the big numbers in X will be randomly lined up with the big and small numbers in Y, making the sum of the products, somewhere in the middle.*

Co-variance, the measure

We took some time to see what happens when you multiply sets of numbers together. We found that $big * big =$ **bigger** and $small * small =$ **still small**, and $big * small =$ **in the middle**. The purpose of this was to give you some conceptual idea of how the co-variance between two measures is reflected in the sum of their products. We did something very straightforward. We just multiplied X with Y, and looked at how the product sums get big and small, as X and Y co-vary in different ways.

Now, we can get a little bit more formal. In statistics, **co-variance** is not just the straight multiplication of values in X and Y. Instead, it's the multiplication of the deviations in X from the mean of X, and the deviation in Y from the mean of Y. Remember

those difference scores from the mean we talked about last chapter? They're coming back to haunt you know, but in a good way like Casper the friendly ghost.

Let's see what this look like in a table:

subject	chocolate	happiness	C_d	H_d	Cd_x_Hd
1	1	1	-3	-3.2	9.6
2	2	2	-2	-2.2	4.4
3	2	3	-2	-1.2	2.4
4	3	3	-1	-1.2	1.2
5	3	3	-1	-1.2	1.2
6	5	5	1	0.8	0.8
7	4	6	0	1.8	0
8	5	5	1	0.8	0.8
9	9	5	5	0.8	4
10	6	9	2	4.8	9.6
Sums	40	42	0	0	34
Means	4	4.2	0	0	3.4

We have computed the deviations from the mean for the chocolate scores (column `C_d`), and the deviations from the mean for the happiness scores (column `H_d`). Then, we multiplied them together (last column). Finally, you can see the mean of the products listed in the bottom right corner of the table, the official **the covariance**.

The formula for the co-variance is:

$$cov(X,Y) = \frac{\sum_i^n (x_i - \bar{X})(y_i - \bar{Y})}{N}$$

OK, so now we have a formal single number to calculate the relationship between two variables. This is great, it's what we've been looking for. However, there is a problem. Remember when we learned how to compute just the plain old **variance**. We looked at that number, and we didn't know what to make of it. It was squared, it wasn't in the same scale as the original data. So, we square rooted the **variance** to produce the **standard deviation**, which gave us a more interpretable number in the range of our data. The **co-variance** has a similar problem. When you calculate the co-variance as we just did, we don't know immediately know its scale. Is a 3 big? is a 6 big? is a 100 big? How big or small is this thing?

From our prelude discussion on the idea of co-variance, we learned the sum of products between two measures ranges between a maximum and minimum value. The same is true of the co-variance. For a given set of data, there is a maximum possible positive value for the co-variance (which occurs when there is perfect positive correlation). And, there is a minimum possible negative value for the co-variance (which occurs when there is a perfect negative correlation). When there is zero co-variation, guess what happens. Zeroes. So, at the very least, when we look at a co-variation statistic, we can see what direction it points, positive or negative. But, we don't know how big or small it is compared to the maximum or minimum possible value, so we don't know the relative size, which means we can't say how strong the correlation is. What to do?

Pearson's r we there yet

Yes, we are here now. Wouldn't it be nice if we could force our measure of co-variation to be between -1 and +1?

-1 would be the minimum possible value for a perfect negative correlation. +1 would be the maximum possible value for a perfect positive correlation. 0 would mean no correlation. Everything in between 0 and -1 would be increasingly large negative

correlations. Everything between 0 and +1 would be increasingly large positive correlations. It would be a fantastic, sensible, easy to interpret system. If only we could force the co-variation number to be between -1 and 1. Fortunately, for us, this episode is brought to you by Pearson's r, which does precisely this wonderful thing.

Let's take a look at a formula for Pearson's r:

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{cov(X,Y)}{SD_X SD_Y}$$

We see the symbol σ here, that's more Greek for you. σ is often used as a symbol for the standard deviation (SD). If we read out the formula in English, we see that r is the co-variance of X and Y, divided by the product of the standard deviation of X and the standard deviation of Y. Why are we dividing the co-variance by the product of the standard deviations. This operation has the effect of **normalizing** the co-variance into the range -1 to 1.

Note

But, we will fill this part in as soon as we can…promissory note to explain the magic. FYI, it's not magic. Brief explanation here is that dividing each measure by its standard deviation ensures that the values in each measure are in the same range as one another.

For now, we will call this mathematical magic. It works, but we don't have space to tell you why it works right now.

> *It's worth saying that there are loads of different formulas for computing Pearson's r. You can find them by Googling them. We will probably include more of them here, when we get around to it. However, they all give you the same answer. And, they are all not as pretty as each other. Some of them might even look scary. In other statistics textbook you will often find formulas that are easier to use for calculation purposes. For example, if you only had a pen and paper, you might use one or another formula because it helps you compute the answer faster by hand. To be honest, we are not very interested in teaching you how to plug numbers into formulas. We give one lesson on that here: Put the numbers into the letters, then compute the answer. Sorry to be snarky. Nowadays you have a computer that you should use for this kind of stuff. So, we are more interested in teaching you what the calculations mean, rather than how to do them. Of course, every week we are showing you how to do the calculations in lab with computers, because that is important to.*

Does Pearson's r really stay between -1 and 1 no matter what? It's true, take a look at the following simulation. Here I randomly ordered the numbers 1 to 10 for an X measure, and did the same for a Y measure. Then, I computed Pearson's r, and repeated this process 1000 times. As you can see all of the dots are between -1 and 1. Neat huh.

```
library(ggplot2)
simulated_sums <- length(0)
for(sim in 1:1000){
  X <- sample(1:10)
  Y <- sample(1:10)
  simulated_sums[sim] <- cor(X,Y)
}
sim_df <- data.frame(sims=1:1000,simulated_sums)
ggplot(sim_df, aes(x = sims, y = simulated_sums))+
  geom_point()+
```

```
  theme_classic()+
  geom_hline(yintercept = -1)+
  geom_hline(yintercept = 1)+
  ggtitle("Simulation of 1000 r values")
```

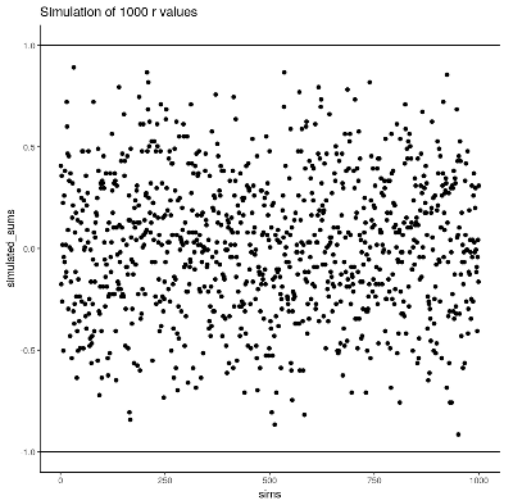


Figure **12.3.1**: A simulation of of correlations. Each dot represents the r-value for the correlation between an X and Y variable that each contain the numbers 1 to 10 in random orders. The figure ilustrates that many r-values can be obtained by this random process.

This page titled 12.3: Turning the numbers into a measure of co-variance is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew J. C. Crump via source content that was edited to the style and standards of the LibreTexts platform.

- **3.3: Turning the numbers into a measure of co-variance** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

12.4: Examples with Data

In the lab for correlation you will be shown how to compute correlations in real data-sets using software. To give you a brief preview, let's look at some data from the world happiness report (2018). This report measured various attitudes across people from different countries. For example, one question asked about how much freedom people thought they had to make life choices. Another question asked how confident people were in their national government. Here is a scatterplot showing the relationship between these two measures. Each dot represents means for different countries.

```
library(data.table)
library(ggplot2)
suppressPackageStartupMessages(library(dplyr))
whr_data <- fread("https://stats.libretexts.org/@api/deki/files/10477/WHR2018.csv")
# select DVs and filter for NAs
smaller_df <- whr_data %>%
               dplyr::select(country,
                      `Freedom to make life choices`,
                      `Confidence in national government`) %>%
               dplyr::filter(!is.na(`Freedom to make life choices`),
                      !is.na(`Confidence in national government`))
# plot the data with best fit line
ggplot(smaller_df, aes(x=`Freedom to make life choices`,
                     y=`Confidence in national government`))+
  geom_point(alpha=.5)+
  geom_smooth(method=lm, se=FALSE, formula=y ~ x)+
  theme_classic()
```

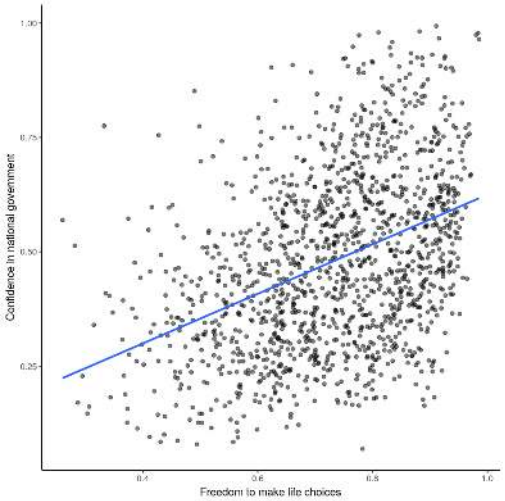


Figure 12.4.1: Relationship between freedom to make life choices and confidence in national government. Data from the world happiness report for 2018.

We put a blue line on the scatterplot to summarize the positive relationship. It appears that as “freedom to make life choices goes up”, so to does confidence in national government. It’s a positive correlation.

The actual correlation, as measured by Pearson’s r is:

```
library(data.table)
suppressPackageStartupMessages(library(dplyr))
```

```
whr_data <- fread("https://stats.libretexts.org/@api/deki/files/10477/WHR2018.csv")
# select DVs and filter for NAs
smaller_df <- whr_data %>%
              dplyr::select(country,
                     `Freedom to make life choices`,
                     `Confidence in national government`) %>%
              dplyr::filter(!is.na(`Freedom to make life choices`),
                     !is.na(`Confidence in national government`))
# calculate correlation
cor(smaller_df$`Freedom to make life choices`,
    smaller_df$`Confidence in national government`)
```

0.408096292505333

You will do a lot more of this kind of thing in the lab. Looking at the graph you might start to wonder: Does freedom to make life choices cause changes how confident people are in their national government? Our does it work the other way? Does being confident in your national government give you a greater sense of freedom to make life choices? Or, is this just a random relationship that doesn't mean anything? All good questions. These data do not provide the answers, they just suggest a possible relationship.

This page titled 12.4: Examples with Data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew J. C. Crump via source content that was edited to the style and standards of the LibreTexts platform.

- **3.4: Examples with Data** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

12.5: Regression — A mini intro

We're going to spend the next little bit adding one more thing to our understanding of correlation. It's called **linear regression**. It sounds scary, and it really is. You'll find out much later in your Statistics education that everything we will be soon be talking about can be thought of as a special case of regression. But, we don't want to scare you off, so right now we just introduce the basic concepts.

First, let's look at a linear regression. This way we can see what we're trying to learn about. Here's some scatter plots, same one's you've already seen. But, we've added something new! Lines.

```
library(ggplot2)
subject_x<-1:100
chocolate_x<-round(1:100*runif(100,.5,1))
happiness_x<-round(1:100*runif(100,.5,1))
df_positive<-data.frame(subject_x,chocolate_x,happiness_x)
subject_x<-1:100
chocolate_x<-round(1:100*runif(100,.5,1))
happiness_x<-round(100:1*runif(100,.5,1))
df_negative<-data.frame(subject_x,chocolate_x,happiness_x)
subject_x<-1:100
chocolate_x<-round(runif(100,0,100))
happiness_x<-round(runif(100,0,100))
df_random<-data.frame(subject_x,chocolate_x,happiness_x)
all_data<-rbind(df_positive,df_negative,df_random)
all_data<-cbind(all_data,correlation=rep(c("positive","negative","random"),each=100))
ggplot(all_data,aes(x=chocolate_x,y=happiness_x))+
  geom_point()+
  theme_classic()+
  geom_smooth(method="lm",se=F,formula=y ~ x)+
  facet_wrap(~correlation)+
  xlab("chocolate supply")+
  ylab("happiness")
```

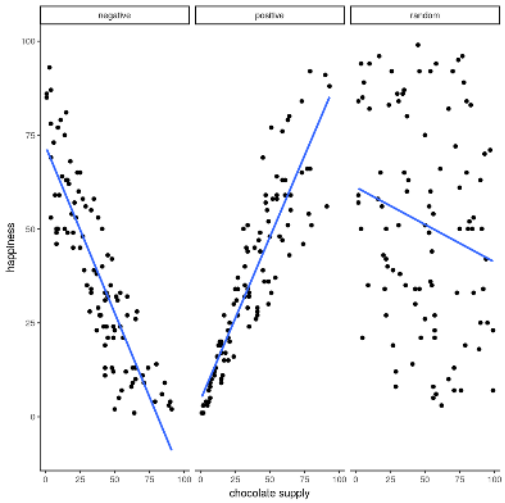


Figure 12.5.1: Three scatterplots showing negative, positive, and a random correlation (where the r-value is expected to be 0), along with the best fit regression line

The best fit line

Notice anything about these blue lines? Hopefully you can see, at least for the first two panels, that they go straight through the data, just like a kebab skewer. We call these lines **best fit** lines, because according to our definition (soon we promise) there are no other lines that you could draw that would do a better job of going straight throw the data.

One big idea here is that we are using the line as a kind of mean to describe the relationship between the two variables. When we only have one variable, that variable exists on a single dimension, it's 1D. So, it is appropriate that we only have one number, like the mean, to describe it's central tendency. When we have two variables, and plot them together, we now have a two-dimensional space. So, for two dimensions we could use a bigger thing that is 2d, like a line, to summarize the central tendency of the relationship between the two variables.

What do we want out of our line? Well, if you had a pencil, and a printout of the data, you could draw all sorts of straight lines any way you wanted. Your lines wouldn't even have to go through the data, or they could slant through the data with all sorts of angles. Would all of those lines be very good a describing the general pattern of the dots? Most of them would not. The best lines would go through the data following the general shape of the dots. Of the best lines, however, which one is the best? How can we find out, and what do we mean by that? In short, the best fit line is the one that has the least error.

> **Note**
>
> R code for plotting residuals thanks to Simon Jackson's blog post: https://drsimonj.svbtle.com/visualising-residuals

Check out this next plot, it shows a line through some dots. But, it also shows some teeny tiny lines. These lines drop down from each dot, and they land on the line. Each of these little lines is called a **residual**. They show you how far off the line is for different dots. It's measure of error, it shows us just how wrong the line is. After all, it's pretty obvious that not all of the dots are on the line. This means the line does not actually represent all of the dots. The line is wrong. But, the best fit line is the least wrong of all the wrong lines.

```
library(ggplot2)
d <- mtcars
fit <- lm(mpg ~ hp, data = d)
d$predicted <- predict(fit)   # Save the predicted values
d$residuals <- residuals(fit) # Save the residual values
ggplot(d, aes(x = hp, y = mpg)) +
  geom_smooth(method = "lm", se = FALSE,
              color = "lightblue", formula=y ~ x) +  # Plot regression slope
  geom_segment(aes(xend = hp, yend = predicted,
                   color="red"), alpha = .5) +  # alpha to fade lines
  geom_point() +
  geom_point(aes(y = predicted), shape = 1) +
  theme_classic()+
  theme(legend.position="none")+
  xlab("X")+ylab("Y")
# Quick look at the actual, predicted, and residual values
#library(dplyr)
#d %>% select(mpg, predicted, residuals) %>% head()
```

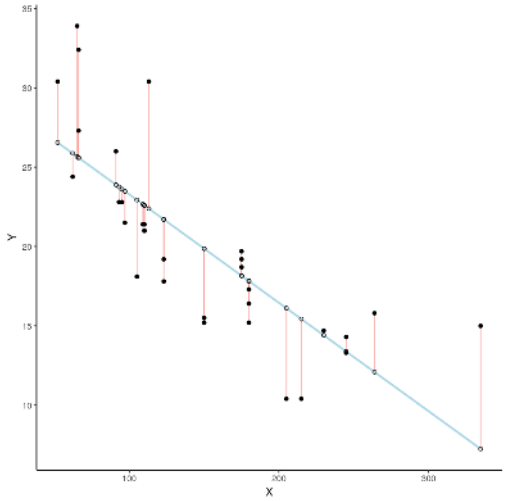


Figure 12.5.2: Black dots represent data points. The blue line is the best fit regression line. The white dots are repesent the predicted location of each black dot. The red lines show the error between each black dot and the regression line. The blue line is the best fit line because it minimizes the error shown by the red lines.

There's a lot going on in this graph. First, we are looking at a scatter plot of two variables, an X and Y variable. Each of the black dots are the actual values from these variables. You can see there is a negative correlation here, as X increases, Y tends to decrease. We drew a regression line through the data, that's the blue line. There's these little white dots too. This is where the line thinks the black dots should be. The red lines are the important residuals we've been talking about. Each black dot has a red line that drops straight down, or straight up from the location of the black dot, and lands directly on the line. We can already see that many of the dots are not on the line, so we already know the line is "off" by some amount for each dot. The red line just makes it easier to see exactly how off the line is.

The important thing that is happening here, is that the the blue line is drawn is such a way, that it minimizes the total length of the red lines. For example, if we wanted to know how wrong this line was, we could simply gather up all the red lines, measure how long they are, and then add all the wrongness together. This would give us the total amount of wrongness. We usually call this the error. In fact, we've already talked about this idea before when we discussed standard deviation. What we will actually be doing with the red lines, is computing the sum of the squared deviations from the line. That sum is the total amount of error. Now, this blue line here minimizes the sum of the squared deviations. Any other line would produce a larger total error.

Here's an animation to see this in action. The animations compares the best fit line in blue, to some other possible lines in black. The black line moves up and down. The red lines show the error between the black line and the data points. As the black line moves toward the best fit line, the total error, depicted visually by the grey area shrinks to it's minimum value. The total error expands as the black line moves away from the best fit line.

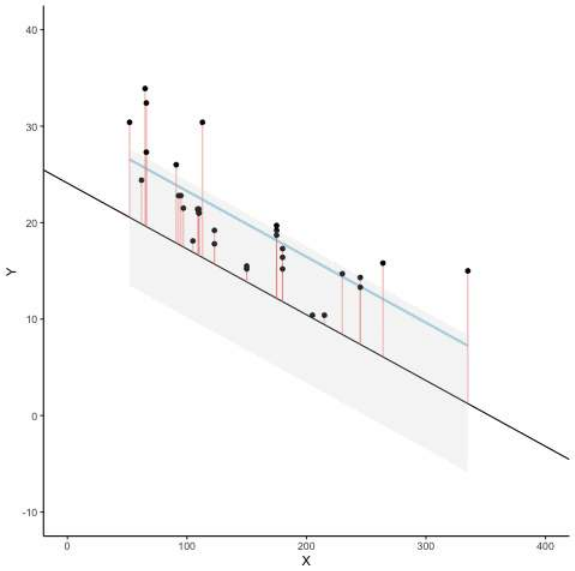


Figure **12.5.3**: The blue line is the best fit regression line explaining the co-variation among the black dots. The black line moves up and down showing alternative lines that could be drawn. The red lines show the amount of error between each data point and the black line. The total amount of error is depicted by the shaded grey area. The size of the grey area expands as the black line moves away from the best fit line, and shrinks to a minimum as the black line moves toward the best fit line.

Whenever the black line does not overlap with the blue line, it is worse than the best fit line. The blue regression line is like Goldilocks, it's just right, and it's in the middle.

This next graph shows a little simulation of how the sum of squared deviations (the sum of the squared lengths of the red lines) behaves as we move the line up and down. What's going on here is that we are computing a measure of the total error as the black line moves through the best fit line. This represents the sum of the squared deviations. In other words, we square the length of each red line from the above animation, then we add up all of the squared red lines, and get the total error (the total sum of the squared deviations). The graph below shows what the total error looks like as the black line approaches then moves away from the best fit line. Notice, the dots in this graph start high on the left side, then they swoop down to a minimum at the bottom middle of the graph. When they reach their minimum point, we have found a line that minimizes the total error. This is the best fit regression line.

```
library(ggplot2)
d <- mtcars
fit <- lm(mpg ~ hp, data = d)
d$predicted <- predict(fit)   # Save the predicted values
d$residuals <- residuals(fit) # Save the residual values
coefs<-coef(lm(mpg ~ hp, data = mtcars))
#coefs[1]
#coefs[2]
x<-d$hp
move_line<-seq(-5,5,.5)
total_error<-c(length(move_line))
cnt<-0
for(i in move_line){
  cnt<-cnt+1
  predicted_y <- coefs[2]*x + coefs[1]+i
  error_y <- (predicted_y-d$mpg)^2
  total_error[cnt]<-sum(error_y)
}
```

```
df<-data.frame(move_line,total_error)
ggplot(df,aes(x=move_line,y=total_error))+
  geom_point()+
  theme_classic()+
  ylab("sum of squared deviations")+
  xlab("change to y-intercept")
```

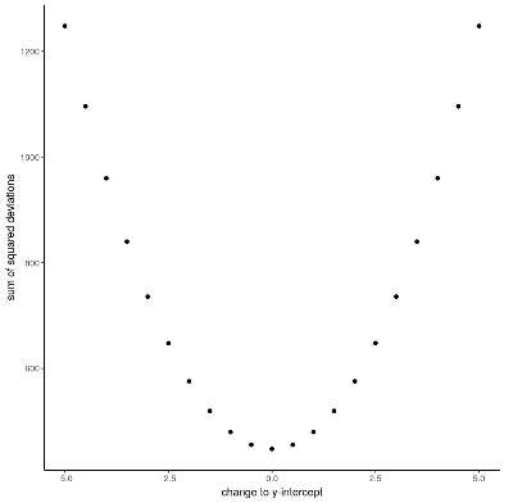


Figure 12.5.4: A plot of the sum of the squared deviations for different lines moving up and down, through the best fit line. The best fit line occurs at the position that minimizes the sum of the sqaured deviations.

OK, so we haven't talked about the y-intercept yet. But, what this graph shows us is how the total error behaves as we move the line up and down. The y-intercept here is the thing we change that makes our line move up and down. As you can see the dots go up when we move the line down from 0 to -5, and the dots go up when we move the line up from 0 to +5. The best line, that minimizes the error occurs right in the middle, when we don't move the blue regression line at all.

Lines

OK, fine you say. So, there is one magic line that will go through the middle of the scatter plot and minimize the sum of the squared deviations. How do I find this magic line? We'll show you. But, to be completely honest, you'll almost never do it the way we'll show you here. Instead, it's much easier to use software and make your computer do it for. You'll learn how to that in the labs.

Before we show you how to find the regression line, it's worth refreshing your memory about how lines work, especially in 2 dimensions. Remember this?

$y = ax + b$, or also $y = mx + b$ (sometimes a or m is used for the slope)

This is the formula for a line. Another way of writing it is:

$$y = slope * x + \text{y-intercept}$$

The slope is the slant of the line, and the y-intercept is where the line crosses the y-axis. Let's look at some lines:

```
library(ggplot2)
ggplot()+
  geom_abline(slope=1,intercept=5,color="blue")+
  geom_abline(slope=-1, intercept=15, color="red")+
  lims(x = c(1,20), y = c(0,20))+
  theme_classic()
```

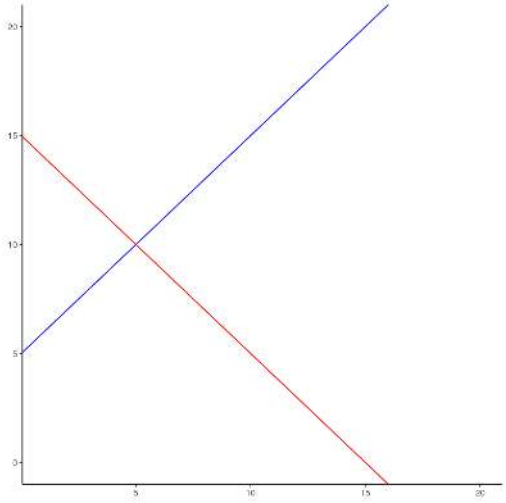


Figure 12.5.5: Two different lines with different y-intercepts (where the line crosses the y-axis), and different slopes. A positive slope makes the line go up from left to right. A negative slope makes the line go down from left to right.

So there is two lines. The formula for the blue line is $y = 1 * x + 5$. Let's talk about that. When x = 0, where is the blue line on the y-axis? It's at five. That happens because 1 times 0 is 0, and then we just have the five left over. How about when x = 5? In that case y =10. You just need the plug in the numbers to the formula, like this:

$$y = 1 * x + 5$$

$$y = 1 * 5 + 5 = 5 + 5 = 10$$

The point of the formula is to tell you where y will be, for any number of x. The slope of the line tells you whether the line is going to go up or down, as you move from the left to the right. The blue line has a positive slope of one, so it goes up as x goes up. How much does it go up? It goes up by one for everyone one of x! If we made the slope a 2, it would be much steeper, and go up faster. The red line has a negative slope, so it slants down. This means y goes down, as x goes up. When there is no slant, and we want to make a perfectly flat line, we set the slope to 0. This means that y doesn't go anywhere as x gets bigger and smaller.

That's lines.

Computing the best fit line

If you have a scatter plot showing the locations of scores from two variables, the real question is how can you find the slope and the y-intercept for the best fit line? What are you going to do? Draw millions of lines, add up the residuals, and then see which one was best? That would take forever. Fortunately, there are computers, and when you don't have one around, there's also some handy formulas.

Note

It's worth pointing out just how much computers have changed everything. Before computers everyone had to do these calculations by hand, such a chore! Aside from the deeper mathematical ideas in the formulas, many of them were made for convenience, to speed up hand calculations, because there were no computers. Now that we have computers, the hand calculations are often just an exercise in algebra. Perhaps they build character. You decide.

We'll show you the formulas. And, work through one example by hand. It's the worst, we know. By the way, you should feel sorry for me as I do this entire thing by hand for you.

Here are two formulas we can use to calculate the slope and the intercept, straight from the data. We won't go into why these formulas do what they do. These ones are for "easy" calculation.

$$intercept = b = \frac{\sum y \sum x^2 - \sum x \sum xy}{n \sum x^2 - (\sum x)^2}$$

$$slope = m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

In these formulas, the x and the y refer to the individual scores. Here's a table showing you how everything fits together.

```
suppressPackageStartupMessages(library(dplyr))
scores<-c(1,2,3,4,5,6,7)
x<-c(1,4,3,6,5,7,8)
y<-c(2,5,1,8,6,8,9)
x_squared<-x^2
y_squared<-y^2
xy<-x*y
all_df<-data.frame(scores,x,y,x_squared,y_squared,xy)
all_df <- all_df %>%
  rbind(c("Sums",colSums(all_df[1:7,2:6])))
slope=((sum(y)*sum(x_squared))-(sum(x)*sum(xy)))/((7*sum(x_squared))-sum(x)^2)
intercept=(7*sum(xy)-sum(x)*sum(y))/(7*sum(x_squared)-sum(x)^2)
knitr::kable(all_df)
```

scores	x	y	x_squared	y_squared	xy
1	1	2	1	4	2
2	4	5	16	25	20
3	3	1	9	1	3
4	6	8	36	64	48
5	5	6	25	36	30
6	7	8	49	64	56
7	8	9	64	81	72
Sums	34	39	200	275	231

We see 7 sets of scores for the x and y variable. We calculated x^2 by squaring each value of x, and putting it in a column. We calculated y^2 by squaring each value of y, and putting it in a column. Then we calculated xy, by multiplying each x score with each y score, and put that in a column. Then we added all the columns up, and put the sums at the bottom. These are all the number we need for the formulas to find the best fit line. Here's what the formulas look like when we put numbers in them:

$$intercept = b = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2} = \frac{39 * 200 - 34 * 231}{7 * 200 - 34^2} = -.221$$

$$slope = m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{7 * 231 - 34 * 39}{7 * 275 - 34^2} = 1.19$$

Great, now we can check our work, let's plot the scores in a scatter plot and draw a line through it with slope = 1.19, and a y-intercept of -.221. It should go through the middle of the dots.

```
library(ggplot2)
x<-c(1,4,3,6,5,7,8)
y<-c(2,5,1,8,6,8,9)
plot_df<-data.frame(x,y)
```

```
ggplot(plot_df,aes(x=x,y=y))+
  geom_point()+
  geom_smooth(method="lm",se=FALSE, formula=y ~ x)+
  theme_classic()
```

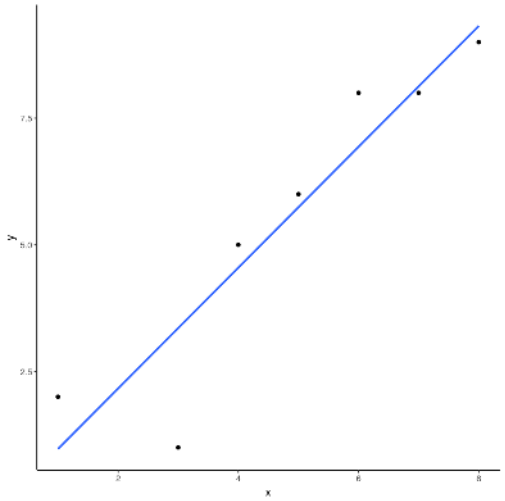


Figure **12.5.6**: An example regression line with confidence bands going through a few data points in a scatterplot.

```
x<-c(1,4,3,6,5,7,8)
y<-c(2,5,1,8,6,8,9)
plot_df<-data.frame(x,y)
coef(lm(y~x,plot_df))
```

(Intercept): -0.221311475409834 **x:** 1.19262295081967

This page titled 12.5: Regression — A mini intro is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew J. C. Crump via source content that was edited to the style and standards of the LibreTexts platform.

- **3.5: Regression — A mini intro** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

12.6: Interpreting Correlations

What does the presence or the absence of a correlation between two measures mean? How should correlations be interpreted? What kind of inferences can be drawn from correlations? These are all very good questions. A first piece of advice is to use caution when interpreting correlations. Here's why.

Correlation does not equal causation

Perhaps you have heard that correlation does not equal causation. Why not? There are lots of reasons why not. However, before listing some of the reasons let's start with a case where we would expect a causal connection between two measurements. Consider, buying a snake plant for your home. Snake plants are supposed to be easy to take care of because you can mostly ignore them.

Like most plants, snake plants need some water to stay alive. However, they also need just the right amount of water. Imagine an experiment where 1000 snake plants were grown in a house. Each snake plant is given a different amount of water per day, from zero teaspoons of water per day to 1000 teaspoons of water per day. We will assume that water is part of the causal process that allows snake plants to grow. The amount of water given to each snake plant per day can also be one of our measures. Imagine further that every week the experimenter measures snake plant growth, which will be the second measurement. Now, can you imagine for yourself what a scatter plot of weekly snake plant growth by tablespoons of water would look like?

Even when there is causation, there might not be obvious correlation

The first plant given no water at all would have a very hard time and eventually die. It should have the least amount of weekly growth. How about the plants given only a few teaspoons of water per day. This could be just enough water to keep the plants alive, so they will grow a little bit but not a lot. If you are imagining a scatter plot, with each dot being a snake plant, then you should imagine some dots starting in the bottom left hand corner (no water & no plant growth), moving up and to the right (a bit of water, and a bit of growth). As we look at snake plants getting more and more water, we should see more and more plant growth, right? "Sure, but only up to a point". Correct, there should be a trend for a positive correlation with increasing plant growth as amount of water per day increases. But, what happens when you give snake plants too much water? From personal experience, they die. So, at some point, the dots in the scatter plot will start moving back down again. Snake plants that get way too much water will not grow very well.

The imaginary scatter plot you should be envisioning could have an upside U shape. Going from left to right, the dot's go up, they reach a maximum, then they go down again reaching a minimum. Computing Pearson's r for data like this can give you r values close to zero. The scatter plot could look something like this:

```
library(ggplot2)
water<-seq(0,999,1)
growth<-c(seq(0,10,(10/499)),seq(10,0,-(10/499)))
noise<-runif(1000,-2,2)
growth<-growth+noise
snake_df<-data.frame(growth,water)
ggplot(snake_df, aes(x=water,y=growth))+
  geom_point()+
  theme_classic()+
  xlab("Water (teaspoons)")+
  ggtitle("Imaginary snake plant growth \n as a function of water")
```

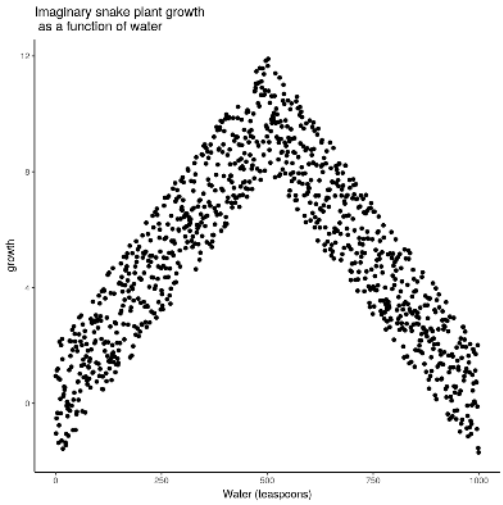


Figure 12.6.1: Illustration of a possible relationship between amount of water and snake plant growth. Growth goes up with water, but eventually goes back down as too much water makes snake plants die.

Granted this looks more like an inverted V, than an inverted U, but you get the picture right? There is clearly a relationship between watering and snake plant growth. But, the correlation isn't in one direction. As a result, when we compute the correlation in terms of Pearson's r, we get a value suggesting no relationship.

```
water<-seq(0,999,1)
growth<-c(seq(0,10,(10/499)),seq(10,0,-(10/499)))
noise<-runif(1000,-2,2)
growth<-growth+noise
cor(growth,water)
```

-0.0051489425461363

What this really means is there is no linear relationship that can be described by a single straight line. When we need lines or curves going in more than one direction, we have a nonlinear relationship.

This example illustrates some conundrums in interpreting correlations. We already know that water is needed for plants to grow, so we are rightly expecting there to be a relationship between our measure of amount of water and plant growth. If we look at the first half of the data we see a positive correlation, if we look at the last half of the data we see a negative correlation, and if we look at all of the data we see no correlation. Yikes. So, even when there is a causal connection between two measures, we won't necessarily obtain clear evidence of the connection just by computing a correlation coefficient.

> *Pro Tip: This is one reason why plotting your data is so important. If you see an upside U shape pattern, then a correlation analysis is probably not the best analysis for your data.*

Confounding variable, or Third variable problem

Anybody can correlate any two things that can be quantified and measured. For example, we could find a hundred people, ask them all sorts of questions like:

1. how happy are you
2. how old are you
3. how tall are you
4. how much money do you make per year
5. how long are your eyelashes
6. how many books have you read in your life
7. how loud is your inner voice

Let's say we found a positive correlation between yearly salary and happiness. Note, we could have just as easily computed the same correlation between happiness and yearly salary. If we found a correlation, would you be willing to infer that yearly salary causes happiness? Perhaps it does play a small part. But, something like happiness probably has a lot of contributing causes. Money could directly cause some people to be happy. But, more likely, money buys people access to all sorts of things, and some of those things might contribute happiness. These "other" things are called **third** variables. For example, perhaps people living in nicer places in more expensive houses are more happy than people in worse places in cheaper houses. In this scenario, money isn't causing happiness, it's the places and houses that money buys. But, even is this were true, people can still be more or less happy in lots of different situations.

The lesson here is that a correlation can occur between two measures because of a third variable that is not directly measured. So, just because we find a correlation, does not mean we can conclude anything about a causal connection between two measurements.

Correlation and Random chance

Another very important aspect of correlations is the fact that they can be produced by random chance. This means that you can find a positive or negative correlation between two measures, even when they have absolutely nothing to do with one another. You might have hoped to find zero correlation when two measures are totally unrelated to each other. Although this certainly happens, unrelated measures can accidentally produce **spurious** correlations, just by chance alone.

Let's demonstrate how correlations can occur by chance when there is no causal connection between two measures. Imagine two participants. One is at the North pole with a lottery machine full of balls with numbers from 1 to 10. The other is at the south pole with a different lottery machine full of balls with numbers from 1 to 10. There are an endless supply of balls in the machine, so every number could be picked for any ball. Each participant randomly chooses 10 balls, then records the number on the ball. In this situation we will assume that there is no possible way that balls chosen by the first participant could causally influence the balls chosen by the second participant. They are on the other side of the world. We should assume that the balls will be chosen by chance alone.

Here is what the numbers on each ball could look like for each participant:

```
Ball<-1:10
North_pole<-round(round(runif(10,1,10)))
South_pole<-round(round(runif(10,1,10)))
the_df_balls<-data.frame(Ball,North_pole,South_pole)
#the_df_balls <- the_df_balls %>%
#  rbind(c("Sums",colSums(the_df_balls[1:10,2:3]))) %>%
#  rbind(c("Means",colMeans(the_df_balls[1:10,2:3])))
knitr::kable(the_df_balls)
```

Ball	North_pole	South_pole
1	3	1
2	7	7
3	8	8
4	6	9
5	4	6
6	10	10
7	2	2
8	5	8
9	2	5
10	2	3

In this one case, if we computed Pearson's r, we would find that $r =$

```
North_pole<-round(round(runif(10,1,10)))
South_pole<-round(round(runif(10,1,10)))
cor(North_pole,South_pole)
```

0.0803444730711034

But, we already know that this value does not tell us anything about the relationship between the balls chosen in the north and south pole. We know that relationship should be completely random, because that is how we set up the game.

The better question here is to ask what can random chance do? For example, if we ran our game over and over again thousands of times, each time choosing new balls, and each time computing the correlation, what would we find? First, we will find fluctuation. The r value will sometimes be positive, sometimes be negative, sometimes be big and sometimes be small. Second, we will see what the fluctuation looks like. This will give us a window into the kinds of correlations that chance alone can produce. Let's see what happens.

Monte-carlo simulation of random correlations

It is possible to use a computer to simulate our game as many times as we want. This process is often termed **monte-carlo simulation**.

Below is a script written for the programming language R. We won't go into the details of the code here. However, let's briefly explain what is going on. Notice, the part that says `for(sim in 1:1000)`. This creates a loop that repeats our game 1000 times. Inside the loop there are variables named `North_pole` and `South_pole`. During each simulation, we sample 10 random numbers (between 1 to 10) into each variable. These random numbers stand for the numbers that would have been on the balls from the lottery machine. Once we have 10 random numbers for each, we then compute the correlation using `cor(North_pole,South_pole)`. Then, we save the correlation value and move on to the next simulation. At the end, we will have 1000 individual Pearson r values.

```
library(ggplot2)
simulated_correlations <- length(0)
for(sim in 1:1000){
  North_pole <- runif(10,1,10)
  South_pole <- runif(10,1,10)
```

```
  simulated_correlations[sim] <- cor(North_pole,South_pole)
}
sim_df <- data.frame(sims=1:1000,simulated_correlations)
ggplot(sim_df, aes(x = sims, y = simulated_correlations))+
  geom_point()+
  theme_classic()+
  geom_hline(yintercept = -1)+
  geom_hline(yintercept = 1)+
  ggtitle("Simulation of 1000 r values")
```

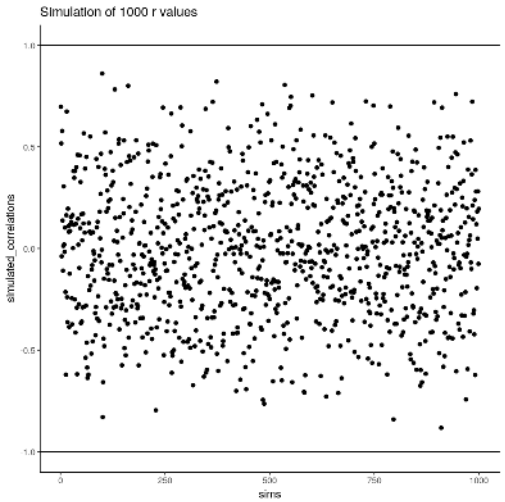


Figure 12.6.2: Another figure showing a range of r-values that can be obtained by chance.

Let's take a look at all of the 1000 Pearson r values. Does the figure below look familiar to you? It should, we have already conducted a similar kind of simulation before. Each dot in the scatter plot shows the Pearson r for each simulation from 1 to 1000. As you can see the dots are all over of the place, in between the range -1 to 1. The important lesson here is that random chance produced all of these correlations. This means we can find "correlations" in the data that are completely meaningless, and do not reflect any causal relationship between one measure and another.

Let's illustrate the idea of finding "random" correlations one more time, with a little movie. This time, we will show you a scatter plot of the random values sampled for the balls chosen from the North and South pole. If there is no relationship we should see dots going everywhere. If there happens to be a positive relationship (purely by chance), we should see the dots going from the bottom left to the top right. If there happens to be a negative relationship (purely by chance), we should see the dots going from the top left down to the bottom right.

On more thing to prepare you for the movie. There are three scatter plots below, showing negative, positive, and zero correlations between two variables. You've already seen this graph before. We are just reminding you that the blue lines are helpful for seeing the correlation.Negative correlations occur when a line goes down from the top left to bottom right. Positive correlations occur when a line goes up from the bottom left to the top right. Zero correlations occur when the line is flat (doesn't go up or down).

```
library(ggplot2)
subject_x<-1:100
chocolate_x<-round(1:100*runif(100,.5,1))
happiness_x<-round(1:100*runif(100,.5,1))
df_positive<-data.frame(subject_x,chocolate_x,happiness_x)
subject_x<-1:100
chocolate_x<-round(1:100*runif(100,.5,1))
```

```
happiness_x<-round(100:1*runif(100,.5,1))
df_negative<-data.frame(subject_x,chocolate_x,happiness_x)
subject_x<-1:100
chocolate_x<-round(runif(100,0,100))
happiness_x<-round(runif(100,0,100))
df_random<-data.frame(subject_x,chocolate_x,happiness_x)
all_data<-rbind(df_positive,df_negative,df_random)
all_data<-cbind(all_data,correlation=rep(c("positive","negative","random"),each=100))
ggplot(all_data,aes(x=chocolate_x,y=happiness_x))+
  geom_point()+
  geom_smooth(method=lm,se=FALSE, formula=y ~ x)+
  theme_classic()+
  facet_wrap(~correlation)+
  xlab("chocolate supply")+
  ylab("happiness")
```

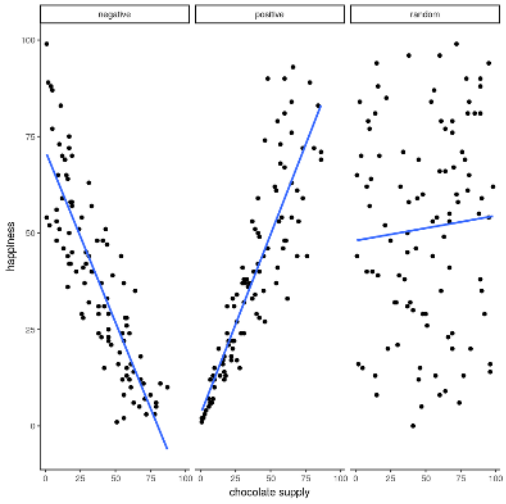


Figure **12.6.3**: A reminder of what positive, negative, and zero correlation looks like.

OK, now we are ready for the movie. You are looking at the process of sampling two sets of numbers randomly, one for the X variable, and one for the Y variable. Each time we sample 10 numbers for each, plot them, then draw a line through them. Remember, these numbers are all completely random, so we should expect, on average that there should be no correlation between the numbers. However, this is not what happens. You can the line going all over the place. Sometimes we find a negative correlation (line goes down), sometimes we see a positive correlation (line goes up), and sometimes it looks like zero correlation (line is more flat).

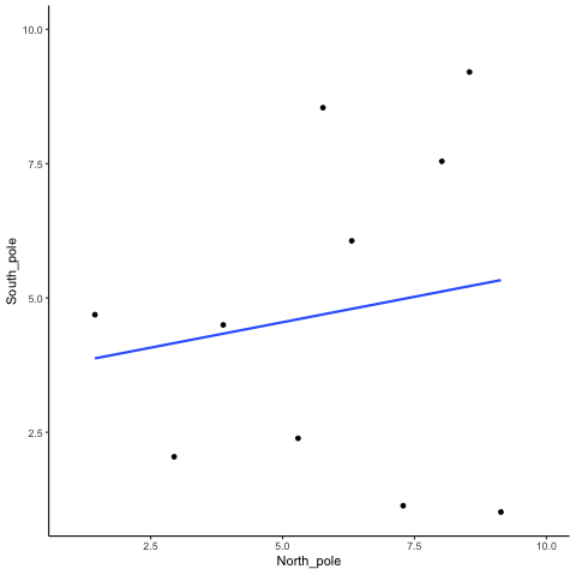


Figure **12.6.4**: Completely random data points drawn from a uniform distribution with a small sampl-size of 10. The blue line twirls around sometimes showing large correlations that are produced by chance.

You might be thinking this is kind of disturbing. If we know that there should be no correlation between two random variables, how come we are finding correlations? This is a big problem right? I mean, if someone showed me a correlation between two things, and then claimed one thing was related to another, how could know I if it was true. After all, it could be chance! Chance can do that too.

Fortunately, all is not lost. We can look at our simulated data in another way, using a histogram. Remember, just before the movie, we simulated 1000 different correlations using random numbers. By, putting all of those r values into a histogram, we can get a better sense of how chance behaves. We can see what kind of correlations chance is likely or unlikely to produce. Here is a histogram of the simulated r values.

```
simulated_correlations <- length(0)
for(sim in 1:1000){
  North_pole <- runif(10,1,10)
  South_pole <- runif(10,1,10)
  simulated_correlations[sim] <- cor(North_pole,South_pole)
}
hist(simulated_correlations,breaks=seq(-1,1,.1))
```

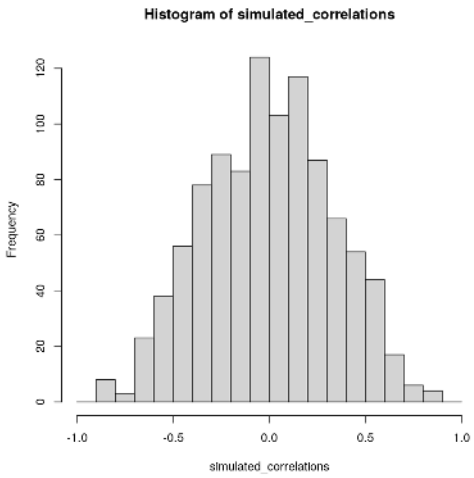


Figure 12.6.5: A histogram showing the frequency distribution of r-values for completely random values between an X and Y variable (sample-size=10). A rull range of r-values can be obtained by chance alone. Larger r-values are less common than smaller r-values.

Notice that this histogram is not flat. Most of the simulated r values are close to zero. Notice, also that the bars get smaller as you move away from zero in the positive or negative direction. The general take home here is that chance can produce a wide range of correlations. However, not all correlations happen very often. For example, the bars for -1 and 1 are very small. Chance does not produce nearly perfect correlations very often. The bars around -.5 and .5 are smaller than the bars around zero, as medium correlations do not occur as often as small correlations by chance alone.

You can think of this histogram as the window of chance. It shows what chance often does, and what it often does not do. If you found a correlation under these very same circumstances (e.g., measured the correlation between two sets of 10 random numbers), then you could consult this window. What should you ask the window? How about, could my observed correlation (the one that you found in your data) have come from this window. Let's say you found a correlation of $r = .1$. Could a .1 have come from the histogram? Well, look at the histogram around where the .1 mark on the x-axis is. Is there a big bar there? If so, this means that chance produces this value fairly often. You might be comfortable with the inference: Yes, this .1 could have been produced by chance, because it is well inside the window of chance. How about $r = .5$? The bar is much smaller here, you might think, "well, I can see that chance does produce .5 some times, so chance could have produced my .5. Did it? Maybe, maybe not, not sure". Here, your confidence in a strong inference about the role of chance might start getting a bit shakier.

How about an $r = .95$?. You might see that the bar for .95 is very very small, perhaps too small to see. What does this tell you? It tells you that chance does not produce .95 very often, hardly if at all, pretty much never. So, if you found a .95 in your data, what would you infer? Perhaps you would be comfortable inferring that chance did not produce your .95, after .95 is mostly outside the window of chance.

Increasing sample-size decreases opportunity for spurious correlation

Before moving on, let's do one more thing with correlations. In our pretend lottery game, each participant only sampled 10 balls each. We found that this could lead to a range of correlations between the numbers randomly drawn from either sides of the pole. Indeed, we even found some correlations that were medium to large in size. If you were a researcher who found such correlations, you might be tempted to believe there was a relationship between your measurements. However, we know in our little game, that those correlations would be spurious, just a product of random sampling.

The good news is that, as a researcher, you get to make the rules of the game. You get to determine how chance can play. This is all a little bit metaphorical, so let's make it concrete.

We will see what happens in four different scenarios. First, we will repeat what we already did. Each participant will draw 10 balls, then we compute the correlation, and do this over 1000 times and look at a histogram. Second, we will change the game so each participant draws 50 balls each, and then repeat our simulation. Third, and fourth, we will change the game so each participant draws 100 balls each, and then 1000 balls each, and repeat etc.

The graph below shows four different histograms of the Pearson r values in each of the different scenarios. Each scenario involves a different sample-size, from, 10, 50, 100 to 1000.

```
library(ggplot2)
all_df<-data.frame()
for(s_size in  c(10,50,100,1000)){
  simulated_correlations <- length(0)
  for(sim in 1:1000){
    North_pole <- runif(s_size,1,10)
    South_pole <- runif(s_size,1,10)
    simulated_correlations[sim] <- cor(North_pole,South_pole)
  }
sim_df <- data.frame(sample_size=rep(s_size,1000),sims=1:1000,simulated_correlations)
all_df<-rbind(all_df,sim_df)
}
ggplot(all_df,aes(x=simulated_correlations))+
  geom_histogram(bins=30)+
  facet_wrap(~sample_size)+
  theme_classic()
```

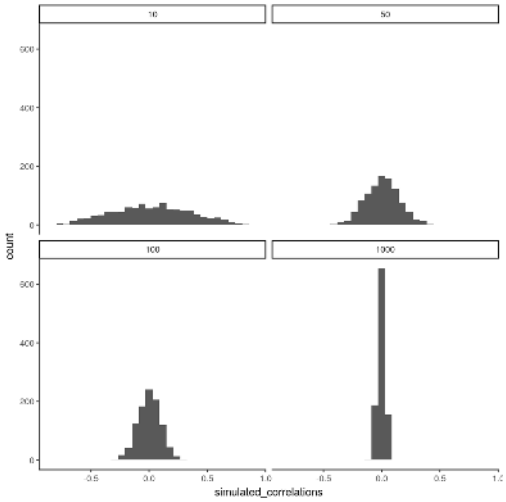


Figure **12.6.6**: Four histograms showing the frequency distributions of r-values between completely random X and Y variables as a function of sample-size. The width of the distributions shrink as sample-size increases. Smaller sample-sizes are more likely to produce a wider range of r-values by chance. Larger sample-sizes always produce a narrow range of small r-values.

By inspecting the four histograms you should notice a clear pattern. The width or range of each histogram shrinks as the sample-size increases. What is going on here? Well, we already know that we can think of these histograms as windows of chance. They tell us which r values occur fairly often, which do not. When our sample-size is 10, lots of different r values happen. That histogram is very flat and spread out. However, as the sample-size increases, we see that the window of chance gets pulled in. For example, by the time we get to 1000 balls each, almost all of the Pearson r values are very close to 0.

One take home here, is that increasing sample-size narrows the window of chance. So, for example, if you ran a study involving 1000 samples of two measures, and you found a correlation of .5, then you can clearly see in the bottom right histogram that .5 does not occur very often by chance alone. In fact, there is no bar, because it didn't happen even once in the simulation. As a result, when you have a large sample size like n = 1000, you might be more confident that your observed correlation (say of .5) was not a spurious correlation. If chance is not producing your result, then something else is.

Finally, notice how your confidence about whether or not chance is mucking about with your results depends on your sample size. If you only obtained 10 samples per measurement, and found $r = .5$, you should not be as confident that your correlation reflects a real relationship. Instead, you can see that r's of .5 happen fairly often by chance alone.

> *Pro tip: when you run an experiment you get to decide how many samples you will collect, which means you can choose to narrow the window of chance. Then, if you find a relationship in the data you can be more confident that your finding is real, and not just something that happened by chance.*

Some more movies

Let's ingrain these idea with some more movies. When our sample-size is small (N is small), sampling error can cause all sort "patterns" in the data. This makes it possible, and indeed common, for "correlations" to occur between two sets of numbers. When we increase the sample-size, sampling error is reduced, making it less possible for "correlations" to occur just by chance alone. When N is large, chance has less of an opportunity to operate.

Watching how correlation behaves when there is no correlation

Below we randomly sample numbers for two variables, plot them, and show the correlation using a line. There are four panels, each showing the number of observations in the samples, from 10, 50, 100, to 1000 in each sample.

Remember, because we are randomly sampling numbers, there should be no relationship between the X and Y variables. But, as we have been discussing, because of chance, we can sometimes observe a correlation (due to chance). The important thing to watch is how the line behaves across the four panels. The line twirls around in all directions when the sample size is 10. It is also moves around quite a bit when the sample size is 50 or 100. It still moves a bit when the sample size is 1000, but much less. In all cases we expect that the line should be flat, but every time we take new samples, sometimes the line shows us pseudo patterns.

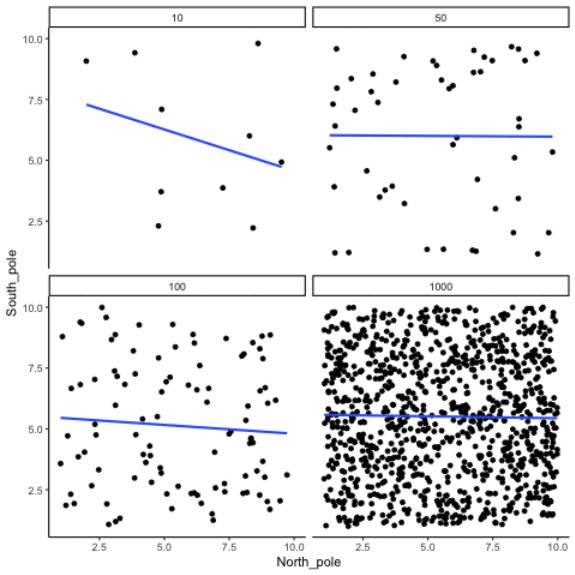


Figure **12.6.7**: Animation of how correlation behaves for completely random X and Y variables as a function of sample size. The best fit line is not very stable for small sample-sizes, but becomes more reliably flat as sample-size increases.

Which line should you trust? Well, hopefully you can see that the line for 1000 samples is the most stable. It tends to be very flat every time, and it does not depend so much on the particular sample. The line with 10 observations per sample goes all over the place. The take home here, is that if someone told you that they found a correlation, you should want to know how many observations they hand in their sample. If they only had 10 observations, how could you trust the claim that there was a correlation? You can't!!! Not now that you know samples that are that small can do all sorts of things by chance alone. If instead, you found out the sample was very large, then you might trust that finding a little bit more. For example, in the above movie you can see that

when there are 1000 samples, we never see a strong or weak correlation; the line is always flat. This is because chance almost never produces strong correlations when the sample size is very large.

In the above example, we sampled numbers random numbers from a uniform distribution. Many examples of real-world data will come from a normal or approximately normal distribution. We can repeat the above, but sample random numbers from the same normal distribution. There will still be zero actual correlation between the X and Y variables, because everything is sampled randomly. But, we still see the same behavior as above. The computed correlation for small sample-sizes fluctuate wildly, and large sample sizes do not.

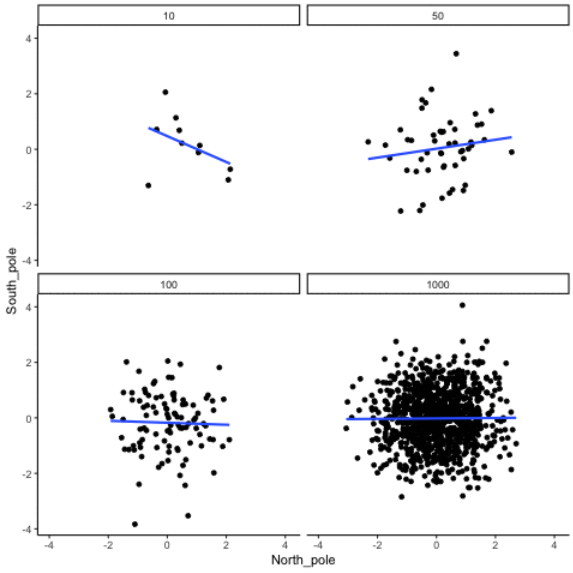


Figure **12.6.8**: Animation of correlation for random values sampled from a normal distribution, rather than a uniform distribution.

OK, so what do things look like when there actually is a correlation between variables?

Watching correlations behave when there really is a correlation

Sometimes there really are correlations between two variables that are not caused by chance. Below, we get to watch a movie of four scatter plots. Each shows the correlation between two variables. Again, we change the sample-size in steps of 10, 50 100, and 1000. The data have been programmed to contain a real positive correlation. So, we should expect that the line will be going up from the bottom left to the top right. However, there is still variability in the data. So this time, sampling error due to chance will fuzz the correlation. We know it is there, but sometimes chance will cause the correlation to be eliminated.

Notice that in the top left panel (sample-size 10), the line is twirling around much more than the other panels. Every new set of samples produces different correlations. Sometimes, the line even goes flat or downward. However, as we increase sample-size, we can see that the line doesn't change very much, it is always going up showing a positive correlation.

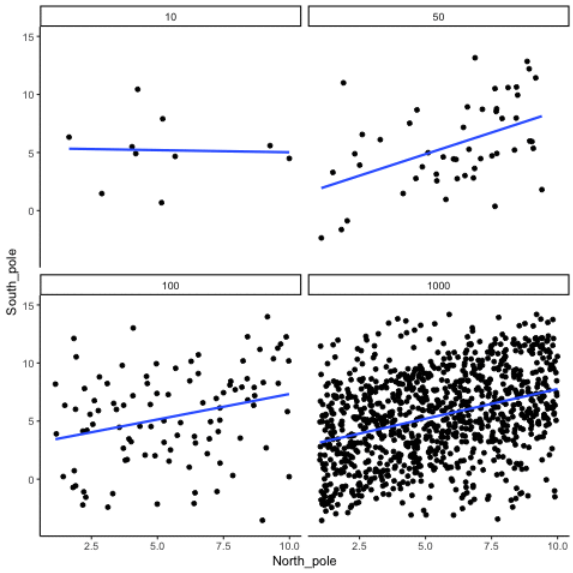


Figure 12.6.9: How correlation behaves as a function of sample-size when there is a true correlation between X and Y variables.

The main takeaway here is that even when there is a positive correlation between two things, you might not be able to see it if your sample size is small. For example, you might get unlucky with the one sample that you measured. Your sample could show a negative correlation, even when the actual correlation is positive! Unfortunately, in the real world we usually only have the sample that we collected, so we always have to wonder if we got lucky or unlucky. Fortunately, if you want to remove luck, all you need to do is collect larger samples. Then you will be much more likely to observe the real pattern, rather the pattern that can be introduced by chance.

This page titled 12.6: Interpreting Correlations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew J. C. Crump via source content that was edited to the style and standards of the LibreTexts platform.

- **3.6: Interpreting Correlations** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

12.7: Summary

In this section we have talked about correlation, and started to build some intuitions about **inferential statistics**, which is the major topic of the remaining chapters. For now, the main ideas are:

1. We can measure relationships in data using things like correlation
2. The correlations we measure can be produced by numerous things, so they are hard to to interpret
3. Correlations can be produced by chance, so have the potential to be completely meaningless.
4. However, we can create a model of exactly what chance can do. The model tells us whether chance is more or less likely to produce correlations of different sizes
5. We can use the chance model to help us make decisions about our own data. We can compare the correlation we found in our data to the model, then ask whether or not chance could have or was likely to have produced our results.

This page titled 12.7: Summary is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Matthew J. C. Crump via source content that was edited to the style and standards of the LibreTexts platform.

- **3.7: Summary** by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

12.8: References

Salsburg, David. 2001. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Macmillan.

12.8: References is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

- **3.8: References** has no license indicated.

CHAPTER OVERVIEW

13: Chi-Square

13.1: Introduction to Chi-Square

13.1.1: Assumptions of the Test(s)

13.2: Introduction to Goodness-of-Fit Chi-Square

13.2.1: Critical Values of Chi-Square Table

13.2.2: Interpretation of the Chi-Square Goodness-of-Fit Test

13.3: Goodness of Fit χ^2 Formula

13.4: Practice Goodness of Fit- Pineapple on Pizza

13.5: Introduction to Test of Independence

13.6: Practice Chi-Square Test of Independence- College Sports

13.6.1: Practice- Fast Food Meals

13.7: RM Chi-Square- The McNemar Test

13.8: Choosing the Correct Test- Chi-Square Edition

This page titled 13: Chi-Square is shared under a not declared license and was authored, remixed, and/or curated by Michelle Oja.

13.1: Introduction to Chi-Square

I don't know about you, but I am TIRED. We've learned SO MUCH.

Can you stay with me for one more chapter, though? See, we've covered the appropriate analyses when we have means for different groups and when we have two different quantitative variables. We've also briefly covered when we have ranks or medians for different groups, and when we have two binary or ranked variables. But what we haven't talked about yet is when we only have qualitative variables. When we have things with names, and all that we can do is count them. For those types of situations, the Chi-Square (χ^2) analysis steps in! It's pronounced like "kite" not like "Chicago" or "chai tea".

Let's practice a little to remind ourselves about qualitative and quantitative variables; it's been a minute since we first introduced these types of variables (and scales of measurement)!

? Exercise 13.1.1

What type is each of the following? Qualitative or Quantitative?

1. Hair color
2. Ounces of vodka
3. Type of computer (PC or Mac)
4. MPG (miles per gallon)
5. Type of music

Answer

1. Hair color: Qualitative (it's a quality, a name, not a number)
2. Ounces of vodka: Quantitative (it's a number that measures something)
3. Type of computer (PC or Mac): Qualitative
4. MPG: Quantitative
5. Type of music: Qualitative

? Exercise 13.1.2

Do you use means to find the average of qualitative or quantitative variables?

Answer

Quantitative. Means are mathematical averages, so the variable has to be a number that measures something.

Instead of means, you use counts with qualitative variables.

Frequency counts: Counts of how many things are in each level of the categories.

Introducing Chi-Square

Our data for the χ^2 test (the chi is a weird-looking X) are quantitative, (also known as nominal) variables. Recall from our discussion of scales of measurement that nominal variables have no specified order (no ranks) and can only be described by their names and the frequencies with which they occur in the dataset. Thus, we can only count how many "things" are in each category. 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 **13.1.1**: Pet Preferences

	Cat	Dog	Other	Total
Observed	14	17	5	36
Expected	*12*	*12*	*12*	*36*

Table **13.1.1** gives an example of a contingency 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 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. This is our actualy data. 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:

$$E = \frac{N}{C}$$

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). Thank the Higher Power of Statistics, formulas with symbols that finally mean something! 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 13.1: Introduction to Chi-Square is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **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.

13.1.1: Assumptions of the Test(s)

All statistical tests make assumptions about your variables, data, and the distributions that they come from. Usually, it's not a huge deal if your sample's data doesn't fit some of the assumptions. Chi-square is not one of those tests. For the chi-square tests discussed so far in this chapter, the assumptions are:

- *Expected frequencies are sufficiently large*. All of the expected frequencies need to be reasonably big for the statistics to use the correction distribution in the background. How big is reasonably big? Opinions differ, but the default assumption seems to be that you generally would like to see all your expected frequencies larger than about 5, though for larger tables you would probably be okay if at least 80% of the the expected frequencies are above 5 and none of them are below 1 (meaning, no categories are empty). However, from what Dr. Navarro has been able to discover , these seem to have been proposed as rough guidelines, not hard and fast rules; and they seem to be somewhat conservative (Larntz, 1973).
- *Data are independent of one another*. One somewhat hidden assumption of the chi-square test is that you have to genuinely believe that the observations are independent. Here's what I mean. Suppose I'm interested in proportion of babies born at a particular hospital that are boys. I walk around the maternity wards, and observe 20 girls and only 10 boys. Seems like a pretty convincing difference, right? But later on, it turns out that I'd actually walked into the same ward 10 times, and in fact I'd only seen 2 girls and 1 boy. Not as convincing, is it? My original 30 *observations* were massively non-independent because I really only had 3 observation. Obviously this is an extreme (and extremely silly) example, but it illustrates the basic issue. Non-independence messes things up. Sometimes it causes you to falsely reject the null, as the silly hospital example illustrates, but it can go the other way too. To give a slightly less stupid example, let's consider what would happen if we asked 50 people to select 4 cards. One possibility would be that *everyone* selects one heart, one club, one diamond and one spade. This is highly non-random behavior from people, but in this case, I would get an observed frequency of 50 for all four suits. For this example, the fact that the observations are non-independent (because the four cards that you pick will be related to each other) actually leads to the opposite effect… falsely retaining the null.

If you happen to find yourself in a situation where independence is violated, it may be possible to use the McNemar test (which we'll discuss) or the Cochran test (which we won't). Similarly, if your expected cell counts are too small, check out the Fisher exact test (which we also won't discuss).

Our first stop in the tour of Chi-Square is the Goodness of Fit test. See you there!

Reference

Larntz, K. Small-sample comparison of exam levels for Chi-Squared Goodness-of-Fit statistics. *Journal of the American Statistical Association, 73* (362), 253-263.

This page titled 13.1.1: Assumptions of the Test(s) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **12.5: Assumptions of the Test(s)** by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.
- **16.1.1: Assumptions of the Test(s)** by Michelle Oja is licensed CC BY-SA 4.0.

13.2: Introduction to Goodness-of-Fit Chi-Square

The first of our two χ^2 tests, the Goodness of Fit test, assesses the distribution of frequencies into different categories of one quantitative variable against any specific distribution. Usually this is equal frequency distributions because that's what we would expect to get if categorization was completely random, but it can also be a specific distribution. For example, if Dr. MO wanted to compare a specific class's frequency of each ethnicity to a specific distribution, it would make more sense to compare her class to the ethnic demographics of the college rather than an assumption that all of the ethnic groups would have the same number of students in the target class.

Hypotheses for Chi-Square

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. This makes it more difficult to have mathematical statements for χ^2 hypotheses (symbols showing which group is bigger or whatever). The next section will walk through the mathematical hypotheses. For now, we will learn how to still state our hypotheses verbally.

Research Hypothesis

The research hypothesis is that we expect a pattern of difference, and then we explain that pattern of difference.

Using Dr. MO's sample class, she works at a college that is designated as a Hispanic-Serving Institution (HSI), so we would expect a pattern of difference such that there will be more students who are Hispanic in her class than students from any other ethnic group.

Null Hypotheses

For goodness-of-fit χ^2 tests, our null hypothesis is often that there is an equal number of observations in each category. That is, there is no pattern of difference between the frequencies in each category. Unless we're looking at the situation above in which we have a distribution of frequencies that we are comparing our sample to, the null hypothesis is that each group will be the same size.

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:

$$df = k - 1$$

Do you remember what "k" stood for when we discussed ANOVAs?

? Exercise 13.2.1

What does "k" stand for?

Answer

The letter "k" usually stands for the number of groups. In Chi-Square, this would be the number of different categories.

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 next, or can be found through the Common Critical Value Tables at the end of this book.

This page titled 13.2: Introduction to Goodness-of-Fit Chi-Square is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **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.

13.2.1: Critical Values of Chi-Square Table

A new type of statistical analysis, a new table of critical values!

Chi-Square Distributions

As you know, there is a whole family of t-distributions, each one specified by a parameter called the degrees of freedom (df). Similarly, all the chi-square distributions form a family, and each of its members is also specified by a its own df. Chi (like "kite," not like "chai" or "Chicago") is a Greek letter denoted by the symbol χ and chi-square is often denoted by χ^2. It looks like a wiggly X, but is not an X. Figure 13.2.1.1 shows several χ-square distributions for different degrees of freedom.

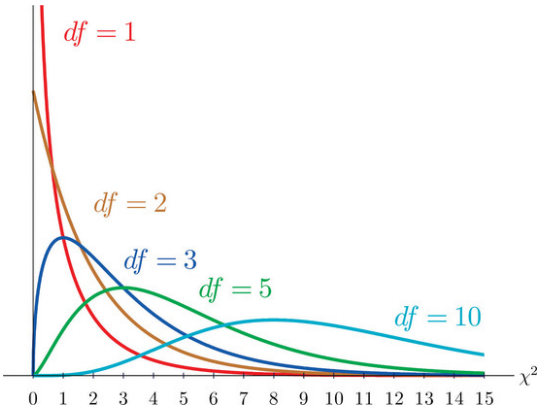


Figure 13.2.1.1: Many χ Distributions (CC-BY-NC-SA Shafer & Zhang)

Like all tables of critical values, this one provides the value in which you should reject the null hypothesis if your calculated value is bigger than the critical value i the table. For chi-square, the null hypothesis is that there is no pattern of relationship, but the process of Null Hypothesis Signficance Testing is the same as we've been learning.

Note

(Critical < Calculated) = Reject null = There is a pattern of relationship. = $p < .05$

(Critical > Calculated) = Retain null = There is no pattern of relationship. = $p > .05$

Illustrated in Figure 13.2.1.2, the value of the chi-square that cuts off a right tail of area c is denoted χ^2_c and is called a critical value (Figure 13.2.1.2).

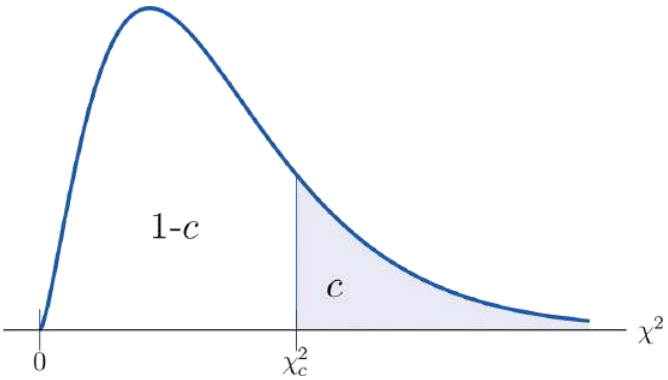


Figure 13.2.1.2: χ^2_c Illustration of Critical Values of Chi-Square. (CC-BY-NC-SA Shafer & Zhang)

Table of Critical Values for χ^2_c

Table 13.2.1.1 below gives values of χ^2_c for various values of c and under several chi-square distributions with various degrees of freedom.

Table 13.2.1.1- Critical Values on the Right Side of the Distribution of Chi-Square

df	p = 0.10	p = 0.05	p = 0.01
1	2.706	3.841	6.635
2	4.605	5.991	9.210
3	6.251	7.815	11.345
4	7.779	9.488	13.277
5	9.236	11.070	15.086
6	10.645	12.592	16.812
7	12.017	14.067	18.475
8	13.362	15.507	20.090
9	14.684	16.919	21.666
10	15.987	18.307	23.209
11	17.275	19.675	24.725
12	18.549	21.026	26.217
13	19.812	22.362	27.688
14	21.064	23.685	29.141
15	22.307	24.996	30.578
16	23.542	26.296	32.000
17	24.769	27.587	33.409
18	25.989	28.869	34.805
19	27.204	30.144	36.191
20	28.412	31.410	37.566
100	118.498	124.342	135.807

Degrees of Freedom

Like with the t-test and ANOVA, the degrees of freedom are based on which kind of analysis you are conducting.

- χ^2_{GoF} Goodness of Fit: $k-1$
  - k is the number of categories.
- χ^2_{ToI} Test of Independence: $(R-1)\times(C-1)$
  - R is the number of rows
  - C is the number of columns
- Kruskal-Wallis Test: $k-1$
  - k is the number of groups

This page titled 13.2.1: Critical Values of Chi-Square Table is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **11.1: Chi-Square Tests for Independence** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

13.2.2: Interpretation of the Chi-Square Goodness-of-Fit Test

This is a long section, but it really helps you understand where the the Chi-Square Goodness of Fit comes from and what it means.

Goodness of Fit Test

The χ^2 Goodness of fit test is one of the oldest hypothesis tests around: it was invented by Karl Pearson around the turn of the century (Pearson 1900), with some corrections made later by Sir Ronald Fisher (Fisher 1922a). To introduce the statistical problem that it addresses, let's start with some psychology…

Over the years, there have been a lot of studies showing that humans have a lot of difficulties in simulating randomness. Try as we might to "act" random, we *think* in terms of patterns and structure, and so when asked to "do something at random", what people actually do is anything but random. As a consequence, the study of human randomness (or non-randomness, as the case may be) opens up a lot of deep psychological questions about how we think about the world. With this in mind, let's consider a very simple study. Suppose I asked people to imagine a shuffled deck of cards, and mentally pick one card from this imaginary deck "at random". After they've chosen one card, I ask them to mentally select a second one. For both choices, what we're going to look at is the suit (hearts, clubs, spades or diamonds) that people chose. After asking, say, N=200 people to do this, I'd like to look at the data and figure out whether or not the cards that people *pretended* to select were really random. The data on the card that people pretended to select are shown in Table **13.2.2.1**, called "Observed" for reasons that will become clear very soon:

Table **13.2.2.1**- Observed Card Suits

Card Suit Pretended to Choose:	Clubs ♣	Diamonds ♢	Hearts ♡	Spades ♠
Observed:	35	51	64	50

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

Excellent. It's also worth nothing that mathematicians prefer to talk about things in general rather than specific things, so you'll also see the notation O_i, which refers to the number of observations that fall within the i-th category (where it could be 1, 2, or 3, or x, y, or z). This textbook has used these type of subscripts in the formulas, but then replaced them with letters relating to the name of the groups in the practice. So, if we want to refer to the set of all observed frequencies, statisticians group all of observed values into one variable, which we'll refer to as O for Observed.

$$O=(O_1,O_2,O_3,O_4)$$

or

$$O=(O_C,O_D,O_H,O_S)$$

Again, there's nothing new or interesting here: it's just notation. If we say that O = (35,51,64,50) all we're doing is describing the table of observed frequencies (i.e., Table **13.2.2.1**), but we're referring to it using mathematical notation.

Hypotheses

Null Hypothesis

We'll start with null hypotheses for Chi-Square Goodness of Fit test because the null hypothesis will help us understand our more limited research hypothesis.

Our research question is whether people choose cards randomly or not. What we're going to want to do now is translate this into some statistical hypotheses, and construct a statistical test of those hypotheses. In this case, the null hypothesis in words is that there is no pattern of relationship in the suits that participants pretended to choose; in other words, all four suits will be chosen with equal probability.

Now, because this is statistics, we have to be able to say the same thing in a mathematical way. If the null hypothesis is true, then each of the four suits has a 25% chance of being selected: in other words, our null hypothesis claims that PC=.25, P_D=.25, P_H=.25 and finally that P_S=.25. However, in the same way that we can group our observed frequencies into a variable called O that summarizes the entire data set, we can use P to refer to the probabilities that correspond to our null hypothesis. So if I let the P= (P_1,P_2,P_3,P_4), or P = (PC, P_D ,P_H ,P_S) refers to the collection of probabilities that describe our null hypothesis, then we have

Null Hypothesis for $P = (0.25, 0.25, 0.25, 0.25)$

In this particular instance, our null hypothesis corresponds to a probabilities P in which all of the probabilities are equal to one another. But this doesn't have to be the case. For instance, if the experimental task was for people to imagine they were drawing from a deck that had twice as many clubs as any other suit, then the null hypothesis would correspond to something like P=(0.4, 0.2, 0.2, 0.2). As long as the probabilities are all positive numbers, and they all sum to 1, them it's a perfectly legitimate choice for the null hypothesis. However, the most common use of the Goodness of Fit test is to test a null hypothesis that all of the categories are equally likely, so we'll stick to that for our example.

Research Hypothesis

What about our research hypothesis? All we're really interested in is demonstrating that the probabilities involved aren't all identical (that is, people's choices weren't completely random). As a consequence, the "human friendly" versions of our research hypothesis is that there is a pattern of relationship in the suits that participants pretended to choose. Another way to say this is that at least one of the suit-choice probabilities *isn't* 0.25. This leads to the mathematical research hypothesis of:

Research Hypothesis of $P \neq (0.25, 0.25, 0.25, 0.25)$

The "Goodness of Fit" Test Statistic

At this point, we have our observed frequencies O and a collection of probabilities P corresponding the null hypothesis that we want to test. What we now want to do is construct a test of the null hypothesis. The basic trick that a Goodness of Fit test uses is to construct a test statistic that measures how "close" the data are to the null hypothesis. If the data don't resemble what you'd "expect" to see if the null hypothesis were true, then it probably isn't true. Okay, if the null hypothesis were true, what would we expect to see? Or, to use the correct terminology, what are the expected frequencies. There are N=200 observations, and (if the null is true) the probability of any one of them choosing a heart is P_H = 0.25, so I guess we're expecting 200 x 0.25=50 hearts, right? Or, more specifically, if we let E refer to "the number of responses from any one category that we'd expect if the null is true", then

$$E = N \times P$$

This test is pretty easy to calculate, and we'll cover in the next sections. But to focus on what's happening: if there are 200 observation that can fall into four categories, and we think that all four categories are equally likely, then on average we'd expect to see 50 observations in each category, right?

Now, how do we translate this into a test statistic? Clearly, what we want to do is compare the *expected* number of observations in each category (E_i) with the *observed* number of observations in that category (O_i). And on the basis of this comparison, we ought to be able to come up with a good test statistic. To start with, let's calculate the difference between what the null hypothesis expected us to find and what we actually did find. That is, we calculate the "observed minus expected" difference score, E-O (Expected minus Observed) for each category. In this scenario, the categories are the car suits. This is illustrated in the following table (Table **13.2.2.2**).

Table **13.2.2.2**- Contingency Table of Cards Pretended to Select

	Clubs ♣	Diamonds ♢	Hearts ♡	Spades ♠
Observed Frequency (O)	35	51	64	50
Expected Frequency (E)	50	50	50	50
Difference Score (O Minus E)	15	-1	-14	0

It's clear that people chose more hearts and fewer clubs than the null hypothesis predicted. However, a moment's thought suggests that these raw differences aren't quite what we're looking for. Intuitively, it feels like it's just as bad when the null hypothesis predicts too few observations (which is what happened with hearts) as it is when it predicts too many (which is what happened with clubs). So it's a bit weird that we have a negative number for clubs and a positive number for hearts. One easy way to fix this is to square everything, so that we now calculate the squared differences, $(E_i - O_i)^2$. Let's see what that looks like in Table **13.2.2.3**.

Table **13.2.2.3**-Contingency Table with Differences Squared

	Clubs ♣	Diamonds ♢	Hearts ♡	Spades ♠
Observed Frequency (O)	35	51	64	50
Expected Frequency (E)	50	50	50	50
Difference Score (O Minus E)	15	-1	-14	0
Difference Score Squared	$15^2 = 225$	$(-1)^2 = 1$	$(-14)^2 = 196$	$0^2 = 0$

Now we're making progress. What we've got now is a collection of numbers that are big whenever the null hypothesis makes a bad prediction (clubs and hearts), but are small whenever it makes a good one (diamonds and spades). Next, for some technical reasons that I'll explain in a moment, let's also divide all these numbers by the expected frequency E_i, so we're actually calculating:

$$\frac{(E-O)^2}{E}$$

Since E=50 for all categories in our example, it's not a very interesting calculation, but let's do it anyway. The results are shown in another table (Table **13.2.2.4**)!

Table **13.2.2.4**- Contingency Table with Observed, Expected, Difference, Difference Squared, and Divided by Expected

	Clubs ♣	Diamonds ♢	Hearts ♡	Spades ♠
Observed Frequency (O)	35	51	64	50
Expected Frequency (E)	50	50	50	50
Difference Score (O Minus E)	15	-1	-14	0
Difference Score Squared	$15^2 = 225$	$(-1)^2 = 1$	$(-14)^2 = 196$	$0^2 = 0$
Diff^2 divided by Expected	$\frac{225}{50} = 4.50$	$\frac{1}{50} = 0.02$	$\frac{196}{50} = 3.92$	$\frac{0}{50} = 0.00$

In effect, what we've got here are four different "error" scores, each one telling us how big a "mistake" the null hypothesis made when we tried to use it to predict our observed frequencies. So, in order to convert this into a useful test statistic, one thing we could do is just add these numbers up. The result is called the *Goodness of Fit* statistic, conventionally referred to either as χ^2 r GOF. What's cool about this is that it's easy to calculate if each of the Expected is a different amount of suits, too.

Intuitively, it's clear that if χ^2 s small, then the observed data are very close to what the null hypothesis predicted expected values, so we're going to need a large χ^2 tatistic in order to reject the null. As we've seen from our calculations, in our cards data set we've got a value of χ^2=8.44 $(4.50 + 0.02 + 3.92 + 0.00 = 8.44)$. So now the question becomes, is this a big enough value to reject the null? The simple answer is that we compare our calculated χ^2=8.44 to a critical values in the Critical Values of Chi-Square Table. The longer answer is below.

The Sampling Distribution of the Goodness of Fit Statistic (advanced)

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

Okay, let's suppose that the null hypothesis is actually true. If so, then the true probability that an observation falls in the i-th category is P_i – after all, that's pretty much the definition of our null hypothesis. Let's think about what this actually means. If you think about it, this is kind of like saying that "nature" makes the decision about whether or not the observation ends up in category by flipping a weighted coin (i.e., one where the probability of getting a head is P). And therefore, we can think of our observed frequency O by imagining that nature flipped N of these coins (one for each observation in the data set)… and exactly O_i of them came up heads. Obviously, this is a pretty weird way to think about the experiment but makes sense if you remember anything about the binomial distribution. Now, if you remember from our discussion of the central limit theorem, the binomial distribution starts to look pretty much identical to the normal distribution, especially when N is large and when the probability isn't *too* close to 0 or 1. In other words as long as N×P is large enough – or, to put it another way, when the expected frequency is large enough – the theoretical distribution of O is approximately normal. Better yet, if O is normally distributed, then so is $\frac{(O-E)}{\sqrt{E}}$… since E is a fixed value, subtracting off E and dividing by $\sqrt{E}$ changes the mean and standard deviation of the normal distribution; but that's all it does.

Okay, so now let's have a look at what our Goodness of Fit statistic actually *is*. What we're doing is taking a bunch of things that are normally-distributed, squaring them, and adding them up. Wait. We've seen that before too!

When you take a bunch of things that have a standard normal distribution (i.e., mean 0 and standard deviation 1), square them, then add them up, then the resulting quantity has a Chi-Square distribution. So now we know that the null hypothesis predicts that the sampling distribution of the Goodness of Fit statistic is a Chi-Square distribution. Cool. There's one last detail to talk about, namely the degrees of freedom. What we're supposed to be looking at is the number of genuinely *independent* things that are getting added together. And, as I'll go on to talk about in the next section, even though there's k things that we're adding, only k−1 of them are truly independent; and so the degrees of freedom is actually only k−1. If you continue learning about statistics, it will be explained.If you're interested, the next section describes why. However, it's fine to calculate the statistics, even to interpret them, without fully understanding all of these concepts.

Degrees of Freedom

Looking Figure 13.2.2.1 you can see that if we change the degrees of freedom, then the Chi-Square distribution changes shape quite substantially.

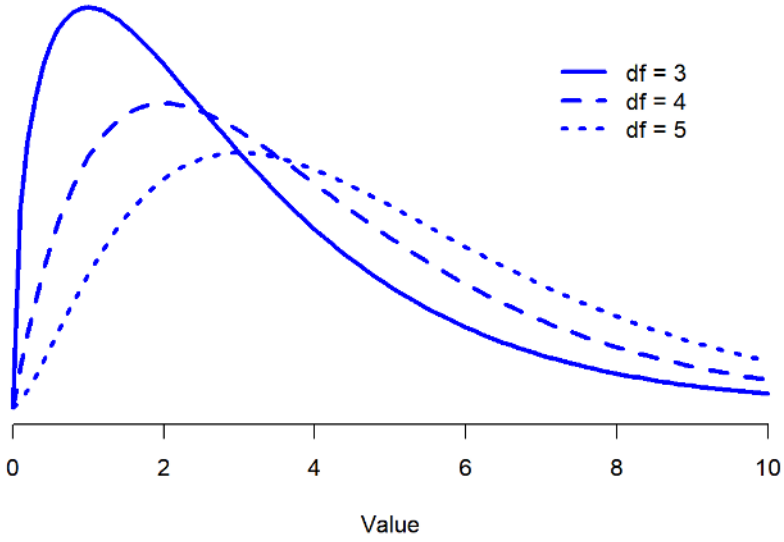


Figure 13.2.2.1- Chi-Square Distributions with Different df's(CC-BY-SA- Danielle Navarro from Learning Statistics with R).

But what exactly *is* it? It's the number of "normally distributed variables" that are being squared and added together. But, for most people, that's kind of abstract, and not entirely helpful. What we really need to do is try to understand degrees of freedom in terms of our data. So here goes.

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

Testing the Null Hypothesis

The final step in the process of constructing our hypothesis test is to figure out what the rejection region is. That is, what values of χ^2 would lead is to reject the null hypothesis. As we saw earlier, large values of X^2 imply that the null hypothesis has done a poor job of predicting the data from our experiment, whereas small values of X^2 imply that it's actually done pretty well. Therefore, a pretty sensible strategy would be to say there is some critical value, such that if χ^2 is bigger than the critical value we reject the null; but if χ^2 is smaller than this value we retain the null. In other words, to use the same language that we've been using! The chi-squared goodness of fit test is always a one-sided test. Right, so all we have to do is figure out what this critical value is. And it's pretty straightforward. If we want our test to have significance level of α=.05 (that is, we are willing to tolerate a Type I error rate of 5%), then we have to choose our critical value so that there is only a 5% chance that χ^2 could get to be that big if the null hypothesis is true. That is to say, we want the 95th percentile of the sampling distribution. This is illustrated inFigure 13.2.2.2.

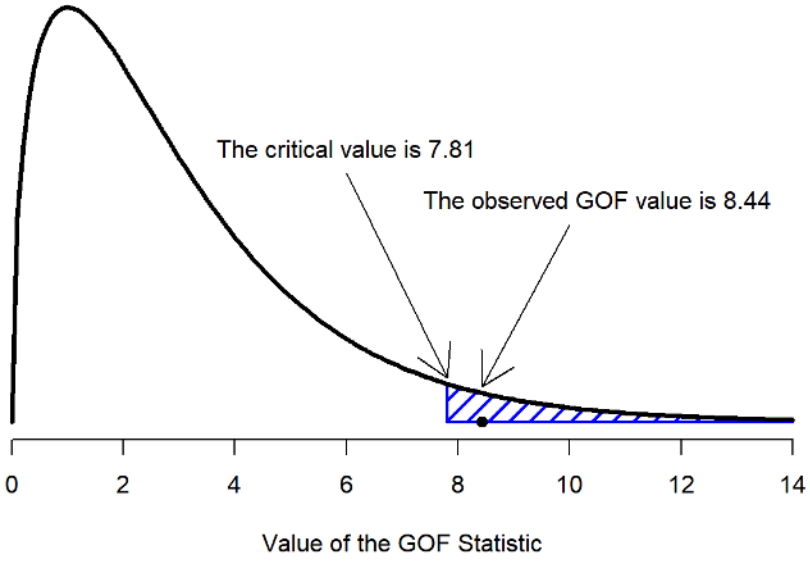


Figure 13.2.2.2- Illustration of how the hypothesis testing works for the chi-square goodness of fit test (CC-BY-SA- Danielle Navarro from Learning Statistics with R).

So if our χ^2 statistic is bigger than 7.81 or so (7.815 from our Critical Values of Chi-Square Table, then we can reject the null hypothesis. Since we actually calculated that before (i.e., χ^2 = 8.44) we can reject the null. So, in this case we would reject the null hypothesis, since p<.05. And that's it, basically. You now know "Pearson's χ^2 test for the goodness of fit". Lucky you.

Example 13.2.2.1

What do you think that the statistical sentence would look like for this scenario?

Solution

The statistical sentence would be $\chi^2(3) = 8.44, p < .05$

Summary

That is a lot of detailed information. If you like to know the "why" behind things, I hope that helped! If you just want the nuts and bolts of how to calculate and then interpret a Chi-Square, I hope that you skimmed this prior section. You will learn how to use the full formula next, and practice calculating and interpreting what it means!

This page titled 13.2.2: Interpretation of the Chi-Square Goodness-of-Fit Test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **16.2.2: Interpretation of the Chi-Square Goodness-of-Fit Test** by Michelle Oja is licensed CC BY-SA 4.0.
- **12.1: The χ2 Goodness-of-fit Test** by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.

13.3: Goodness of Fit χ^2 Formula

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 qualitative 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:

$$\chi^2 = \sum_{Each} \left(\frac{(E-O)^2}{E} \right)$$

This formula is telling us to find the difference, square it, then divide by the Expected value for that category, and then add together that number for each categor.

Huh? Let's continue to use our pet preference data, shown in Table **13.3.1** to see what that means. We'll first use the table to do all of the calculations described in the formula, then use the formula alone.

Table **13.3.1**- Pet Preference Observations & Expectations

	Cat	Dog	Other	Total
Observed Frequencies	14	17	5	36
Expected Frequencies	*12*	*12*	*12*	*36*
Difference Score (E Minus O)				
Difference Score Squared				
Diff^2 divided by Expected				

Let's look at Table **13.3.1** a little closer first. The Total column is the sum of the frequencies in that row. In this case, the Total is also our N because each person could only choose one type of pet. To determine the Expected frequencies, we used to Total, and divided it by how many groups we have (k = 3, which are Cats, Dogs, Other):

$$\frac{Total}{k} = \frac{36}{3} = 12$$

Okay, now that know where the numbers come from so far, fill in the rest of the table.

Example 13.3.1

Calculate the formula to complete Table **13.3.1**.

Solution

Table **13.3.2**- Pet Preference Observations & Expectations

	Cat	Dog	Other	Total
Observed Frequencies	14	17	5	36
Expected Frequencies	*12*	*12*	*12*	*36*
Difference Score (E Minus O)	-2	-5	7	0
Difference Score Squared ($Diff^2$)	4	25	49	78
$Diff^2$ divided by Expected	0.33	2.08	4.08	6.49

What would this look like with our Chi-Square formula?

$$\chi^2 = \frac{(14-12)^2}{12} + \frac{(17-12)^2}{12} + \frac{(5-12)^2}{12} = 0.33 + 2.08 + 4.08 = 6.49$$

For each category's calculation, the expected value in the numerator and the expected value in the denominator are the same value whether we used the table or the formula. As you have noticed, the result is also *the same* whether you use the table to do the calculations, or did it all with the formula. The table is explaining each step of the formula, but they are exactly the same process. It's Statisticians Choice how you would like to calculate Chi-Square (table of formula).

Let's now take a look at an example from start to finish.

This page titled 13.3: Goodness of Fit χ^2 Formula is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **14.3: χ^2 Statistic** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.
- **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.

13.4: Practice Goodness of Fit- 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 Like/Dislike answer.

Step 1: State the Hypotheses

We start, as always, with our hypotheses. Chi-Square focuses on patterns of relationship, so that's what the hypotheses in words should talk about. Let's go through research hypothesis to see how this all works out.

Example 13.4.1

What is the research hypothesis in words for this scenario? Make sure to list which group you think will have a higher frequency.

Solution

- Research hypothesis in words: There will be a pattern of difference such that there will be more people who dislike pineapple on their pizza than people who like pineapple on their pizza.

The hypotheses in symbols focus on probabilities, but because of how Chi-Square works, we can only say that the probabilities will not be equal.

- Reseach hypothesis in symbols: $P_{Like} \neq 0.50, P_{Dislike} \neq 0.50, or P \neq (0.50, 0.05)$

The probability of 0.50 (which means a 50% chance) was found by knowing that we only have two options: Like or Dislike. All probabilities add up to 100% chance, so with only two options, we find $\frac{100}{2} = 50$ which means that the P (probability) is 0.50.

If this research hypothesis in symbols doesn't make sense, it might be easier to start with a null hypothesis in words and symbols, then figure out how that works out for the research hypothesis.

Example 13.4.2

What is the null hypothesis in words and symbols for this scenario?

Solution

- Null hypothesis in words: There is no pattern of difference based on liking pineapple on pizza.
- Null hypothesis in symbols: $P_{Like} \neq 0.50, P_{Dislike} \neq 0.50, or P \neq (0.50, 0.05)$

Let's move on to an easier step!

Step 2: Find the Critical Value

Per usual, we will leave α at its typical level of 0.05. You can find the Critical Values of Chi-Square Table earlier in this chapter, or look for the link in the Common Critical Values page at the end of this book.

Exercise 13.4.1

What is the critical value for this scenario?

Answer

We have two options in our data (Like or Dislike), which will give us two categories (k=2). The Degrees of Freedom is found through k-1, so we will have 1 df (k-1=2-1=1). From our χ^2 table of critical values, we find a critical value of 3.841 for our α of p=0.05.

See, that was easy! How, the slightly-less-easy-but-not-that-hard step of calculating the Chi-Square test statistic.

Step 3: Calculate the Test Statistic

The results of the data collection are presented in Table 13.4.1.

Table 13.4.1: Results of Data collection

	Like	Dislike	Total
Observed	19	26	19+26=45

First, let's find the Expected values, then we'll fill have what we need to complete the full calculation.

Example 13.4.3

With two categories and 45 scores, what is the Expected frequency?

Solution

$$E = \frac{45}{2} = 22.50$$

We can use the Observed and the Expected frequencies to calculate our χ^2 statistic either through a table or individually in the equation:

$$\chi^2 = \sum_{Each} \left(\frac{(E-O)^2}{E} \right)$$

The first example will use the table.

Example 13.4.4

Complete the calculations labeled to fill in Table 13.4.2.

Table 13.4.2- Table to Complete Chi-Square Formula

	Like	**Dislike**	**Total**
Observed	19	26	45.00
Expected	22.50	22.50	22.50+22.50=45.00
Difference Score (E Minus O)			
Difference Score Squared			
Diff^2 divided by Expected			

Solution

Table 13.4.3- Table to Complete Chi-Square Formula

	Like	Dislike	Total
Observed	19	26	45.00
Expected	22.50	22.50	45.00
Difference Score (O Minus E)	22.50-19=3.50	22.50-26=-3.50	N/A
Difference Score Squared	$3.50^2 = 12.25$	$-3.50^2 = 12.25$	N/A
Diff^2 divided by Expected	$\frac{12.25}{22.50} = 0.54$	$\frac{12.25}{22.50} = 0.54$	0.54+0.54=1.08

You might have noticed that there are still two empty cells. You can add up the Difference Scores (they equal zero in this example) and the squared Difference Scores (they equal 24.50), but we don't use them for the χ^2 formula, so you can save some time and not calculate them.

Also, if you used a spreadsheet, the final sum of $\frac{Diff^2}{E} = 1.09$; those darn rounding differences!

What would this look like in the Chi-Square formula?

Example 13.4.5

Use the χ^2 formula with the Observed frequencies and Expected frequencies to calculate the test statistic for χ^2:

$$\chi^2 = \sum_{Each}\left(\frac{(E-O)^2}{E}\right)$$

Solution

$$\chi^2 = \frac{(22.50-19)^2}{22.50} + \frac{(22.50-26)^2}{22.50} = 0.54 + 0.54 = 1.08$$

Using the table to calculate the χ^2 and using the formula resulted in the same result (because you are doing the same things mathematically). It's your choice which option that you prefer. It seems easier to use the formula when there are so few categories (k), but the table seems easier to use when there are more categories. The table is also easier to use if you're using a spreadsheet.

Now that we have the calculated χ^2, we can make the decision!

Step 4: Make the Decision

Our observed test statistic had a value of 1.08 and our critical value was 3.84. What do we do if this is still true?

Note

Slightly modified from earlier versions to fit the hypotheses, but the idea is the same:

(Critical < Calculated) = Reject null = There is a pattern of relationship. = $p < .05$

(Critical > Calculated) = Retain null = There is no pattern of relationship. = $p > .05$

Based on this note...

Example 13.4.6

Do we retain or reject the null hypothesis?

Solution

Because our critical value is larger than our calculated value, we retain the null hypothesis.

The debate rages on.

Exercise 13.4.2

What would our results look like in the statistical sentence?

Answer

$\chi^2(1)=1.08$, $p>.05$

Write-Up

How might we write this up? We can't quite fulfill the four requirements for reporting results because there are no means to include. Instead, let's include the Observed frequencies.

Example 13.4.7

Report the results in a concluding paragraph that includes the four requirements (but use Observed frequencies instead of descriptive statistics).

Solution

The research hypothesis was that there would be a pattern of difference such that more people would dislike pineapples on pizza than like pineapples on pizza. This research hypothesis was not supported ($\chi^2(1)=1.08$, $p>.05$). There does not seem to be a pattern of difference; of our 45 participants, 19 people like pineapple on pizza, and 26 people dislike pineapple on pizza.

That's it! If you want more practice, check out this blog post about the frequency of the different colors of M&Ms.

We now move on to the other kind of Chi-Square analysis, the Test of Independence.

This page titled 13.4: Practice Goodness of Fit- Pineapple on Pizza is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **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.

13.5: Introduction to Test of Independence

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 qualitative data (nominal scale of measurement), which violates the assumptions of traditional correlation analylses. Although we learned in the correlation chapter that there is a type of correlation (Phi correlation) when there are two binary variables (two variables that each only have two options), the χ^2 test for independence comes in handy when the variables have more than two categories or levels. The χ^2 test performed when there are two variables is known as the Test of Independence. In this analysis, we are looking to see if the values of each qualitative variable (that is, the frequency of their levels) is related to or independent of the values of the other qualitative variable.

As noted previously, our only description for qualitative data is frequency, so we will again present our observations in a contingency table showing frequencies. When we have two categorical variables, each combination of levels from each categorical variable are presented. This type of called a contingency table because it shows the frequency of each category in one variable, contingent upon the specific level of the other variable.

Example 13.5.1

How are contingency tables different from factorial design tables?

Solution

The difference is what's in the cells. In factorial design squares, the variables are crossed and the means for each cell (and the margins) are what we're interested in. For contingency tables, what's in the cells is the actual frequencies (not means).

An example contingency table is shown in Table **13.5.1**, 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 **13.5.1**: Contingency Table of College Sports and College Decision

College Sports	Primary Affect	Somewhat Affected	Did Not Affect Decision	*Total*
Watched	47	26	14	$\sum_{Row}$ = *87*
Did Not Watch	21	23	37	$\sum_{Row}$ = *81*
Total	$\sum_{Column}$ = *68*	$\sum_{Column}$ = *49*	$\sum_{Column}$ = *51*	$\sum_{Column}$ = *168*

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 **13.5.1** 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 for Rows and C for Columns, 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 or R1,C1. Next to it, which has 26 people who watched college sports as a child but had sports only somewhat affect their decision, cell is 1,2 or R1,C2, 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. Ooh, that sounds familiar! Marginal values are the total values for a single category of one variable, added up across levels of the other variable. In Table **13.5.1**, 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, 50 considered it somewhat, and 50 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. Yay for calculation checks!

Expected Values of Contingency Tables

Because these crossed contingency tables have more data, we don't usually start out with the Expected values in any row or column like we included in the Goodness of Fit tables. However, the 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:

$$E_{EachCell} = \frac{RT * CT}{N}$$

in which RT stands for Row Total and CT stands for Column Total. This allows for calculating an Expected frequency for each cell. Using the data from Table **13.5.1**, we can calculate the expected frequency for cell R1,C1, the college sport watchers who used sports at their primary criteria, to be:

$$E_{R1,C1} = \frac{87 * 68}{168} = 35.21$$

We can follow the same math to find all the expected values for all of Expected frequencies in each cell.

Example 13.5.1

Use the Observed frequencies from Table **13.5.1** to find the Expected frequencies for each combination of the two qualitative variables.

Solution

Table **13.5.2**: Table of EXPECTED Frequences of College Sports and College Decision

College Sports	Primary Affect	Somewhat Affected	Did Not Affect Decision	*Total*
Watched	$E_{R1,C1} = \frac{87 \times 68}{168} = 35.21$	$E_{R1,C2} = \frac{87 \times 49}{168} = 25.38$	$E_{R1,C1} = \frac{87 \times 51}{168} = 26.41$	$\sum_{Row} = 87$
Did Not Watch	32.79	23.63	24.59	$\sum_{Row} = 81$
Total	$\sum_{Column} = 68$	$\sum_{Column} = 49$	$\sum_{Column} = 51$	$\sum_{Row} = 168$

Notice that the marginal values still add up to the same totals as before. Crazy! This is because the expected frequencies are just row and column averages simultaneously. Our total N will also add up to the same value.

These 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 can do that, though, we should look at the hypotheses and the critical value table.

This page titled 13.5: Introduction to Test of Independence is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **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.

13.6: Practice Chi-Square Test of Independence- College Sports

We will be using the frequencies of watching college sports and how much college sports teams affect the choice of college attendance again, so here is the contingency table of the Observed frequencies and the table with the Expected frequencies that have been presented before. They are being presented again so that you don't have to go back and forth to find all of the data.

Here's the contingency table with the Observed frequencies:

Table **13.6.1**: Contingency Table of College Sports and College Decision

College Sports	Primary Affect	Somewhat Affected	Did Not Affect Decision	*Total*
Watched	47	26	14	*87*
Did Not Watch	21	23	37	*81*
Total	*68*	*49*	*51*	*168*

and the Expected frequencies:

Table **13.6.2**: Table of EXPECTED Frequencies of College Sports and College Decision

College Sports	Primary Affect	Somewhat Affected	Did Not Affect Decision	*Total*
Watched	35.21	25.38	26.41	*87*
Did Not Watch	32.79	23.63	24.59	*81*
Total	*68*	*49*	*51*	*168*

Now we're ready to follow the same 4-step procedure that you've come to know.

Step 1: State the Hypotheses

Chi-Square tests on patterns of relationship, and this doesn't change when we have more than one variable. The research hypothesis does not need to specify how each cell relates to each other like in a factorial ANOVA because we generally don't do pairwise comparisons for Chi-Square analyses. Instead, you can describe a general pattern of relationship between the two variables.

Example 13.6.1

What is the research hypothesis in words for this scenario? Make sure to describe a general pattern.

Solution

- Research hypothesis in words: There will be a pattern of difference such that there will be more people whose college decision was affected by college sports AND who watched college sports.

Remember, the hypotheses in symbols we can only say that the probabilities will not be equal. To determine that, let's figure out what the probability would be if all of the cells were equal. To find that out, we would divide a probability of 100% by the number of cells. There are six cells (Affected Decision=3; Watched=2; $\mathbf{2 \times 3 = 6}$).

$$\frac{100}{6} = 16.67$$

So the probability that any random participant will fall into a specific cell is 0.167 for each cell.

Example 13.6.2

What is the research hypothesis in symbols for this scenario?

Solution

- Research hypothesis in symbols: $P_{EachCell} \neq 0.167$.

If this research hypothesis in symbols doesn't make sense, it might be easier to start with a null hypothesis in words and symbols, then figure out how that works out for the research hypothesis. But honestly, it's not a huge deal if you don't get the probability part of the hypotheses. The important point is that you are testing the null hypothesis that all frequencies will be similar (no pattern of relationship), but that you actually expect a particular pattern (research hypothesis).

Example 13.6.3

What is the null hypothesis in words and symbols for this scenario?

Solution

- Null hypothesis in words: There is no pattern of difference in watching college sports affect college decisions.
- Null hypothesis in symbols: $P_{EachCell} = 0.167$.

Before we move on to an easier step, let's stop and remind ourselves that, just like correlations, Chi-Square tests cannot show that watching sports *caused* people to choose their college differently. This scenario is set up to make you think that the IV is watching college sports and the DV is the person's choice of college, but a Chi-Square can't test whether one thing causes another; these statistical analyses can only show if there's a *pattern of relationship or not*. The design of the experiment (how we collect the data to rule out alternative causes) is how we can show that one variable causes changes in another. This scenario is basically asking the participants *if they think* that college sports affected their choice of college. It's a small distinction, but an important one.

Step 2: Find the Critical Value

Okay, moving on from the "correlation doesn't equal causation" rant applied to Chi-Square!

Our critical value will come from the same table that we used for the Goodness of Fit Chi-Square 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. This is described at the bottom of the Critical Value of Chi-Square Table page, and looks like this:

- χ^2_{ToI} Test of Independence: $(R-1) \times (C-1)$
  - R is the number of rows
  - C is the number of columns

What this means is that the number of rows minus one is multiplied by the number of columns minus one. In our example:

$$df = (2-1)(3-1) = 1 \times 2 = 2$$

Exercise 13.6.1

What is the critical value for this scenario from the p = 0.05 column?

Answer

With our df = 2 ($df = (2-1) * (3-1) = 1 \times 2 = 2$), the critical value is 5.991.

Step 3: Calculate the Test Statistic

You probably won't believe it, but you finally have caught a break in learning formulas. The formula for Chi-Square's Goodness of Fit test is the same formula for the Chi-Square Test of Independence!

$$\chi^2 = \sum_{Each} \left(\frac{(E-O)^2}{E} \right)$$

If you find a way to combine Table **13.6.1** and Table **13.6.2** with the Differences, Differences Squared and divided by the Expected frequences into one table, then you are a data visualization wizard! For now, we'll create a new table for each step of the formula.

Example 13.6.4

Use the previous two tables (Table **13.6.1** and Table **13.6.2**) to create a table of differences by subtracting the Observed frequencies from the Expected frequencies for each cell.

Solution

Here is a table of differences:

Table **13.6.3**: Table of Difference of Observed & Expected Frequencies of College Sports and College Decision

College Sports	Primary Affect	Somewhat Affected	Did Not Affect Decision	*Total*
Watched	35.21-47=-11.79	25.38-26=-0.62 (or -0.63)	26.41-14=12.47	0
Did Not Watch	32.79-21=11.79	23.63-23=0.63	24.59-37=-12.41	0
Total	0	0 (or 0.01)	0	0 (or 0.01)

Notice that the row and column Totals are zero (or nearly so, depending on rounding). This is another good calculation check!

So far, we've accomplished the part in the parentheses of the formula:

$$\chi^2 = \sum_{Each} \left(\frac{(E-O)^2}{E} \right)$$

What is the next step in this formula?

Example 13.6.4

Square the difference scores in each cell of Table **13.6.3**.

Solution

Table **13.6.4**: Table of Squared Differences of Observed & Expected Frequencies of College Sports and College Decision

College Sports	Primary Affect	Somewhat Affected	Did Not Affect Decision
Watched	139.00	0.38	154.01
Did Not Watch	139.00	0.40	154.01

The row and column for the Total was removed because those sums aren't used for anything. You can calculate them for completeness, but it won't help you finish the formula. Unless it helps you not get lost!

Now we've finished the numerator of the formula:

$$\chi^2 = \sum_{Each} \left(\frac{(E-O)^2}{E} \right)$$

What is the next step?

Example 13.6.5

Using the squared differences in Table **13.6.4**:, complete the formula by dividing each cell by it's own Expected frequency (found in Table **13.6.2**). Then, add up all of the Total rows and columns to get the calculated χ^2.

Solution

Table **13.6.5**: Table of Squared Difference of Observed & Expected Frequencies of College Sports and College Decision Divided by Expected

Frequencies

College Sports	Primary Affect	Somewhat Affected	Did Not Affect Decision	*Total*
Watched	$\frac{139}{35.21} = 3.95$	$\frac{0.38}{25.38} = 0.01$	$\frac{154.01}{26.41} = 5.83$	$\sum_{Row} = 9.79$
Did Not Watch	$\frac{139}{32.79} = 4.24$	$\frac{0.40}{23.63} = 0.02$	$\frac{154.01}{24.59} = 6.26$	$\sum_{Row} = 10.52$
Total	$\sum_{Column} = 8.19$	$\sum_{Column} = 0.03$	$\sum_{Column} = 12.09$	$\sum_{Column} = 20.31$ $\sum_{Row} = 20.31$

The Total for summing the Totals for the columns is the same as the sum of the Totals for the rows, so we did it correctly!

Okay, let's see if, for the Test of Independence, doing your calculations in the formula might be easier.

Example \(\PageIndex{6}\)

Use the Chi-Square formula to calculate the χ^2 statistic

Solution

Using the information in Table **13.6.1** and Table **13.6.2**, we find:

$$\chi^2 = \frac{(35.21-47)^2}{35.21} + \frac{(25.38-26)^2}{25.38} + \frac{(26.41-14)^2}{26.41} + \frac{(32.79-21)^2}{32.79} + \frac{(23.62-23)^2}{23.62} + \frac{(24.59-37)^2}{24.59}$$

$$\chi^2_{Diff} = \frac{(-11.79)^2}{35.21} + \frac{(-0.62)^2}{25.38} + \frac{(12.41)^2}{26.41} + \frac{(11.79)^2}{32.79} + \frac{(0.63)^2}{23.62} + \frac{(-12.41)^2}{24.59}$$

$$\chi^2_{DiffSquared} = \frac{139}{35.21} + \frac{0.38}{25.38} + \frac{154.01}{26.41} + \frac{139.00}{32.79} + \frac{0.40}{23.62} + \frac{154.01}{24.59}$$

$$\chi^2_{Division} = 3.95 + 0.01 + 5.83 + 4.24 + 0.02 + 6.26 = 20.31$$

$$\chi^2 = 20.31$$

What do you think? Was it easier to do the calculations in five different tables, or do it all in one formula? There's no right answer for this, it really is what's easier for you.

But now, we're ready to make a decision!

Step 4: Make the Decision

What is the final decision?

Exercise 13.6.2

Should the null hypothesis be retained or rejected?

Answer

Our calculated χ^2=20.31, and the critical χ^2 was 5.991, so we would reject the null hypothesis. Our calculated value is so extreme that we would expect it less than 5% of the time if there really was no pattern of relationship between the two qualitative variables.

So what would the statistical sentence look like?

? Exercise 13.6.3

What would our results look like in the statistical sentence?

Answer

$\chi^2(2)=20.31$, $p<.05$

Let's use all that we've done to let people know what we found in...

The Write-Up

Can you write this up with the four requirements for reporting results but without descriptive statistics? You can include all of the Observed frequencies, but that gets clunky. A good way around that is to refer to the original table of Observed frequencies.

✓ Example 13.6.7

Report the results in a concluding paragraph that includes the four requirements.

Solution

The research hypothesis was that there will be a pattern of difference such that there will be more people whose college decision was affected by college sports AND who watched college sports A pattern of difference was found ($\chi^2(2)=20.31$, $p<.05$). As can be seen in Table **13.6.1**, this research hypothesis was not supported. People who watched college sports seem to believe that they used that to choose their college, and people who didn't watch college sports seem to believe that they did not use college sports to make their decision about which college to choose.

Did you notice all of the "seems like" and "they believe" in that concluding paragraph? Yeah, that's how scientists write. Because science is cumulative, each of one us adds one piece of evidence to a pile that supports one idea. In this case, the idea was that people thought that their choice of college was affected by whether they watched college sports or not. In the Goodness of Fit example, the idea that was supported was that there are about the same amount of people who like and dislike pineapples on pizza. But one study is never conclusive. Instead, many, many scientists conduct many, many studies. Some of them show reality, but some of them ($p<.05$) find results from their sample that do not match the reality in the population. It can be hard for non-scientists because they might just see us being wishy-washy about our results when we are really following the guidelines of the null hypothesis significance testing procedure.

Let's try one more example so that we've got this Chi-Square thing down.

This page titled 13.6: Practice Chi-Square Test of Independence- College Sports is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **14.7: College Sports** by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.
- **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.
- **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.
- **11.1: Chi-Square Tests for Independence** by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/beginning-statistics.

13.6.1: Practice- Fast Food Meals

This example again uses data from fast food restaurants from OpenIntro.org. We will be looking at all of the non-chicken meals this time, and frequencies of sandwich type with the two categories being hamburgers (hamburgers, cheeseburgers, and sliders) or sandwich (sandwiches, wraps, or rolls) by whether they are considered a kids/juniors meal or for adults Table **13.6.1.1** shows a contingency table with the Observed frequencies:

Table **13.6.1.1**: Table of Frequencies of Sandwich Type and Meal Type

Number of Meals	Adult Meal	Kids Meal	Total
Hamburger	74	6	$\sum_{Row}$ =80
Sandwich	98	5	$\sum_{Row}$ =103
Total	$\sum_{Column}$ =172	$\sum_{Column}$ =11	$\sum_{Row}$ =183 $\sum_{Column}$ =183

We're cutting it close to violating the assumptions of this type of analysis with only five meals for kids that are sandwiches, but we have the minimum that is required.

Step 1: State the Hypotheses

We'll start with the research hypothesis. Look at Table **13.6.1.1**. In general, what kind of meal is more common (Adult or Kids) for which sandwich type (hamburger or sandwich)? Use that to determine the research hypothesis.

Example 13.6.1.1

What is the research hypothesis in words for this scenario? Make sure to describe a general pattern.

Solution

- Research hypothesis in words: There will be a pattern of difference such that there will be more kids meals with hamburgers than with sandwiches.

The research hypothesis in symbols is that the probabilities will *not* be equal to each other. To determine that, let's figure out what the probability would be if all of the cells were equal. To find that out, we would divide a probability of 100% by the number of cells. There are four cells (Sandwich Type = 2; Meal Type = 2; $\mathbf{2 \times 2 = 4}$).

$$\frac{100}{4} = 25.00$$

So the probability that any random participant will fall into a specific cell is = 0.25 for each cell. This works out to be:

- Research hypothesis in symbols: $\boldsymbol{P_{EachCell} \neq 0.25}$.

Sometimes it's easier to start with a null hypothesis in words and symbols, then figure out how that works out for the research hypothesis.

Exercise 13.6.1.1

What is the null hypothesis in words and symbols for this scenario?

Answer

- Null hypothesis in words: There is no pattern of difference by type of meals and type of sandwich.
- Null hypothesis in symbols: (P_{EachCell} = 0.25\)

Step 2: Find the Critical Value

Let's find the critical value, the easiest step in our process!

? Exercise 13.6.1.2

Using the Critical Value of Chi-Square Table page (or found in the Common Critical Value Tables page at the back of this book), determine the critical value for this scenario from the p = 0.05 column.

Answer

To determine the critical value, we much first find the correct Degrees of Freedom. The formula (found on the bottom of the page for the Critical Values of Chi-Square Table)

Degrees of Freedom (df) for χ^2 Test of Independence: $(R-1) \times (C-1)$

Since each group has two levels, that's: $df = (2-1) \times (2-1) = 1 \times 1 = 1$

With our df=1, the critical value is 3.841.

Step 3: Calculate the Test Statistic

We will again calculate the χ^2 statistics through the use of multiple tables, and then again by filling in the formula for each category:

$$\chi^2 = \sum_{Each} \left(\frac{(E-O)^2}{E} \right)$$

If you prefer one way to calculate over the other, feel free to skip the one that you don't like!

Calculating with Tables

We will start with the method using tables. To complete the χ^2 formula:

$$\chi^2 = \sum_{Each} \left(\frac{(E-O)^2}{E} \right)$$

We first need to find the difference between the Expected frequency and the Observed frequency. In case you forgot, that's calculated by multiplying the row total by the column total, then dividing by the total N for each cell:

$$E_{EachCell} = \frac{RT * CT}{N}$$

✓ Example 13.6.1.2

Find the Expected frequencies for the Observed frequencies in Table 13.6.1.1.

Solution

Table 13.6.1.2: Table of Expected Frequencies of Sandwich Type and Meal Type

Number of Meals	Adult Meal	Kids Meal	Total
Hamburger	$E = \frac{80 * 172}{183} = 75.19$	$E = \frac{80 * 11}{183} = 4.81$	$\sum_{Row}$ =80
Sandwich	$E = \frac{103 * 172}{183} = 96.81$	$E = \frac{103 * 11}{183} = 6.19$	$\sum_{Row}$ =103
Total	$\sum_{Column}$ =172	$\sum_{Column}$ =11	$\sum_{Row}$ =183 $\sum_{Column}$ =183

Make sure to compute the row and column totals. They should be the same as the Totals in Table 13.6.1.1. If they are not, then you made a computation mistake. It's easier to fix it now than after the whole process when you figure out that your final answer isn't the same as everyone else's.

Based on the formula, what would we do next?

$$\chi^2 = \sum_{Each} \left(\frac{(E-O)^2}{E} \right)$$

Right, E-O, which means to subtract the Observed frequency from each cell from the Expected frequency from that same cell. If you're doing this all on your calculator, it might be easier to do the following step at the same time. Based on the formula, what would we do after subtracting the Observed from the Expected frequencies?

So in the following table, you can do the subtraction and squaring in the same table.

Example 13.6.1.3

Find the difference scores by subtracting the Observed frequencies from the Expected frequencies for each cell, then square each of those difference scores.

Solution

Table 13.6.1.3: Table of Differences Squared

Number of Meals	Adult Meal	Kids Meal
Hamburger	$75.19–74 = 1.19^2 = 1.42$	$4.81–6 = -1.19^2 = 1.42$
Sandwich	$96.81–98 = -1.19^2 = 1.42$	$6.19–5 = 1.19^2 = 1.42$

How come all of the differences scores are the same?! Maybe that happens when one group has so few frequencies? Maybe that happens when there's so little difference between at least two of the cells? It'll be interesting to see what that means for the final calculation!

If you complete two separate tables, one for the subtraction and one for the squaring, the row and column Totals for the differences (subtraction) would be zero. Dr. MO did separate tables on her own, and that's what she found!

Looking at the formula, we've now completed all of the calculations for the numerator. Can you see what the next step is?

$$\chi^2 = \sum_{Each} \left(\frac{(E-O)^2}{E} \right)$$

Example 13.6.1.4

Complete the formula by dividing each cell by its own Expected frequency, then, add up all of the Total rows and columns to get the calculated χ^2.

Solution

Table 13.6.1.4: Final Table Calculating Chi-Square

Number of Meals	Adult Meal	Kids Meal	Total
Hamburger	$E = \frac{80 * 172}{183} = 75.19$	$E = \frac{80 * 11}{183} = 4.81$	$\sum_{Row}$ =80
Sandwich	$E = \frac{103 * 172}{183} = 96.81$	$E = \frac{103 * 11}{183} = 6.19$	$\sum_{Row}$ =103
Total	$\sum_{Column}$ =172	$\sum_{Column}$ =11	$\sum_{Row}$ =183 $\sum_{Column}$ =183

The Total for summing the Totals for the columns is the same as the sum of the Totals for the rows, so we did it correctly!

Ooh, that looks small. Let's see if we use the formula for the calculations gets the same answer.

Calculating with the Formula

Example 13.6.1.5

Use the Chi-Square formula to calculate the χ^2 statistic.

Solution

Using the information in Table **13.6.1.1** and Table **13.6.1.2**, we find:

$$\chi^2 = \frac{(75.19-74)^2}{75.19} + \frac{(4.81-6)^2}{4.81} + \frac{(96.81-98)^2}{96.81} + \frac{(6.19-5)^2}{6.19}$$

$$\chi^2_{Diff} = \frac{(1.19)^2}{75.19} + \frac{(-1.19)^2}{4.81} + \frac{(-1.19)^2}{96.81} + \frac{(1.19)^2}{6.19}$$

$$\chi^2_{Squared} = \frac{1.42}{75.19} + \frac{1.42}{4.81} + \frac{1.42}{96.81} + \frac{1.42}{6.19}$$

$$\chi^2_{Division} = 0.02 + 0.30 + 0.01 + 0.23 = 0.56$$

$$\chi^2 = 0.56$$

The good news is that both methods of calculation got us the same answer! Now, we're ready to make a decision about the null hypothesis.

Step 4: Make the Decision

What do you think when this is still true:

$$\textbf{(Critical} < \textbf{Calculated) = Reject null = There is a pattern of relationship.} = p < .05$$

$$\textbf{(Critical} > \textbf{Calculated) = Retain null = There is no pattern of relationship.} = p > .05$$

Exercise 13.6.1.3

Should the null hypothesis be retained or rejected? What should the statistical sentence look like?

Answer

Our calculated χ^2=0.56 is much smaller than the critical χ^2 of 3.841, so we would retain the null hypothesis. This leads to the statistical sentence of χ^2(1)=0.56, p>.05.

We have all we need to write-up a conclusion now.

The Write-Up

Example 13.6.1.6

Report the results in a concluding paragraph with the four requirements for reporting results but refer to Table **13.6.1.1**: for the Observed frequencies since we don't have descriptive statistics.

Solution

The research hypothesis was that there will be a pattern of difference such that there will be more kids meals with hamburgers than with sandwiches. This was not supported; there is no statistically significant pattern of difference (χ^2(1)=0.56, p>.05). As can be seen in Table **13.6.1.1**, although there were more kids meals with hamburgers than sandwiches, there was about the same amount of kids meals with hamburgers as there were with sandwiches.

And that’s it! That’s Pearson’s Chi-Square Test of Independence! Not so scary, right? One last type of Chi-Square, and then we’ll talk about choosing the appropriate statistical test again.

Contributors and Attributions

- Foster et al. (University of Missouri-St. Louis, Rice University, & University of Houston, Downtown Campus)

This page titled 13.6.1: Practice- Fast Food Meals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

13.7: RM Chi-Square- The McNemar Test

Suppose you've been hired to work for the *Generic Political Party* (GPP), and part of your job is to find out how effective the GPP political advertisements are. So, what you do, is you put together a sample of N=100 people, and ask them to watch the GPP ads. Before they see anything, you ask them if they intend to vote for the GPP; and then after showing the ads, you ask them again, to see if anyone has changed their minds. Obviously, if you're any good at your job, you'd also do a whole lot of other things too, but let's consider just this one simple experiment. One way to describe your data is via the following contingency table:

Table **13.7.1**- Voting and Advertisement Counts

Voting and Ads	Before	After	Total
Yes Vote	30	10	40
No Vote	70	90	160
Total	100	100	200

At first pass, you might think that this situation lends itself to the Pearson χ^2 Test of Independence. However, a little bit of thought reveals that we've got a problem: we have 100 participants, but 200 observations. This is because each person has provided us with an answer in *both* the before column and the after column. What this means is that the 200 observations aren't independent of each other: if voter A says "yes" the first time and voter B says "no" the first time, then you'd expect that voter A is more likely to say "yes" the second time than voter B. The consequence of this is that the usual χ^2 test won't give trustworthy answers due to the violation of the independence assumption (found in the section on Assumptions of Chi-Square tests). Now, if this were a really uncommon situation, I wouldn't be bothering to waste your time talking about it. But it's not uncommon at all: this is a *standard* repeated measures design, and none of the tests we've considered so far can handle it. (You might immediately think about the Phi correlation, Dr. MO certainly did! But according to MathCracker. com, Phi is a χ^2 but with an extra step, so it would have the same assumptions as all Chi-Square analyses- no dependent data).

Eek.

The solution to the problem was published by McNemar (1947). The trick is to start by tabulating your data in a slightly different way:

Table **13.7.2**- Rearranged Voting and Advertisement Counts

	Before: Yes	Before: No	Total
After: Yes	5	5	10
After: No	25	65	90
Total	30	70	100

This is exactly the same data, but it's been rewritten so that each of our 100 participants appears in only one cell. Because we've written our data this way, the independence assumption is now satisfied, and this is a contingency table that we *can* use to construct an χ^2 Goodness of Fit statistic. However, as we'll see, we need to do it in a slightly nonstandard way. To see what's going on, it helps to label the entries in our table a little differently:

Table **13.7.3**- Cells Labeled

	Before: Yes	Before: No	Total
After: Yes	a	b	a+b
After: No	c	d	c+d
Total	a+c	b+d	n

Next, let's think about what our null hypothesis is: it's that the "before" test and the "after" test have the same proportion of people saying "Yes, I will vote for GPP". Because of the way that we have rewritten the data, it means that we're now testing the hypothesis that the *row totals* and *column totals* come from the same distribution. Thus, the null hypothesis in McNemar's test is that we have "marginal homogeneity," meaning that the row totals and column totals have the same distribution: $P_a+P_b=P_a+P_c$, and similarly that $P_c+P_d=P_b+P_d$. Notice that this means that the null hypothesis actually simplifies to $P_b=P_c$. In other words, as far as the McNemar test is concerned, it's only the off-diagonal entries in this table (i.e., b and c) that matter! After noticing this, the McNemar test of marginal homogeneity is not that different to a usual χ^2 test.

Since the calculation is so similar to χ^2 we won't be going over it. If we ran a McNemar's test to determine if people were just as likely to vote GPP after the ads as they were before hand, we would find statistically significant difference ($\chi^2(1)=12.04,p<.001$), suggesting that the groups were not just as likely to vote GPP after the as as before. But look closely before you recommend dumping money into the advertising budget! It looks like the ads had a *negative* effect: people were *less* likely to vote GPP after seeing the ads. (Which makes a lot of sense when you consider the quality of a typical political advertisement.)

As always, if you are doing statistics for graduate school or your job, you'll have software that will do all of this for you. For now, you are learning the formulas for two reasons:

1. The formulas show you what is happening (mathematically) so that you understand the results better.
2. Being able to work through a formula helps with your logic, reasoning, and critical thinking skills.

Speaking of critical thinking, let's get to the final section of this chapter: Choosing the Correct Statistical Analysis!

Reference

McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*, 153-157.

This page titled 13.7: RM Chi-Square- The McNemar Test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **12.8: The McNemar Test** by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.
- **16.7: RM Chi-Square- The McNemar Test** by Michelle Oja is licensed CC BY-SA 4.0.

13.8: Choosing the Correct Test- Chi-Square Edition

We've discussed which test to choose previously (after learning about t-tests, and again after learning about ANOVAs), and we will again soon! This section is just to help decide which Chi-Square analysis to use when you only have qualitative variables.

Which Chi-Square?

You have seen the χ^2 test statistic used in two different circumstances. The bulleted list is a summary that will help you decide which χ^2 test is the appropriate one to use.

- **Goodness-of-Fit:** Use the goodness-of-fit test to decide whether a population with an unknown distribution "fits" a known distribution. In this case there will be a single qualitative survey question or a single outcome of an experiment from a single population. Goodness-of-Fit is typically used to see if the population is uniform (all outcomes occur with equal frequency), the population is normal, or the population is the same as another population with a known distribution. In short:
  - Use the χ^2 Goodness of Fit when you only have one qualitative variable, and when
  - You are testing if the pattern of frequencies is about equal in all of the categories.
- **Independence:** Use the test for independence to decide whether two qualitative variables (factors) are independent or dependent. In this case there will be two qualitative survey questions or experiments and a contingency table will be constructed. The goal is to see if the two variables are unrelated (independent) or related (dependent). In short:
  - Use the χ^2 Test of Independence when you only have two or more qualitative variables, and when
  - You are testing if the pattern of frequencies is about equal in *all* of the combined categories.

In sum, the Goodness of Fit test is typically used to determine if data fits a particular distribution while the Test of Independence makes use of a contingency table to determine the independence of two factors.

Practice

? Exercise 13.8.1

Which test do you use to decide whether an observed distribution of frequencies of a qualitative variable is the same as an expected distribution?

Answer

Goodness of Fit test

? Exercise 13.8.2

Which test would you use to decide whether two qualitative factors factors have a pattern of relationship?

Answer

Test of Independence

On to some practice problems, then a wrap-up of everything that you've learned! You should be so impressed with yourself. I know that I am!

This page titled 13.8: Choosing the Correct Test- Chi-Square Edition is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- **11.6: Comparison of the Chi-Square Tests** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

CHAPTER OVERVIEW

14: Wrapping it Up

14.1: Introduction to Wrapping Up

14.2: Choosing the Test

14.3: Why did you take this class?

This page titled 14: Wrapping it Up is shared under a not declared license and was authored, remixed, and/or curated by Michelle Oja.

14.1: Introduction to Wrapping Up

You've learned the what, how, and why of introductory behavioral statistics! You've done so great to get so far! Before you move on to your next class, there are a couple things that we'll reinforce in this chapter.

The first thing that we'll reinforce is how to decide which statistical test to run. Then, you can practice interpreting reported results just like you would see a research article. We'll end by circling back to why you took this class in the first place.

This page titled 14.1: Introduction to Wrapping Up is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

14.2: Choosing the Test

You might remember that we already talked about which statistical test to run at the end of the chapters on ANOVA. In that section, Dr. MO said that there are three things about your variables that determined which statistical analysis is most appropriate; these were:

- The type of DV (qualitative or quantitative or ranked/ordinal)
- The number of groups
- Whether the groups are independent or dependent or the population

In an Example, you described what types of variables (based on these three descriptions) fit all of the analyses that we had covered at that point:

- One-Sample t-test
  - DV: Compares means, so quantitative
  - Number of Groups: Compares a sample to the population
  - Type of Group: The population
- Two-sample independent t-test
  - DV: Compares means, so quantitative
  - Number of Groups: Compares two samples
  - Type of Group: Independent (unrelated)
- Two-sample dependent t-test
  - DV: Compares means, so quantitative
  - Number of Groups: Compares two samples
  - Type of Group: Dependent (related)
- Between Groups ANOVA
  - DV: Compares means, so quantitative
  - Number of Groups: Compares two or more levels of an IV
  - Type of Group: Independent (unrelated)
- Repeated Measures ANOVA (also called Within Groups ANOVA)
  - DV: Compares means, so quantitative
  - Number of Groups: Compares two or more levels of an IV
  - Type of Group: Dependent (related)
- Mann-Whitney U
  - DV: Ranked data, or we assume that the distribution is NOT normally distributed.
  - Number of Groups: Compares two groups
  - Type of Group: Independent (unrelated)
- Wilcoxon Match-Pair Signed-Rank test
  - DV: Ranked data, or we assume that the distribution is NOT normally distributed.
  - Number of Groups: Compares two groups
  - Type of Group: Dependent (related)
- Kruskal-Wallis One-Way ANOVA
  - DV: Ranked data, or we assume that the distribution is NOT normally distributed.
  - Number of Groups: Compares two or more groups
  - Type of Group: Independent (unrelated)
- Friedman's test
  - DV: Ranked data, or we assume that the distribution is NOT normally distributed.
  - Number of Groups: Compares two or more groups

- Type of Group: Dependent (related)

But since then, we've also learned about correlations, regression, and Chi-Squares. What we learned about regression is mostly focused on extending what we learned about correlations, so we won't look at them again right now. However, regression is one of the most powerful predictive tools that statisticians have, so you will learn a LOT more about this type of analysis if you take more advanced statistic courses. For now, let's look at describe the three characteristics for each of the correlations and Chi-Square analyses that we have learned about.

Example 14.2.1

Identify the three criteria (type of DV, number of groups, and type of group) for the following analyses:

- Pearson's Correlation
- Spearman's Rank Correlation
- Phi Correlation
- Chi-Square Goodness of Fit
- Chi-Square Test of Independence
- McNemar Test

Solution

- Pearson's Correlation
  - DV: Two quantitative variables (not two levels of one IV, they are two entirely different variables).
  - Number of Groups: Compares two groups
  - Type of Group: We are testing to see if they are linearly related.
- Spearman's Rank Correlation
  - DV: Two ranked (ordinal) variables.
  - Number of Groups: Compares two groups
  - Type of Group: We are testing to see if they are related.
- Phi Correlation
  - DV: Two binary variables.
  - Number of Groups: Compares two groups
  - Type of Group: We are testing to see if they are related.
- Chi-Square Goodness of Fit
  - DV: One qualitative variable with two or more levels
  - Number of Groups: Compares two or more groups or categories
  - Type of Group: Independent
- Chi-Square Test of Independence
  - DV: Two qualitative variables (not two levels of one IV, they are two entirely different variables).
  - Number of Groups: Compares two groups, each with two or more levels or categories.
  - Type of Group: Independent
- McNemar Test
  - DV: Two qualitative variables (not two levels of one IV, they are two entirely different variables).
  - Number of Groups: Compares two groups, each with two or more levels or categories.
  - Type of Group: Dependent

Another way to think about choosing the most appropriate test is through a flow chart or decision tree. A flow chart or decision tree could cover the three basic questions about the variables, and branch out depending on the answer.

Note

You might consider drawing a flow chart yourself with only the statistical tests that your professor has focused on.

For now, you can check out this PDF of a Decision Tree Handout created by Dr. MO to see what this flow chart might look like with most of the analyses that we've covered so far. Here is the link to the interactive website to help you practice when to use which kind of statistical analysis that was provided in that previous ANOVA chapter. To start, click on the kinds of analyses that you want to be tested on, then hit Submit.

You might think that you don't need to learn about these things because statistical software will know which analysis to run. This is both true and not true. Sometimes, the software won't let you run the analysis with the wrong type of variables. For example, if you try to have SPSS run a Chi-Square with quantitative variables, it will say that it can't. But, if you try to run a Chi-Square with qualitative variables that are related, *it will*. But we know that one of the basic requirements of Chi-Square is that the groups must be independent (meaning that the same people can't be in both groups, the groups can't be related). So, you can't trust the software to know what's it's doing. You have to be the one who knows what you're doing!

What the software definitely won't do is tell you what it means. Let's turn to that idea next.

Next, let's finish the textbook by revisiting why you had to take this class. This will also be a great refresher on all that you've learned!

This page titled 14.2: Choosing the Test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

14.3: Why did you take this class?

Why did you take this class?

Do you remember, way back, months and months ago, when you first read Chapter 1? There was a whole section about why this course might be required for your major or helpful for you. In this section, we'll review those ideas to see if you agree that you learned the information and it might be important.

> *First reason that you are taking a statistics class: Understanding statistics will help make you better in your future career.*
>
> *Second reason that you are taking a statistics class: You will need to report evidence to show that what you are doing helps your clients/patients/customers.*

Do you think that you might be better at your career because you can understand any statistics that you come across related to your field? In particular, you now can calculate, interpret, and report descriptive statistics and graphs, and statistically compare means or find relationships. Let's break that apart a little.

Descriptive statistics include measures of central tendency, measures of variability, and graphs.

Example 14.3.1

What are the three typical measures of central tendency? What is the typical measure of variability?

Solution

The three typical measures of central tendency are the mean, median, and mode. The typical measure of variability is the standard deviation, although the range can also be useful.

Do you remember how to calculate them? We won't be testing on this because that's what the first unit was all about, but you can go back to some of the practice pages or exercises in the first unit to make sure that you understand the formulas. That would also be a great way to show yourself how much you've learned! It will also support Dr. MO's belief that you now understand statistics and can report them.

Hopefully you can see that reporting the descriptive statistics of your clients/patients/customers can be useful in any career to take a "temperature check" on how things are going now. The next step is to make some changes and use these statistical analyses to compare group means, or find relationships between variables to decide what changes might meaningfully affect your outcome variable (DV).

Understanding Statistics

Now, here's the clincher. If you were provided (or calculated) descriptive statistics for a topic in your (future) career field, do you think that you could use them to understand the topic? Again, we won't be testing on this, but hopefully you can see how useful it is to understand what descriptive statistics are tell you.

The next part of understanding statistics is about statistically comparing means (t-tests or ANOVAs) or finding relationships (correlations, regression, Chi-Square). Although we definitely won't be seeing if you remember how to calculate all of those, let's see if you remember how to interpret them! We'll start with t-tests.

Example 14.3.2

For future managers: Manager often want to use surveys to understand the thoughts and behaviors of their clients. Although it is cheaper to use online surveys because it saves on printer costs, does it provide as much data as in-person surveys? As a future manager, you might want to know which type of survey provides the most responses. Using the surveys that your agency has implemented in the last three years (three online surveys and three in-person surveys, all aiming for 100 respondents, you find that $t(4)=6.78$, $p<.05$, with the average number of responses for online surveys is 22 and the average

number of responses for the in-person surveys is 88. In the future, would you use online or in-person surveys? Why or why not?

Solution

In-person surveys should be the primary method of surveying clients because the average number of responses is statistically significantly higher. .

That was an independent samples t-test.

? Exercise 14.3.1

What is the t-test with two samples, other than an independent samples t-test?

Answer

There are three kinds of t-tests:

- One-sample (compares a sample to a population)
- Independent samples (compares two independent samples)
- Dependent samples (compares two dependent samples)

So the other kind of t-test with two samples is dependent samples t-test.

Let's try interpreting the results of a dependent samples t-test.

✓ Example 14.3.3

For future therapists: Sleep deprivation is a serious issue that affects physical health, memory, and safety. Some non-pharmaceutical solutions include meditation or exercises. Imagine that a therapist had 50 clients suffering from insomnia, so the therapist had half of the clients try meditating for 10 minutes every day for two weeks, and then those same clients stopped meditating and exercised for 30 minutes every day for two weeks. The other have of the clients started with the exercise condition for two weeks, then stopped exercising and meditated for two more weeks. The time it took the to fall asleep was measured at the end of each two week period for all clients so that we can see which method worked best. If $t(49)=1.23$, $p>.05$, with it taking an average of 65 minutes for all clients to fall asleep after the meditation condition and about 58 minutes, on average, for all clients to fall asleep after the exercise condition, which method to treat insomnia would you recommend, and why?

Solution

This is sort of a trick question because there is no statistical difference between the two groups, so there is not enough information to recommend either method.

The next set of examples will use the results for hypothetical ANOVAs.

✓ Example 14.3.1

For future health care workers: Medical professionals, like doctors and nurses, often admonish their patients to lose weight. If you were a medical professional and read a study that compared weight loss for three groups (exercise, low-fat diet, both exercise and low-fat diet) that found $F(2,87) = 6.78$, $p<.05$, with the average weight loss being:

- Exercise: 3.6 pounds
- Low-Fat Diet: 2.4 pounds
- Exercise & Low-Fat Diet: 11.1 pounds,

What would you recommend to your patients as a good way to lose weight?

Solution

Ack, another trick question! We don't have the pairwise comparisons to know which means are different from one another! Although it might be tempting to say that the combination works best, we can't know whether there's a difference between any of the conditions, but especially the other two conditions, without further information.

Let's turn the insomnia example into a repeated measures ANOVA.

? Exercise 14.3.2

For the future therapists again: The therapist has 50 clients suffering from insomnia. The order of the conditions didn't seem to matter, so this time all of the clients meditated for 10 minutes every day for two weeks, and then those same clients stopped meditating and exercised for 30 minutes every day for two weeks. The time it took to fall asleep was measured at the beginning of the study (before either treatment was started), end of the first set of two weeks, then at the end of the second set of two weeks. If $F(2,98)=2.31$, $p>.05$, with it taking an average of 72 minutes to fall asleep before any of the treatments, 56 minutes to fall asleep after the meditation condition, and an average of 65 minutes to fall asleep after the exercise condition, which method to treat insomnia would you recommend, and why?

Answer

This is a double trick question! There is no statistical difference between the three groups, so neither method can be recommended. If the ANOVA's null hypothesis had been rejected, we still wouldn't know which condition was statistically better because the pairwise comparisons were not analyzed. Maybe meditation worked better than the other conditions for this sample, but we can't be sure unless we statistically compare the mean of each group to each other group.

Factorial designs and regression equations are difficult to interpret, especially when the results are presented in such short paragraphs, so we'll move on to looking at relationships between variables.

✓ Example 14.3.5

For future criminal justice professionals: Crime rates declined in U.S. cities for many decades, but are slowly increasing. What has also been increasing is wealth disparity (the difference between the highest earners and the lowest earners). If these two things seemed to be linearly related, we might be able to decrease crime by decreasing wealth disparity. Based on this hypothetical result looking at crime rates and wealth disparity (measured as the difference between the top 1% and the bottom 1% of U.S. households) $r(48)=0.32$, $p>.05$, would looking at decreasing wealth disparity be one way to reduce crime rates?

Solution

Yes, based on the statistically significant positive correlation, as crime rates increase so does wealth disparity. You might be politically averse to decreasing wealth disparity, but based on the statistical results it might be a way to decrease crime rates.

Let's try this one on your own:

? Exercise 14.3.3

For students and future teachers: As a way to improve grades, some students re-write their notes while other students watch educational videos. Which one seems to work better at passing (or not) their classes if you conducted a study that found $\chi^2(1)=4.83$, $p<.05$, with more people passing who re-wrote their notes than the number of people passing how watched educational videos (and more people failed who watched videos than people who failed who re-wrote their notes)?

Answer

Since there seems to be a pattern of relationship between passing and type of studying such that those who re-wrote their notes were more likely to pass (than fail) compared to those who watched educational videos, it looks like re-writing notes is the best way to study!

use null hypothesis significance testing in your future career.

Dr. MO also wrote: “Finally, some of you will fall in love with statistics, and become a researcher for a living!” This won’t happen to all of you, or even most of you, but there’s always one or two students who realize that they can help people but not in the way that they imagined (like being a therapist or nurse). Was this you?

"Research shows that..."

The first chapter also had a short section on research studies. This section described how research hypotheses with measured IVs and DVS and representative samples tell you more accurate information than any other type information. Now that we’ve gone through many, many examples, hopefully this idea makes a little more sense.

Learn How to Learn

In closing out this whole adventure in behavioral statistics, Dr. MO hopes that not only did you learn how to calculate and interpret statistical analyses, but that you also learned a little how to learn. If you ended up not learning as much as you’d like, here is a summary of information presented in chapter 1 of what you can do in your next class (or your next attempt at this class).

(Re-)Watch this five-part video series from cognitive psychologist Dr. Chew, for how best to learn (and what to do if you fail:

- Part 1: Beliefs that Make You Fail
- Part 2: What Students Should Know about How People Learn
- Part 3: Cognitive Principles for Optimizing Learning
- Part 4: Putting the Principles of Optimizing Learning into Practice
- Part 5: I Blew the Exam, Now What?

In his videos, Dr. Chew emphasizes how important time and effortful thinking is for learning; if it was easy, everyone would be learning a lot! This article by Zanardelli Sickler (2017) provides more activities that you can do before, during, and after the lecture to make sure that you learn and understand the material. Here are even more ideas for what to do during the time you spend studying before or after class:

- Use the Exercises in this book as practice.
- Weekly meetings with a tutor or time in your schools math lab.
- Work in pairs to check each others’ work.

You can also decide what kinds of activities might work best to do on your phone (reading), and what activities you might want to wait until you have a bigger screen (quizzes) or have more space to write (practice problems).

College courses are hard, especially behavioral statistics (which combines math *and* English skills!), and you probably have more than just this one class and many other responsibilities. Be proud of yourself for what you’ve learned! And if you didn’t quite make it this time, use the tips provided to learn the material for your next go around. Dr. MO has taught this class for years, and has **never** met a student that couldn’t pass it (There have been a lot of students who didn’t pass, but **none** who just didn’t have the ability to pass it). That bears repeating: No students lacked the ability to pass. If you haven’t passed yet, try putting more time into studying, and studying in ways that require more effortful thinking. You can do this! And if you did pass this class, you will undoubtedly hit a class or a work activity that is more challenging for you, so try to remember:

- More time
- More effortful thinking

And finally, correlation doesn’t mean causation!

And with that, we’re out!

This page titled 14.3: Why did you take this class? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

Index

A

alternative hypothesis
- 6.10: The Alternative Hypothesis

B

Bonferroni test
- 9.5.1: Pairwise Comparison Post Hoc Tests for Critical Values of Mean Differences

box plots
- 2.9: Graphing Quantitative Data- Boxplots

C

contingency table
- 13.2.1: Critical Values of Chi-Square Table

D

descriptive statistics
- 2.1: Introduction to Looking at Data (This is what too many numbers looks like)

F

Frequency Polygons
- 2.7: Graphing Quantitative Data- Line Graphs

K

Kruskal–Wallis Test
- 9.8: Non-Parametric Analysis Between Multiple Groups

kurtosis
- 2.6: Skew and Kurtosis

L

leptokurtic
- 2.6: Skew and Kurtosis

line graph
- 2.7: Graphing Quantitative Data- Line Graphs

N

Normal distribution
- 4.1: Normal Distributions
- 5.4: Normal Distributions and Probability Distributions

P

PEMDAS
- 1.9: Math Refresher

pie chart
- 2.4: Graphing Qualitative Variables- Pie Charts
- 4.4: Introduction to the z table

pie charts
- 2.4: Graphing Qualitative Variables- Pie Charts

post hoc test
- 9.5.1: Pairwise Comparison Post Hoc Tests for Critical Values of Mean Differences

R

regression
- 12.5: Regression — A mini intro

S

skew
- 2.6: Skew and Kurtosis

skewness
- 2.6: Skew and Kurtosis

standard deviation
- 3.3: Spread and Variability

T

Tests for Independence
- 13.2.1: Critical Values of Chi-Square Table

Tukey’s Honest Significant Difference (HSD)
- 9.5.1: Pairwise Comparison Post Hoc Tests for Critical Values of Mean Differences

V

variability
- 3.3: Spread and Variability

variance
- 3.3: Spread and Variability

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Many of the pages in this book are re-mixes of open content from other books in the LibreTexts Library. Sometimes sections from several sources are used on the same page and there may be no unique open license. Where there is a conflict, the licensing may appear as *undeclared*. On any page, by toggling AutoAttibution on the Tools menu, you can display the detailed attribution on a paragraph level. Material from different books will be color coded and the source displayed by mousing over the paragraph.

Content in the Front and Back Matter, directories and similar can be uncopywriteable, being simple listings and these are also shown as *undeclared*.

Overview

Title: Intro to Stats in the Social Sciences (MPC)

Webpages: 162

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-SA 4.0: 54.9% (89 pages)
- CC BY-NC-SA 4.0: 25.3% (41 pages)
- Undeclared: 16% (26 pages)
- CC BY-NC 4.0: 2.5% (4 pages)
- Other: 0.6% (1 page)
- CC BY 4.0: 0.6% (1 page)

By Page

- Intro to Stats in the Social Sciences (MPC) - *CC BY-NC 4.0*
  - Front Matter - *CC BY-NC 4.0*
    - TitlePage - *CC BY-NC 4.0*
    - InfoPage - *Other*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - Foreword - *CC BY-NC 4.0*
  - 1: An Introduction to Statistics - *Undeclared*
    - 1.1: What are statistics? - *CC BY-NC-SA 4.0*
    - 1.2: Why do we study statistics? - *CC BY-NC-SA 4.0*
    - 1.3: Statistics in Psychology - *CC BY-SA 4.0*
    - 1.4: Putting It Together- Types of Statistical Studies and Producing Data - *CC BY 4.0*
    - 1.5: Introduction to Psychological Measurement - *CC BY-SA 4.0*
    - 1.6: Types of Data and How to Collect Them - *CC BY-NC-SA 4.0*
    - 1.7: Statistics and Science - *CC BY-NC-SA 4.0*
    - 1.8: How to Succeed in Statistics - *CC BY-NC-SA 4.0*
    - 1.9: Math Refresher - *CC BY-SA 4.0*
  - 2: Describing Data using Distributions and Graphs - *Undeclared*
    - 2.1: Introduction to Looking at Data (This is what too many numbers looks like) - *CC BY-SA 4.0*
    - 2.2: Frequency Tables - *CC BY-SA 4.0*
    - 2.3: Graphing Qualitative Variables- Bar Graphs - *CC BY-SA 4.0*
    - 2.4: Graphing Qualitative Variables- Pie Charts - *CC BY-SA 4.0*
    - 2.5: Graphing Quantitative Variables - *CC BY-NC-SA 4.0*
    - 2.6: Skew and Kurtosis - *CC BY-SA 4.0*
    - 2.7: Graphing Quantitative Data- Line Graphs - *CC BY-SA 4.0*
    - 2.8: Graphing Quantitative Data- Histograms - *CC BY-SA 4.0*
    - 2.9: Graphing Quantitative Data- Boxplots - *CC BY-SA 4.0*
    - 2.10: Graphing Quantitative Data- Scatterplots - *CC BY-SA 4.0*
    - 2.11: Summary and Some Honesty - *CC BY-NC-SA 4.0*
  - 3: Measures of Central Tendency and Spread - *Undeclared*
    - 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: Z-scores and the Standard Normal Distribution - *Undeclared*
  - 4.1: Normal Distributions - *CC BY-NC-SA 4.0*
  - 4.2: Introduction to z-scores - *CC BY-NC-SA 4.0*
  - 4.3: Calculating z-scores - *CC BY-SA 4.0*
  - 4.4: Introduction to the z table - *CC BY-SA 4.0*
  - 4.5: Z table of critical values - *CC BY-SA 4.0*
  - 4.6: Predicting Amounts - *CC BY-SA 4.0*
  - 4.7: Summary of z Scores - *CC BY-SA 4.0*
  - 4.8: The Write-Up - *CC BY-SA 4.0*
- 5: Distributions - *Undeclared*
  - 5.1: Introduction to Distributions - *CC BY-SA 4.0*
  - 5.2: What is Probability - *CC BY-NC-SA 4.0*
  - 5.3: The Law of Large Numbers - *CC BY-SA 4.0*
  - 5.4: Normal Distributions and Probability Distributions - *CC BY-SA 4.0*
  - 5.5: Putting it All Together - *CC BY-SA 4.0*
  - 5.6: Summary- The Bigger Picture - *CC BY-NC-SA 4.0*
- 6: Inferential Statistics and an Introduction to Hypothesis Testing - *Undeclared*
  - 6.1: Growth Mindset - *CC BY-SA 4.0*
  - 6.2: People, Samples, and Populations - *CC BY-SA 4.0*
  - 6.3: The Sampling Distribution of Sample Means - *CC BY-NC-SA 4.0*
  - 6.4: Using Standard Error for Probability - *CC BY-NC-SA 4.0*
  - 6.5: Sampling Distribution, Probability and Inference - *CC BY-NC-SA 4.0*
  - 6.6: Logic and Purpose of Hypothesis Testing - *CC BY-NC-SA 4.0*
  - 6.7: The Probability Value - *CC BY-NC-SA 4.0*
  - 6.8: Descriptive versus Inferential Statistics - *CC BY-SA 4.0*
  - 6.9: The Null Hypothesis - *CC BY-NC-SA 4.0*
  - 6.10: The Alternative Hypothesis - *CC BY-SA 4.0*
  - 6.11: Null Hypothesis Significance Testing - *CC BY-SA 4.0*
  - 6.12: Critical Values, p-values, and Significance - *CC BY-SA 4.0*
    - 6.12.1: Critical Values - *CC BY-SA 4.0*
    - 6.12.2: Summary of p-values and NHST - *CC BY-SA 4.0*
  - 6.13: Steps of the Hypothesis Testing Process - *CC BY-SA 4.0*
  - 6.14: Movie Popcorn - *CC BY-NC-SA 4.0*
  - 6.15: Effect Size - *CC BY-NC-SA 4.0*
  - 6.16: Office Temperature - *CC BY-NC-SA 4.0*
  - 6.17: The Two Errors in Null Hypothesis Significance Testing - *CC BY-SA 4.0*
    - 6.17.1: Power and Sample Size - *CC BY-SA 4.0*
    - 6.17.2: The p-value of a Test - *CC BY-SA 4.0*
- 7: Introducing t-tests (One-sample designs) - *Undeclared*
  - 7.1: Introduction to One-Sample t-tests - *Undeclared*
  - 7.2: Hypothesis Testing with t - *CC BY-NC-SA 4.0*
  - 7.3: Confidence Intervals - *CC BY-NC-SA 4.0*
  - 7.4: t table of critical values - *Undeclared*
  - 7.5: Reporting Results - *Undeclared*
    - 7.5.1: Descriptive and Inferential Calculations and Conclusion Example - *Undeclared*
- 8: Two Sample Designs - *Undeclared*
  - 8.1: Repeated Measures - *CC BY-NC-SA 4.0*
    - 8.1.1: Change and Differences - *CC BY-NC-SA 4.0*
    - 8.1.2: Hypotheses of Change and Differences - *CC BY-NC-SA 4.0*
    - 8.1.3: Increasing Satisfaction at Work - *CC BY-NC-SA 4.0*
    - 8.1.4: Bad Press - *CC BY-NC-SA 4.0*
  - 8.2: Independent Samples - *CC BY-NC-SA 4.0*
    - 8.2.1: Difference of Means - *CC BY-NC-SA 4.0*
    - 8.2.2: Research Questions about Independent Means - *CC BY-NC-SA 4.0*
    - 8.2.3: Hypotheses and Decision Criteria - *CC BY-NC-SA 4.0*
    - 8.2.4: Independent Samples t-statistic - *CC BY-NC-SA 4.0*
    - 8.2.5: Standard Error and Pooled Variance - *CC BY-NC-SA 4.0*
    - 8.2.6: Movies and Mood - *CC BY-NC-SA 4.0*
    - 8.2.7: Effect Sizes and Confidence Intervals - *CC BY-NC-SA 4.0*
    - 8.2.8: Homogeneity of Variance - *CC BY-NC-SA 4.0*
- 9: Analysis of Variance- Independent Samples - *Undeclared*
  - 9.1: Why ANOVA? - *Undeclared*
    - 9.1.1: Observing and Interpreting Variability - *CC BY-NC-SA 4.0*
    - 9.1.2: Ratio of Variability - *CC BY-SA 4.0*
  - 9.2: Introduction to ANOVA's Sum of Squares - *CC BY-SA 4.0*

        - 9.2.1: Summary of ANOVA Summary Table - *CC BY-SA 4.0*
    - 9.3: Hypotheses in ANOVA - *Undeclared*
    - 9.4: Practice with Job Applicants - *CC BY-SA 4.0*
        - 9.4.1: Table of Critical F-Scores - *CC BY-SA 4.0*
    - 9.5: Introduction to Pairwise Comparisons - *CC BY-SA 4.0*
        - 9.5.1: Pairwise Comparison Post Hoc Tests for Critical Values of Mean Differences - *CC BY-SA 4.0*
    - 9.6: Practice on Mindset Data - *CC BY-SA 4.0*
    - 9.7: On the Relationship Between ANOVA and the Student t Test - *CC BY-SA 4.0*
    - 9.8: Non-Parametric Analysis Between Multiple Groups - *CC BY-SA 4.0*
- 10: Analysis of Variance- Repeated Measures - *Undeclared*
    - 10.1: Introduction to Repeated Measures ANOVA - *CC BY-SA 4.0*
        - 10.1.1: Things Worth Knowing About RM ANOVAs - *CC BY-SA 4.0*
    - 10.2: ANOVA Summary Table - *CC BY-SA 4.0*
        - 10.2.1: Repeated Measures ANOVA Sum of Squares Formulas - *CC BY-SA 4.0*
    - 10.3: Practice with RM ANOVA Summary Table - *CC BY-SA 4.0*
        - 10.3.1: Practice with Mindset - *CC BY-SA 4.0*
    - 10.4: Non-Parametric RM ANOVA - *CC BY-SA 4.0*
- 11: Factorial ANOVA (Two-Way) - *Undeclared*
    - 11.1: Introduction to Factorial Designs - *CC BY-SA 4.0*
        - 11.1.1: Factorial Notations and Square Tables - *CC BY-SA 4.0*
    - 11.2: Introduction to Main Effects and Interactions - *CC BY-SA 4.0*
        - 11.2.1: Example with Main Effects and Interactions - *CC BY-SA 4.0*
        - 11.2.2: Graphing Main Effects and Interactions - *CC BY-SA 4.0*
        - 11.2.3: Interpreting Main Effects and Interactions in Graphs - *CC BY-SA 4.0*
        - 11.2.4: Interpreting Interactions- Do Main Effects Matter? - *CC BY-SA 4.0*
        - 11.2.5: Interpreting Beyond 2x2 in Graphs - *CC BY-SA 4.0*
    - 11.3: Two-Way ANOVA Summary Table - *CC BY-SA 4.0*
        - 11.3.1: Calculating Sum of Squares for the Factorial ANOVA Summary Table - *CC BY-SA 4.0*
    - 11.4: When Should You Conduct Post-Hoc Pairwise Comparisons? - *CC BY-SA 4.0*
    - 11.5: Practice with a 2x2 Factorial Design- Attention - *CC BY-SA 4.0*
        - 11.5.1: Practice 2x3 Factorial ANOVA on Mindset - *CC BY-SA 4.0*
    - 11.6: Choosing the Correct Analysis- Mean Comparison Edition - *CC BY-SA 4.0*
- 12: Correlation (and a little about Regression) - *CC BY-SA 4.0*
    - 12.1: If something caused something else to change, what would that look like? - *CC BY-SA 4.0*
    - 12.2: Pearson's r - *CC BY-SA 4.0*
    - 12.3: Turning the numbers into a measure of co-variance - *CC BY-SA 4.0*
    - 12.4: Examples with Data - *CC BY-SA 4.0*
    - 12.5: Regression — A mini intro - *CC BY-SA 4.0*
    - 12.6: Interpreting Correlations - *CC BY-SA 4.0*
    - 12.7: Summary - *CC BY-SA 4.0*
    - 12.8: References - *CC BY-SA 4.0*
- 13: Chi-Square - *Undeclared*
    - 13.1: Introduction to Chi-Square - *CC BY-SA 4.0*
        - 13.1.1: Assumptions of the Test(s) - *CC BY-SA 4.0*
    - 13.2: Introduction to Goodness-of-Fit Chi-Square - *CC BY-SA 4.0*
        - 13.2.1: Critical Values of Chi-Square Table - *CC BY-SA 4.0*
        - 13.2.2: Interpretation of the Chi-Square Goodness-of-Fit Test - *CC BY-SA 4.0*
    - 13.3: Goodness of Fit χ^2 Formula - *CC BY-SA 4.0*
    - 13.4: Practice Goodness of Fit- Pineapple on Pizza - *CC BY-SA 4.0*
    - 13.5: Introduction to Test of Independence - *CC BY-SA 4.0*
    - 13.6: Practice Chi-Square Test of Independence- College Sports - *CC BY-SA 4.0*
        - 13.6.1: Practice- Fast Food Meals - *CC BY-SA 4.0*
    - 13.7: RM Chi-Square- The McNemar Test - *CC BY-SA 4.0*
    - 13.8: Choosing the Correct Test- Chi-Square Edition - *CC BY-SA 4.0*
- 14: Wrapping it Up - *Undeclared*
    - 14.1: Introduction to Wrapping Up - *CC BY-SA 4.0*
    - 14.2: Choosing the Test - *CC BY-SA 4.0*
    - 14.3: Why did you take this class? - *CC BY-SA 4.0*

- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*

Detailed Licensing

Overview

Title: Intro to Stats in the Social Sciences (MPC)

Webpages: 162

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-SA 4.0: 54.9% (89 pages)
- CC BY-NC-SA 4.0: 25.3% (41 pages)
- Undeclared: 16% (26 pages)
- CC BY-NC 4.0: 2.5% (4 pages)
- Other: 0.6% (1 page)
- CC BY 4.0: 0.6% (1 page)

By Page

- Intro to Stats in the Social Sciences (MPC) - *CC BY-NC 4.0*
  - Front Matter - *CC BY-NC 4.0*
    - TitlePage - *CC BY-NC 4.0*
    - InfoPage - *Other*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - Foreword - *CC BY-NC 4.0*
  - 1: An Introduction to Statistics - *Undeclared*
    - 1.1: What are statistics? - *CC BY-NC-SA 4.0*
    - 1.2: Why do we study statistics? - *CC BY-NC-SA 4.0*
    - 1.3: Statistics in Psychology - *CC BY-SA 4.0*
    - 1.4: Putting It Together- Types of Statistical Studies and Producing Data - *CC BY 4.0*
    - 1.5: Introduction to Psychological Measurement - *CC BY-SA 4.0*
    - 1.6: Types of Data and How to Collect Them - *CC BY-NC-SA 4.0*
    - 1.7: Statistics and Science - *CC BY-NC-SA 4.0*
    - 1.8: How to Succeed in Statistics - *CC BY-NC-SA 4.0*
    - 1.9: Math Refresher - *CC BY-SA 4.0*
  - 2: Describing Data using Distributions and Graphs - *Undeclared*
    - 2.1: Introduction to Looking at Data (This is what too many numbers looks like) - *CC BY-SA 4.0*
    - 2.2: Frequency Tables - *CC BY-SA 4.0*
    - 2.3: Graphing Qualitative Variables- Bar Graphs - *CC BY-SA 4.0*
    - 2.4: Graphing Qualitative Variables- Pie Charts - *CC BY-SA 4.0*
    - 2.5: Graphing Quantitative Variables - *CC BY-NC-SA 4.0*
    - 2.6: Skew and Kurtosis - *CC BY-SA 4.0*
    - 2.7: Graphing Quantitative Data- Line Graphs - *CC BY-SA 4.0*
    - 2.8: Graphing Quantitative Data- Histograms - *CC BY-SA 4.0*
    - 2.9: Graphing Quantitative Data- Boxplots - *CC BY-SA 4.0*
    - 2.10: Graphing Quantitative Data- Scatterplots - *CC BY-SA 4.0*
    - 2.11: Summary and Some Honesty - *CC BY-NC-SA 4.0*
  - 3: Measures of Central Tendency and Spread - *Undeclared*
    - 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: Z-scores and the Standard Normal Distribution - *Undeclared*
    - 4.1: Normal Distributions - *CC BY-NC-SA 4.0*
    - 4.2: Introduction to z-scores - *CC BY-NC-SA 4.0*
    - 4.3: Calculating z-scores - *CC BY-SA 4.0*
    - 4.4: Introduction to the z table - *CC BY-SA 4.0*
    - 4.5: Z table of critical values - *CC BY-SA 4.0*
    - 4.6: Predicting Amounts - *CC BY-SA 4.0*
    - 4.7: Summary of z Scores - *CC BY-SA 4.0*
    - 4.8: The Write-Up - *CC BY-SA 4.0*
  - 5: Distributions - *Undeclared*
    - 5.1: Introduction to Distributions - *CC BY-SA 4.0*

- 5.2: What is Probability - *CC BY-NC-SA 4.0*
- 5.3: The Law of Large Numbers - *CC BY-SA 4.0*
- 5.4: Normal Distributions and Probability Distributions - *CC BY-SA 4.0*
- 5.5: Putting it All Together - *CC BY-SA 4.0*
- 5.6: Summary- The Bigger Picture - *CC BY-NC-SA 4.0*

- 6: Inferential Statistics and an Introduction to Hypothesis Testing - *Undeclared*
  - 6.1: Growth Mindset - *CC BY-SA 4.0*
  - 6.2: People, Samples, and Populations - *CC BY-SA 4.0*
  - 6.3: The Sampling Distribution of Sample Means - *CC BY-NC-SA 4.0*
  - 6.4: Using Standard Error for Probability - *CC BY-NC-SA 4.0*
  - 6.5: Sampling Distribution, Probability and Inference - *CC BY-NC-SA 4.0*
  - 6.6: Logic and Purpose of Hypothesis Testing - *CC BY-NC-SA 4.0*
  - 6.7: The Probability Value - *CC BY-NC-SA 4.0*
  - 6.8: Descriptive versus Inferential Statistics - *CC BY-SA 4.0*
  - 6.9: The Null Hypothesis - *CC BY-NC-SA 4.0*
  - 6.10: The Alternative Hypothesis - *CC BY-SA 4.0*
  - 6.11: Null Hypothesis Significance Testing - *CC BY-SA 4.0*
  - 6.12: Critical Values, p-values, and Significance - *CC BY-SA 4.0*
    - 6.12.1: Critical Values - *CC BY-SA 4.0*
    - 6.12.2: Summary of p-values and NHST - *CC BY-SA 4.0*
  - 6.13: Steps of the Hypothesis Testing Process - *CC BY-SA 4.0*
  - 6.14: Movie Popcorn - *CC BY-NC-SA 4.0*
  - 6.15: Effect Size - *CC BY-NC-SA 4.0*
  - 6.16: Office Temperature - *CC BY-NC-SA 4.0*
  - 6.17: The Two Errors in Null Hypothesis Significance Testing - *CC BY-SA 4.0*
    - 6.17.1: Power and Sample Size - *CC BY-SA 4.0*
    - 6.17.2: The p-value of a Test - *CC BY-SA 4.0*
- 7: Introducing t-tests (One-sample designs) - *Undeclared*
  - 7.1: Introduction to One-Sample t-tests - *Undeclared*
  - 7.2: Hypothesis Testing with t - *CC BY-NC-SA 4.0*
  - 7.3: Confidence Intervals - *CC BY-NC-SA 4.0*
  - 7.4: t table of critical values - *Undeclared*
  - 7.5: Reporting Results - *Undeclared*
    - 7.5.1: Descriptive and Inferential Calculations and Conclusion Example - *Undeclared*
- 8: Two Sample Designs - *Undeclared*
  - 8.1: Repeated Measures - *CC BY-NC-SA 4.0*
    - 8.1.1: Change and Differences - *CC BY-NC-SA 4.0*
    - 8.1.2: Hypotheses of Change and Differences - *CC BY-NC-SA 4.0*
    - 8.1.3: Increasing Satisfaction at Work - *CC BY-NC-SA 4.0*
    - 8.1.4: Bad Press - *CC BY-NC-SA 4.0*
  - 8.2: Independent Samples - *CC BY-NC-SA 4.0*
    - 8.2.1: Difference of Means - *CC BY-NC-SA 4.0*
    - 8.2.2: Research Questions about Independent Means - *CC BY-NC-SA 4.0*
    - 8.2.3: Hypotheses and Decision Criteria - *CC BY-NC-SA 4.0*
    - 8.2.4: Independent Samples t-statistic - *CC BY-NC-SA 4.0*
    - 8.2.5: Standard Error and Pooled Variance - *CC BY-NC-SA 4.0*
    - 8.2.6: Movies and Mood - *CC BY-NC-SA 4.0*
    - 8.2.7: Effect Sizes and Confidence Intervals - *CC BY-NC-SA 4.0*
    - 8.2.8: Homogeneity of Variance - *CC BY-NC-SA 4.0*
- 9: Analysis of Variance- Independent Samples - *Undeclared*
  - 9.1: Why ANOVA? - *Undeclared*
    - 9.1.1: Observing and Interpreting Variability - *CC BY-NC-SA 4.0*
    - 9.1.2: Ratio of Variability - *CC BY-SA 4.0*
  - 9.2: Introduction to ANOVA's Sum of Squares - *CC BY-SA 4.0*
    - 9.2.1: Summary of ANOVA Summary Table - *CC BY-SA 4.0*
  - 9.3: Hypotheses in ANOVA - *Undeclared*
  - 9.4: Practice with Job Applicants - *CC BY-SA 4.0*
    - 9.4.1: Table of Critical F-Scores - *CC BY-SA 4.0*
  - 9.5: Introduction to Pairwise Comparisons - *CC BY-SA 4.0*
    - 9.5.1: Pairwise Comparison Post Hoc Tests for Critical Values of Mean Differences - *CC BY-SA 4.0*
  - 9.6: Practice on Mindset Data - *CC BY-SA 4.0*
  - 9.7: On the Relationship Between ANOVA and the Student t Test - *CC BY-SA 4.0*

- 9.8: Non-Parametric Analysis Between Multiple Groups - *CC BY-SA 4.0*

- 10: Analysis of Variance- Repeated Measures - *Undeclared*
  - 10.1: Introduction to Repeated Measures ANOVA - *CC BY-SA 4.0*
    - 10.1.1: Things Worth Knowing About RM ANOVAs - *CC BY-SA 4.0*
  - 10.2: ANOVA Summary Table - *CC BY-SA 4.0*
    - 10.2.1: Repeated Measures ANOVA Sum of Squares Formulas - *CC BY-SA 4.0*
  - 10.3: Practice with RM ANOVA Summary Table - *CC BY-SA 4.0*
    - 10.3.1: Practice with Mindset - *CC BY-SA 4.0*
  - 10.4: Non-Parametric RM ANOVA - *CC BY-SA 4.0*
- 11: Factorial ANOVA (Two-Way) - *Undeclared*
  - 11.1: Introduction to Factorial Designs - *CC BY-SA 4.0*
    - 11.1.1: Factorial Notations and Square Tables - *CC BY-SA 4.0*
  - 11.2: Introduction to Main Effects and Interactions - *CC BY-SA 4.0*
    - 11.2.1: Example with Main Effects and Interactions - *CC BY-SA 4.0*
    - 11.2.2: Graphing Main Effects and Interactions - *CC BY-SA 4.0*
    - 11.2.3: Interpreting Main Effects and Interactions in Graphs - *CC BY-SA 4.0*
    - 11.2.4: Interpreting Interactions- Do Main Effects Matter? - *CC BY-SA 4.0*
    - 11.2.5: Interpreting Beyond 2x2 in Graphs - *CC BY-SA 4.0*
  - 11.3: Two-Way ANOVA Summary Table - *CC BY-SA 4.0*
    - 11.3.1: Calculating Sum of Squares for the Factorial ANOVA Summary Table - *CC BY-SA 4.0*
  - 11.4: When Should You Conduct Post-Hoc Pairwise Comparisons? - *CC BY-SA 4.0*
  - 11.5: Practice with a 2x2 Factorial Design- Attention - *CC BY-SA 4.0*
    - 11.5.1: Practice 2x3 Factorial ANOVA on Mindset - *CC BY-SA 4.0*
  - 11.6: Choosing the Correct Analysis- Mean Comparison Edition - *CC BY-SA 4.0*
- 12: Correlation (and a little about Regression) - *CC BY-SA 4.0*
  - 12.1: If something caused something else to change, what would that look like? - *CC BY-SA 4.0*
  - 12.2: Pearson's r - *CC BY-SA 4.0*
  - 12.3: Turning the numbers into a measure of co-variance - *CC BY-SA 4.0*
  - 12.4: Examples with Data - *CC BY-SA 4.0*
  - 12.5: Regression — A mini intro - *CC BY-SA 4.0*
  - 12.6: Interpreting Correlations - *CC BY-SA 4.0*
  - 12.7: Summary - *CC BY-SA 4.0*
  - 12.8: References - *CC BY-SA 4.0*
- 13: Chi-Square - *Undeclared*
  - 13.1: Introduction to Chi-Square - *CC BY-SA 4.0*
    - 13.1.1: Assumptions of the Test(s) - *CC BY-SA 4.0*
  - 13.2: Introduction to Goodness-of-Fit Chi-Square - *CC BY-SA 4.0*
    - 13.2.1: Critical Values of Chi-Square Table - *CC BY-SA 4.0*
    - 13.2.2: Interpretation of the Chi-Square Goodness-of-Fit Test - *CC BY-SA 4.0*
  - 13.3: Goodness of Fit χ^2 Formula - *CC BY-SA 4.0*
  - 13.4: Practice Goodness of Fit- Pineapple on Pizza - *CC BY-SA 4.0*
  - 13.5: Introduction to Test of Independence - *CC BY-SA 4.0*
  - 13.6: Practice Chi-Square Test of Independence- College Sports - *CC BY-SA 4.0*
    - 13.6.1: Practice- Fast Food Meals - *CC BY-SA 4.0*
  - 13.7: RM Chi-Square- The McNemar Test - *CC BY-SA 4.0*
  - 13.8: Choosing the Correct Test- Chi-Square Edition - *CC BY-SA 4.0*
- 14: Wrapping it Up - *Undeclared*
  - 14.1: Introduction to Wrapping Up - *CC BY-SA 4.0*
  - 14.2: Choosing the Test - *CC BY-SA 4.0*
  - 14.3: Why did you take this class? - *CC BY-SA 4.0*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*