BEHAVIORAL STATISTICS



Michelle Oja
Taft College



Taft College

PSYC 2200: Elementary Statistics for Behavioral and Social Sciences

Michelle Oja

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.

The LibreTexts name, logo, and book covers are trademarked by LibreTexts, Inc. (a 501(c)(3) not-for-profit organization) and protected against unauthorized use.

This text was compiled on 08/24/2026

TABLE OF CONTENTS

Acknowledgements

Licensing

Licensing

For Faculty

1.1: Changes in Second Edition

1: Introduction to Behavioral Statistics

1.1: Why are you taking this course?

1.2: Making Meaning of Numbers

1.3: How Do You Know?

1.3.1: The Scientific Method

1.4: Types of Data

1.4.1: IV and DV

1.4.2: Types of Variables

1.4.3: Scales of Measurement

1.5: Populations and Samples

1.6: Collecting Data- More Practice with Populations and Samples

1.7: "Research shows that..."

1.8: Learning (Statistics)

2: What Do Data Look Like? (Graphs)

2.1: Introduction to Looking at Data

2.2: Frequency Tables

2.3: APA Style Tables

2.4: Graphing Qualitative Variables- Bar Graphs

2.5: Graphing Qualitative Variables- Pie Charts

2.6: Graphing Quantitative Variables

2.7: Skew and Kurtosis

2.8: Graphing Quantitative Data- Line Graphs

2.9: Graphing Quantitative Data- Histograms

2.10: Graphing Quantitative Data- Means Plot

2.11: Graphing Quantitative Data- Boxplots

2.12: Graphing Quantitative Data- Scatterplots

2.13: Summary and Some Honesty

2.14: APA Style Charts

3: Measures of Central Tendency

3.1: Introduction to Measures of Central Tendency

3.2: Math Refresher

3.2.1: Decimals
3.3: Mode
3.4: Median
3.5: Mean
3.6: Interpreting All Three
3.7: Summary of Measures of Central Tendency

4: Measures of Variability

4.1: Introduction to Measures of Variability
4.2: Introduction to Standard Deviation Calculations
4.3: Practice SD on Wait Times
4.4: Interpreting Standard Deviations
4.5: Fractions, Decimals, and Percents, Oh My!
4.6: Putting All of the Measures Together

5: Distributions

5.1: Introduction to Distributions
5.2: Introduction to Probability
5.3: The Binomial Distribution
5.4: The Law of Large Numbers
5.5: Three Important Distributions Together
5.6: Sampling Distributions and the Central Limit Theorem
5.7: Putting it All Together

6: Using z

6.1: Introduction to z-scores
6.2: Calculating z-scores
6.3: Practice z-score Calculation on IQ Scores
6.4: Introduction to the z table
6.5: Practice Using the z Table
6.6: Predicting Amounts
6.7: Summary of z Scores

7: APA Style

7.1: APA and APA Style
7.2: APA Style Resources
7.3: General Paper Format
7.4: Formatting by Section
7.5: Tables and Figures

8: Inferential Statistics and Hypothesis Testing

8.1: Growth Mindset
8.2: Samples and Populations Refresher
8.3: Can Samples Predict Populations?

8.4: Descriptive versus Inferential Statistics
8.5: The Research Hypothesis and the Null Hypothesis
8.6: Null Hypothesis Significance Testing
8.7: Critical Values
8.7.1: Rejection Region Visualization
8.8: Practice with p-values and NHST
8.9: Steps of the Hypothesis Testing Process
8.10: The Two Errors in Null Hypothesis Significance Testing
8.11: Power and Sample Size
8.12: Views of NHST

9: One Sample t-test
9.1: Introduction to One-Sample t-tests
9.2: One-Sample t-test Calculations
9.3: Reporting Results
9.4: Practice One-Sample t-tests on Mindset Scores
9.5: Confidence Intervals

10: Independent Samples t-test
10.1: Introduction to Independent Sample t-tests
10.2: Independent Samples t-test Equation
10.3: Hypotheses with Two Samples
10.4: Practice Independent t-test on Movies and Mood
10.5: Practice Independent t-test on Growth Mindset
10.6: When to NOT use the Independent Samples t-test
10.7: Non-Parametric Independent Sample t-Test

11: Dependent Samples t-test
11.1: Introduction to Dependent Samples
11.2: Dependent Sample t-test Calculations
11.3: Practice Dependent t-test on Job Satisfaction
11.4: Practice Dependent t-test on Changes in Mindset
11.5: Non-Parametric Analysis of Dependent Samples
11.6: Choosing Which Statistic- t-test Edition

12: BG ANOVA
12.1: Why ANOVA?
12.1.1: Observing and Interpreting Variability
12.1.2: Ratio of Variability
12.2: Introduction to ANOVA's Sum of Squares
12.3: Summary of ANOVA Summary Table
12.4: Hypotheses in ANOVA
12.5: Practice BG ANOVA on Job Applicants

12.6: Introduction to Pairwise Comparisons
12.7: Pairwise Comparison Post Hoc Tests
12.8: Practice BG ANOVA and Pairwise Comparisons on Mindset Data
12.9: ANOVA versus t-test
12.10: Non-Parametric Analysis Between Multiple Groups

13: RM ANOVA

13.1: Introduction to Repeated Measures ANOVA
13.2: RM ANOVA Summary Table
13.3: Repeated Measures ANOVA Sum of Squares Formulas
13.4: Practice RM ANOVA Summary Table on Typing Speed
13.5: Practice RM ANOVA on Mindset Scores
13.6: Non-Parametric RM ANOVA

14: Factorial ANOVA (Two-Way)

14.1: Introduction to Factorial Designs
14.2: Factorial Notations and Square Tables
14.3: Introduction to Main Effects and Interactions
14.4: Example with Main Effects and Interactions
14.5: Graphing Main Effects and Interactions
14.6: Interpreting Main Effects and Interactions in Graphs
14.7: Do Main Effects Matter?
14.8: Beyond 2x2 in Graphs
14.9: Two-Way ANOVA Summary Table
14.9.1: Calculating Sum of Squares for the Factorial ANOVA Summary Table
14.10: When Should You Conduct Post-Hoc Pairwise Comparisons?
14.11: Practice Factorial on Attention
14.12: Practice Factorial on Mindset Scores
14.13: Choosing the Correct Analysis- Mean Comparison Edition

15: Correlations

15.1: Refresh to Prepare
15.2: What do Two Quantitative Variables Look Like?
15.3: Introduction to Pearson's r
15.4: Correlation versus Causation
15.4.1: Correlation versus Causation in Graphs
15.5: Strength, Direction, and Linearity
15.6: Hypotheses
15.7: Correlation Formula- Covariance Divided by Variability
15.8: Practice Correlation on Anxiety and Depression
15.9: Practice Correlation on Nutrition
15.10: Alternatives to Pearson's Correlation

15.11: Final Considerations

16: Regression

16.1: Introduction- Line of Best Fit
16.2: Regression Line Equation
16.3: Using Linear Equations
16.4: Hypothesis Testing- Slope to ANOVAs
16.5: Practice Regression on Health and Happiness
16.6: Practice Regression on Nutrition
16.7: Multiple Regression

17: Chi-Square

17.1: Introduction to Chi-Square
17.2: Assumptions of the Test(s)
17.3: Introduction to Goodness-of-Fit
17.4: Interpretation of the Goodness-of-Fit Test
17.5: Goodness of Fit Formula
17.6: Practice Goodness of Fit on Pineapple on Pizza
17.7: Introduction to Test of Independence
17.8: Practice Test of Independence on College Sports
17.9: Practice Chi-Square on Fast Food Meals
17.10: Phi Correlation is Not a Correlation
17.11: Choosing the Correct Test- Chi-Square Edition

18: Wrap Up

18.1: Choosing the Test
18.2: Why did you take this class?

Common Formulas

Common Critical Value Tables

1.1: Table of Critical Values of z
1.2: Table of Critical Values of t
1.3: Table of Critical Values of F
1.4: Table of Critical Values of r
1.5: Table of Critical Values of Chi-Square

Index

Glossary

Detailed Licensing

Detailed Licensing

This page titled Behavioral Statistics 2e is shared under a license and was authored, remixed, and/or curated by Michelle Oja.

Acknowledgements

The first edition for this book was compiled for a sabbatical project during the COVID-19 pandemic. Through the support of the California Community Colleges Chancellor's Office's Burden-Free Instructional Materials program, my college was able to support this second edition (and other projects!) through Zero Textbook Cost grants (CCCCO ZTC). I am thankful for my college's support and the CCCCO ZTC grants. And none of this would be possible without the dedication of the Academic Senate for California College's Open Educational Resources Initiative (ASCCC's OERI).

I am thankful to Dr. Eveland for suggesting that we write a behavioral statistics together years ago (to which I responding with a resounding "Nope!"). I am also indebted to Tabitha Raber and David Wymore for being the first faculty at Taft College to delve into OER. They gave me advice and support.

The creator and staff from LibreTexts were also instrumental in helping me understand how to develop a textbook in their system, and fix anything that I "broke"! I started the first edition over 5 years ago; a lot has changed at LibreTexts, but their helpfulness has continued. I lost my Mom between the first and second editions, but she was so proud of that first edition because of how profession LibreTexts makes the bound versions.

Through this support, I was able to learn more about online educational resources (OER) and statistics (and coding of equations!). I read several openly-licensed statistics textbooks that became the foundation for this compilation. In particular, Foster et al.'s (2021) An Introduction to Psychological Statistics was the starting point for many of the chapters. I also loved Navarro's (2020) Learning Statistics with R- A Tutorial for Psychology Students and Other Beginners; she wrote a wonderfully enlightening and enjoyable statistics textbook. It would have been perfect if we were looking for a textbook to teach statistical software. Finally, Crump's Answering Questions with Data (also found at https://www.crumplab.com/statistics/) provided helpful pages, and supported most of the chapter on factorial design ANOVAs.

Finally, none of this could have happened without my partner supporting me and taking care of the house; thank you, Gary!

Citations

Crump, M. J. C. (2021). *Answering Questions with Data - Introductory Statistics for Psychology Students*. Retrieved July 3, 2021, from https://stats.libretexts.org/@go/page/7883

Foster, G. C., Lane, D., Scott, D., Hebl, M., Guerra, R., Osherson, D., & Zimmer, H. (2021). *An Introduction to Psychological*. Retrieved July 3, 2021, from https://stats.libretexts.org/@go/page/7077

Navarro, D. (2020). *Learning Statistics with R - A tutorial for Psychology Students and other Beginners*. Retrieved July 3, 2021, from https://stats.libretexts.org/@go/page/3936

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

Licensing

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

This page titled Licensing is shared under a license and was authored, remixed, and/or curated by LibreTexts.

Licensing

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

This page titled Licensing is shared under a license and was authored, remixed, and/or curated by LibreTexts.

For Faculty

Licensing

The licensing for this textbook changes based on the webpage (sub-section). As most of the textbook was modified from Foster et al. (2021), much of the textbook is CC-BY-NC-SA (a Creative Commons license that allows anyone to use the information for non-commercial purposes and the user must license the modified work in the same way). Most of the other authors' work is licensed as CC-BY-SA (a Creative Commons license allowing use that requires that any modifications of the current work is licensed in the same way).

Homework

Homework for the first edition of this textbook housed in Canvas Commons (search PSYC 2200). Dr. MO created worksheets that can be answered via Canvas Quizzes, while other faculty have since created other homework ancillaries for the first edition of this textbook. In the worksheets designed by Dr. MO, more questions were provided than might be reasonable for a student to complete for one assignment in order for faculty to have options to "mix and match" questions to create their own worksheets and Canvas quizzes. This variety will provide a better fit for you and your class, as well as reduce the likelihood of your exact homework leaking into the hands of the internet. If this worksheet quiz format or the data sets don't work for you, there are full courses for behavioral statistics on Canvas Commons that provide quizzes or labs (for this textbook, and in general). You can also look into MyOpenMath, or the Query or ADAPT functions in LibreTexts.

If you want to use any of these options with your own data, or want to create new assignments, below is a list of some sites with free datasets. These are not really in any order, but I did use the data from OpenIntro.org.

- Openly-licensed data sets: https://www.openintro.org/data/
- Open Stats Lab from Psychological Science: https://sites.google.com/view/openstatslab
- Datasets on Google: https://datasetsearch.research.google.com/
  - Blog explaining the process: https://www.blog.google/products/search/making-it-easier-discover-datasets/
- Lab manual for Crump Answering Questions with Data textbook: (https://crumplab.github.io/statistics/index.html): https://crumplab.github.io/statisticsLab/
  - You can just grab the data, or use the whole lab manual.
- Data set for teaching and learning at North Carolina State library: https://www.lib.ncsu.edu/teaching-and-learning-datasets
- Forbes' Big Data list: https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#76e598b0b54d
- Cengage data sets: http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/tvds/frames/frame.html
- http://users.stat.ufl.edu/~winner/datasets.html
- Health data at University of Florida : https://people.clas.ufl.edu/winner/miscellaneous-datasets/
- National Opinion Research Center's General Social Survey: http://sda.berkeley.edu/sdaweb/analysis/?dataset=gss14 or www.tinyurl.com/generalsocialsurvey
- PsyToolkit is a free-to-use toolkit for demonstrating, programming, and running cognitive-psychological experiments and surveys, but also might have data from others' research: https://www.psytoolkit.org/
  - Race IAT: https://osf.io/g7cw8/
- APA maintains an Online Psychology Laboratory, in which you can have your students engage in experiments, or have them use available data for anaylsis: https://opl.apa.org/

Book Features

Units

The first edition of this textbook was originally organized into units with some Examples and Exercises based on a theme for each

unit. The potential units and themes are as follows:

1. Unit 1: Description theme is student success
2. Unit 2: Mean Differences theme is growth mindset
3. Unit 3: Relationships theme is nutrition
4. Unit 4: Wrap-Up theme is future careers

Future editions will break Unit 2 into two units, one based on t-tests and one based on ANOVAs. If you have data that would be good to work with for ANOVAs, please send it to Dr. MO! In particular, she's interested in data that BIPOC students would find interesting and has a strengths focus (not a deficit lens).

Notation

Many of the formulas and notation in this textbook are intentionally mathematically incorrect. Although you and I may know what they mean, adding a bunch of subscripts and extra symbols can confuse students. In this introductory textbook, the notations and formulas were chosen to be easy to understand for students new to mathematical equations, rather than to be 100% mathematically correct.

Articulation

This book was compiled to fulfill the requirements of the Course Identification descriptor (C-ID) for Math 110 for the California Community Colleges and Taft College's Course Outline of Record for PSYC 2200 (Elementary Statistics for the Behavioral and Social Sciences).

Additional Textbooks

This textbook is a compilation of Dr. MO's own work with a lot of support from several CC-BY-SA or CC-BY-NC-SA textbooks on LibreTexts website, including:

- Foster, G. C., Lane, D., Scott, D., Hebl, M., Guerra, R., Osherson, D., & Zimmer, H. (2021). *An Introduction to Psychological*. Retrieved July 3, 2021, from https://stats.libretexts.org/@go/page/7077
- Navarro, D. (2020). *Learning Statistics with R - A tutorial for Psychology Students and other Beginners*. Retrieved July 3, 2021, from https://stats.libretexts.org/@go/page/3936
- Crump, M. J. C. (2021). *Answering Questions with Data - Introductory Statistics for Psychology Students*. Retrieved July 3, 2021, from https://stats.libretexts.org/@go/page/7883

Additional textbooks that might be useful are:

- Poritz, J. A (2021). *Lies, Damned Lies, or Statistics - How to Tell the Truth with Statistics (Poritz)*. Retrieved July 3, 2021, from https://stats.libretexts.org/@go/page/7784
- Illowsky, B., & Dean, S. (2019). Book: Introductory Statistics (OpenStax) With Multimedia and Interactivity. Retrieved July 3, 2021, from https://stats.libretexts.org/@go/page/6885

Finally, this support course to help statistics students with basic math is also listed in Learning (Statistic) section of this textbook: Green, L. (2021). *Support Course for Elementary Statistics*. Retrieved July 3, 2021, from https://stats.libretexts.org/@go/page/4710

Enjoy!

This page titled For Faculty is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

1.1: Changes in Second Edition

As Dr. MO used the first edition of this textbook for a few years, so became aware of important changes that needed to be made. This second edition. The following is an annotated list of the bigger changes:

- The first edition's third chapter covered several descriptive analyses. In this second edition, chapter 3 covers the three measures of central tendency, and a new chapter 4 covers measures of variability (primarily standard deviations).
- The first edition organized the textbook into four units. This second edition removed those unit markers. However, the themes for each unit were still utilized. See the For Faculty page for more information on the original units and the themes.
- The model of the scientific method used in the first edition was replaced with a model that better aligns with a behavioral statistics course. The stages of this updated model were then simplified to be used as steps of the Null Hypothesis Significance Testing process.
- The dependent t-test formula was updated to make it easier for students to follow the equation. This change lost some of the intuitive understanding of what the formula is doing, however.
- There was some redundancy and extraneous sections in the first edition, which were deleted.
- All critical value tables were moved to the Back Matter, rather than in the section in which they were first mentioned. Also, the critical value tables for z and t were simplified.

Additionally, there were minor changes related to coding or typographical errors that were fixed. Feel free to use the "Leave feedback" function on any pages to note other errors.

There are tons of more changes that Dr. MO would like to make, but she ran out of time for this edition. :(In particular, future editions will break Unit 2 into two units, one based on t-tests and one based on ANOVAs. If you have data that would be good to work with for ANOVAs, please send it to Dr. MO! In particular, she's interested in data that BIPOC students would find interesting and has a strengths focus (not a deficit lens). If you have suggestions, you can also use the "Leave feedback" button on the bottom of any change.

This page titled 1.1: Changes in Second Edition is shared under a license and was authored, remixed, and/or curated by LibreTexts.

1: INTRODUCTION TO BEHAVIORAL STATISTICS

1.1: Why are you taking this course?

1.2: Making Meaning of Numbers

1.3: How Do You Know?

1.3.1: The Scientific Method

1.4: Types of Data

1.4.1: IV and DV

1.4.2: Types of Variables

1.4.3: Scales of Measurement

1.5: Populations and Samples

1.6: Collecting Data- More Practice with Populations and Samples

1.7: "Research shows that..."

1.8: Learning (Statistics)

This page titled 1: Introduction to Behavioral Statistics is shared under a license and was authored, remixed, and/or curated by Michelle Oja.

1.1: Why are you taking this course?

Why must you take a stats class when you are not a math major? There are probably dozens of answers to why this class is required, but we'll focus on a couple reasons related to your future career, and a few reasons related to being a critically-thinking adult. The one that is probably most compelling to you right now is:

> *Understanding statistics will help make you better in your future career.*

Whether you become a therapist, a social worker, a police officer, a nurse, or run a business, you should read research from your field, and understand the statistical results. You are being trained to think like a manager or supervisor. To make decisions in your career, you'll need to understand and apply the results of statistical analyses of others in your field. Understanding statistical analyses will help you make decisions based on evidence. What you will learn in this class can help you show that your program (or your decision) is statistically significantly better than the alternatives.

Every career that Dr. MO can think of has people doing research to see how to be better. Whether your goal is to work as a therapist or work in social services, criminal justice, education, health care, business, or pretty much any field, really, there's a whole wealth of research on how to help clients' mental health and life situations, prevent crime and improve jury decision-making, improve learning, improve health, or increase sales. As a profession, you should read the research from your field. That research will often include results of statistical analyses. Rather than having to trust the researchers themselves, or (gasp) a summary generated by AI (artificial intelligence), you will be able to understand the statistical results *yourself* by the end of this textbook. Understanding statistical analyses will help you make decisions based on evidence. What you will learn in this class can help you show that your program (or your decision) is statistically significantly better than the alternatives.

What you might not know right now, but will know as you progress in your career, is that much of what you will be doing is documenting what you did, and how it worked. That leads us to a second major reason for taking this course:

> *You will need to report evidence to show that what you are doing helps your clients/students/patients/customers.*

It turns out that organizations only want to continue funding activities that are effective. It will be much easier to persuade your supervisor you need additional resources or get support from your community because you can show that what you're doing is working. This evidence comes from *analyzing*, **interpreting**, and reporting data from your program.

Historical Context

Florence Nightingale (1820–1910) is considered the mother of modern nursing. As a British nurse, she also urged governments to keep good records and be guided by statistical evidence (Wikipedia). She was influenced by Belgian astronomer Adolphe Quetelet (1796–1874) who named what we call the social sciences as *social physics*, by analogy with *celestial physics* (Encyclopedia.com).

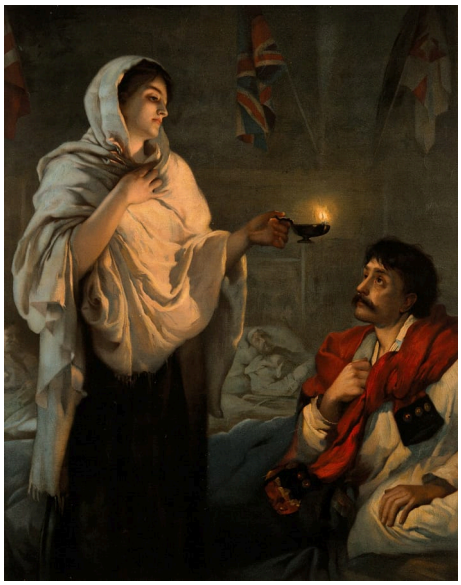


Figure 1.1.1:*The Lady with the Lamp or Miss Nightingale at Scutari*. Popular lithograph reproduction of a painting of Nightingale by Henrietta Rae from Wellcome Collection gallery. CC-BY-4.0, Public Domain

But why do colleges require any math class? After teaching behavioral statistics for almost 15 years, Dr. MO has noticed some outcomes related to how students think at the end of the course. First:

> *Learning statistical analyses develops critical thinking.*

As you learn how to *analyze*, **interpret**, and report data, you'll practice following a process and using available information to make logical decisions. You'll need to decide what analyses you can do with the data that is collected or provided, decide how to apply the equations to your data, decide what the analyses results are telling you, decide the most appropriate way to report your results, and sometimes you'll be asked to make recommendations based on the results. That's a lot of thinking!

While this might sound overwhelming, there are steps that you'll follow. Practicing following a process is another way that you'll be developing your critical thinking. Some of you are naturally more detail-oriented or organized than others, but you *all* can become better at following a process. By doing the work in this textbook and your class, you'll be developing this skill! Being a critical thinker will help you avoid scams and being taken advantage of, and, of course, you'll be better at any job or career if you are good at following directions, staying organized, and identifying any breakdown of logic in processes or when someone's trying to convince you that their way is the right way.

Another way that this class will change how you think is through something that Dr. MO calls "math literacy". Math literacy is like a combination of statistical literacy (the ability to understand and reason with statistics and data) and numeracy (the ability to understand, reason with, and apply simple numerical concepts).

> *Understanding how numbers work will help you make predictions and spot inaccurate information.*

As you practice with some of the more complex equations, you'll want to know if you're doing things correctly before you get to the end. You'll want to make some guesses on what the results should be, then compare your guesses to what your calculator is telling you. You can tell if you've gone off track earlier in the process if you have a sense of how numbers work and how mathematical operations work. For example, sometimes Dr. MO wants to put her class of 35 into groups of three. To know how many groups she will need, she could divide 35 by 3. But before she does that on a calculator, she knows that 36 divided by 3 is 12, so the calculator should show her a number close to 12. The actual number is 11.67 (rounded to two decimal points), which is pretty close to 12!

For more on why your major may require a statistics class, watch this short TED Talk by Arthur Benjamin:

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online



Figure 1.1.1: Arthur Benjamin's TED Talk on statistics before calculus. (CC-BY-NC-ND TED Talks via YouTube or TED Talks directly)

A bonus benefit of learning how to *analyze*, **interpret**, and report data is that some of you will fall in love with statistics, and become a researcher for a living! This isn't a reason why you have to take the course, but it is a happy accident for those of you who want to help people, but tend to be more logical rather than emotional (like Dr. MO!). According to this American Psychological Association (APA) article, there are jobs out there for folks who like statistics.

Note

Compare and Contrast: How are Dr. MO's reasons for taking this course similar to Dr. Navarro's reasons in her Statistics in Psychology page? How do the reasons differ?

What about AI?

You might be wondering how much of this artificial intelligence (AI), like ChatGPT or Gemini or Claude, can do. From Dr. MO's limited experience:

- *Understanding statistics will help make you better in your future career.*
  - AI can summarize reports, but only you can apply those findings to your clients. While AI might act like it can answer questions about your specific clients, only you should be deciding how the research results relate to specific individuals or situations.
- *You will need to report evidence to show that what you are doing helps your clients/students/patients/customers.*
  - AI can provide some data summary and reporting, and will likely be able to do most in the near future. However, some AI results are wildly incorrect. But how would you know that if you didn't do the analyses yourself, too?
- *Learning statistical analyses develops critical thinking.*
  - If you rely on AI to do your thinking, then you will not develop better thinking skills. It's just that simple.
- *Understanding how numbers work will help you make predictions and spot inaccurate information.*
  - Again, if you have AI do the work then you are not developing the skills to understand how numbers work and recognize when something is off.

But sometimes the question isn't "Can AI do this?", but "What are the expectations. For each of the reasons why this course is required, consider:

1. Would your future university want AI to do this goal, rather than you? Why or why not?
2. Would your future employers want AI to do this goal, rather than you? Why or why not?
3. Would your future clients want AI to do this goal, rather than you? Why or why not?

While you might think that Dr. MO's answer to all of these may be "no," that's just not the case. Your university professors may want you to understand how AI works so that you can use it effectively in the future. Your future employer may not care if you can think critically, as long as the project is completed (and is accurate). Your future clients probably want you to help them individually, but you could use AI in the background to help with other tasks.

Now that you've thought about why this course might be required, let's understand a little more about what this course will be about!

This page titled 1.1: Why are you taking this course? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 1.1: On the Psychology of Statistics by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.

1.2: Making Meaning of Numbers

The Crash Course video series are great, and there's one for statistics! Check out this Crach Course Statistics Preview video so that you have an idea of what this class will be about:

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Figure 1.2.1: Crash Course Statistics Preview. (Crash Course: Statistics via YouTube)

You should follow the whole Crash Course Statistics playlist as you go through this textbook's chapters; the chapters won't completely align with the videos, but they will be similar. The Khan Academy also has a great series on statistics.

Statistics are the results of statistical analyses, such as the percentage of psychology majors at your school; the analysis was calculating the percentage, while the result is a statistic that you could share.

Definition: Statistics

The results of statistical analyses

Definition: Statistical Analyses

Procedures to organize and interpret numerical information

To learn more about what statistics are, you can watch Alan Smith's TED Talk about how so many people are bad at estimating statistics.

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Figure 1.2.1: Alan Smith's TED Talk on loving statistics. (CC-BY-NC-ND TED Talks via YouTube or directly from TED Talks)

I have my students watch this video because it shows them that statistics aren't something to be afraid of, and that statistics are really about each of us.

Learning statistical analysis is actually not that similar to learning mathematical analyses. It's more like learning to read and interpret another language. Social scientists know that a language is one of the components of a culture; the culture of statistics is the scientific method and probability. We'll talk about the scientific method next, and will touch on probably when talking about distributions.

This page titled 1.2: Making Meaning of Numbers is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

1.3: How Do You Know?

efore we talk about the scientific method, though, let's look at the bigger picture about what is considered knowledge. ilosophers call these questions (and broader ideas) about the ways of knowing information *epistemology* (link to Wikipedia). ake a minute to ponder some of what you know and how you acquired that knowledge. For example, maybe you have a feeling at summarizing your professor's lecture is more helpful than writing down everything that they say but you don't really have any perience comparing these two ways of taking notes and haven't read about others' research on this topic. Or, perhaps you ow that reading whole chapter of your textbook in detail is important because a parent told you this is what you should do. Or, aybe you've reasoned that face-to-face classes provide information during the class sessions, and learning information will lp you pass the exam, so attendance in the class will help you pass the exam. You could also ask your fiends what study rategies work them to decide which ones may work for you. Or, your professor could randomly assign some students to get inted lecture notes and some students to get online lecture notes, then compare the paper grades to see which type of notes ork best. But should we trust knowledge from these sources?

Methods of Knowing

The paragraph above gave examples of methods of acquiring knowledge. Below are brief descriptions with each method's strengths and weaknesses. Your research methods course will provide you with an in-depth examination of how psychologists use the scientific method to advance our understanding of human behavior and the mind.

Intuition

The first method of knowing that we'll discuss is intuition. When we use our intuition, we are relying on our guts, our emotions, and/or our instincts to guide us. Rather than examining facts or using rational thought, intuition involves believing what feels true. The problem with relying on intuition is that our intuitions can be wrong because they can be driven by cognitive and motivational biases rather than logical reasoning or scientific evidence. While you may feel that summarizing lectures is useful, you don't actually have any evidence that this practice is effective. However, weighing alternatives and thinking of all the different possibilities can be paralyzing for some people and sometimes decisions based on intuition are actually superior to those based on analysis.

Authority

Perhaps one of the most common methods of acquiring knowledge is through authority. This method involves accepting new ideas because some authority figure states that they are true. These authorities include parents, the media, doctors, priests and other religious authorities, the government, and even your professors! While in an ideal world we should be able to trust authority figures, history has taught us otherwise and many instances of atrocities against humanity are a consequence of people unquestioningly following authority (e.g., Salem Witch Trials, Nazi War Crimes). On a more benign level, while reading the whole chapter of your textbook in detail worked when your parents' generation were attending college, this may not make as much sense for students who work full-time, take care of family members, and are used to quick bursts of information through popular media and social media. Instead, students who had had access to the internet their entire lives may be better at skimming large blocks of text to find the key information that they need to accomplish their tasks. These examples illustrate that the problem with using authority to obtain knowledge is that they may be wrong. Nevertheless, much of the information we acquire is through authority because we don't have time to question and independently research every piece of knowledge we learn through authority. But we can learn to evaluate the credentials of authority figures, to evaluate the methods they used to arrive at their conclusions, and evaluate whether they have any reasons to mislead us.

Rationalism (Reason and Logic)

Rationalism involves using logic and reasoning to acquire new knowledge. Using this method premises are stated and logical rules are followed to arrive at sound conclusions. For instance, if you acknowledge the premise that face-to-face classes are designed to share information through in-person lectures and activities and the premise that participating in lectures and activities leads to higher exam scores, then you can come to the rational conclusion that attending class will lead to higher exam scores. The problem with this method is that if the premises are wrong or there is an error in logic then the conclusion will not be valid. For instance, the premise that information is shared in class is incorrect for "flipped classrooms" in which the lecture

information is reviewed individually at home while the homework and activities are completed in class with the instructor. Also, unless formally trained in the rules of logic, it is easy to make an error. Nevertheless, if the premises are correct and logical rules are followed appropriately then this is a sound means of acquiring knowledge.

Empiricism (Data or Experience)

Empiricism involves acquiring knowledge through observation and experience. If you see how your friends study, you may decide that what they are doing is the best option. You could even try something out to see how it works for you. However, this is not a systematic way to gain knowledge. Plus, we are limited in what we, personally, can experience and observe. Moreover, our prior experiences can alter the way we perceive events. Nevertheless, empiricism is at the heart of the scientific method. Science relies on observations. But not just any observations, science relies on structured observations which is known as systematic empiricism.

Scientific Method

The scientific method is a process of systematically collecting and evaluating evidence to test ideas and answer questions. While scientists may use intuition, authority, rationalism, and empiricism to generate new ideas, they don't stop there. Scientists go a step further by using systematic empiricism to make careful observations under various controlled conditions in order to test their ideas and they use rationalism to arrive at valid conclusions. While the scientific method is the most likely of all of the methods to produce valid knowledge, like all methods of acquiring knowledge it also has its drawbacks.

One major problem is that it is not always feasible to use the scientific method; this method can require considerable time and resources. Another problem with the scientific method is that it cannot be used to answer all questions.

Another issue is that the focus of these methods of knowing, and the scientific method in particular, is based on Western (e.g., European) conceptions of science. Others have noted that Indigenous scholars find that their cultural approaches to knowledge to do not fit into these categories, primarily because Indigenous knowledge tends to be relational by emphasizing relationships with the land and relationships with the community (Held, 2019) but the strength of scientific method is controlling the environment to try to get "standard" results. While there is no one set of Indigenous methods that encompass all ways that cultures gain and recognize knowledge due to the diversity of tribal nations and cultures (Simonds & Christopher, 2013), Held (2019) argues that Indigenous research has a common understanding of interconnectedness and interdependence amongst individuals, communities, and their local land. Additionally, storytelling is considered a method of knowing in some Indigenous cultures (Simonds & Christopher, 2013).

Consider for Yourself:

Where would you put storytelling as a method of knowing?

If you would incorporate storytelling into one of the five categories above, which category would you add it to and why? If you would not incorporate storytelling into one of the five categories above, why not?

The title of this section (How do you know?) is based off of the song "That's How You Know" (link to YouTube) from the movie *Enchanted*.

References

Held, M. B. E. (2019). Decolonizing research paradigms in the context of settler colonialism: An unsettling, mutual, and collaborative effort. *International Journal of Qualitative Methods, 18,* 1-16. DOI: 10.1177/1609406918821574

Simonds, V. W., and Christopher, S. (2013). Adapting Western research methods to Indigenous ways of knowing. *American Journal of Public Health, 103*(12), 2185-2192. DOI: 10.2105/AJPH.2012.301157

This page titled 1.3: How Do You Know? is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, and Dana C. Leighton via source content that was edited to the style and standards of the LibreTexts platform.

- 4.2: Ch 2.2 Histogram has no license indicated.

1.3.1: The Scientific Method

I can hear you now, “What? You just convinced me that I will need to know something about statistics to be good in my future career, but, seriously, the scientific method?”

Yup. It turns out that the social sciences are a *science*. And each step of the scientific method is related to what you’ll learn in this course.

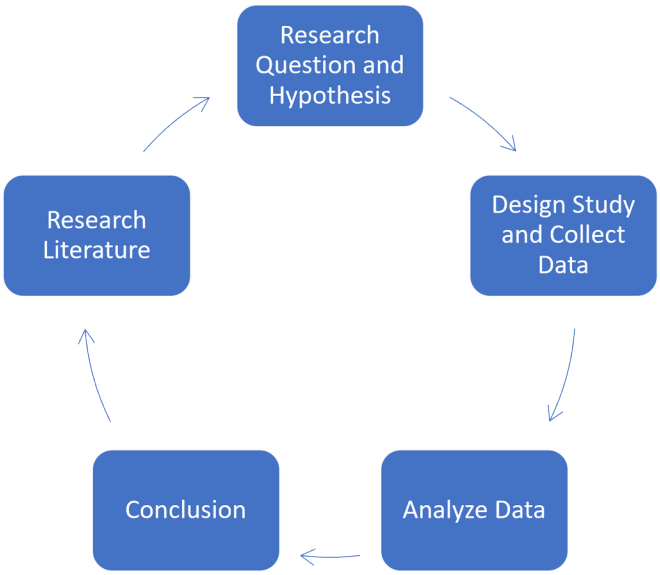


Figure 1.3.1.1: Model of Stages of the Scientific Method. (CC-BY-SA; Michelle Oja via source)

Figure 1.3.1.1 shows a model of the scientific method as a cycle. The basic idea is that to learn about the world, we often start with questions which then lead to developing research hypotheses. Once we have a hypothesis, we can decide how to test that hypothesis by designing a study to collect data to test that hypothesis (and then collecting the data). Some argue that the Research Literature stage should go between these, too. Sometimes, your question has already been answered by others! For researchers, sometimes they start with reading the academic literature, and then that leads them to research questions and hypotheses. For our purposes, we’ll start with a research question that leads to developing a research hypothesis, then design a study and collect data to test that research hypothesis. Once we have some data, we can analyze it in the next stage. In the social sciences, analyzing the data is not the last step. We then use the results to answer the research question, or make a conclusion. If we started with research questions, then the last stage is sharing the conclusion! There are tons of ways to do this, with the most formal mode being an academic research paper or presentation.

Exercise 1.3.1.1

Which of the stages of the scientific method do you think that a behavioral statistics class will focus on?

Answer available at the end of this section.

The following example of using the scientific method (Table 1.3.1.2) is based off of Part 2 (What Students Should Know about How People Learn) and Part 3 (Cognitive Principles for Optimizing Learning) of Dr. Chew's five-part video series on the best way to learn material and pass your classes. Dr. Chew is a cognitive psychologist, so his career is to learn how people think, process material, and learn.

Table 1.3.1.1*: Scientific Method*

Stages of the Scientific Method	Description of the Step of the Scientific Method	Example of the Step of the Scientific Method
Research Question and Hypothesis	The research questions could be pretty broad. This is a sentence that states a predicted relationship between two or more groups. This should not be a question, and it should not include the word "because" or else you actually have two hypotheses (the predicted relationship, and the reason why the relationship is that way).	RQ: "What should I do to pass this class?" RH: "Students who re-organize their notes will earn more points on their first paper than students who re-read their notes."
Design Study and Collect Data	There are whole classes that help you learn how best to design studies to answer your research questions. The important part now is to identify who you are collecting data from (sample) and how you will count or measure the variables in your Research Hypothesis.	"At the end of each class session, half of the class will spend 20 minutes rewriting and organizing their notes, while the other half of the class will spend 20 minutes reviewing (re-reading) their notes. We will then compare their points earned on the first paper between these two groups."
Analyze Data	Remember the second reason that you have to take this class about reporting evidence? That's the first step in this part of the Scientific Method. The other part of this stage of the Scientific Method is to test your Research Hypothesis	That's what the rest of this book will be about!
Make Conclusion	This is when you decide if your Research Hypothesis was supported or not. (We don't say that a Research Hypothesis is "proven;" we'll discuss why not when we talk about Null Hypothesis Significant Testing).	Every chapter after the APA Style chapter will help with this!
Research Literature	Let people know what you found!	Typically, researchers present at a conference or publish a research article. In this example, we could also share our findings with the student newspaper or create a flyer or infographic for tutors.

Conclusion

To the surprise of many students, research and statistics is a fairly significant part of the social sciences. To the surprise of no one, statistics is very rarely the favorite part of one's education. After all, if you really loved the idea of doing statistics, you'd probably be enrolled in a statistics class right now, not a *behavioral* statistics class. So, not surprisingly, there's a pretty large proportion of the student base that isn't happy about the fact that their major has so much research and statistics in it. By the end of this chapter, if not this textbook, hopefully you'll understand why research and statistics are so important!

Answers

Exercise 1.3.1.1

Most students new to this course think that the emphasis will be on Analyze Data, but that's just a step to get to Making Conclusions and sharing our findings.

In this text, we will focus on Analyzing Data and writing up our findings (Research Literature), but writing up our results involves Making Conclusions that test Research Hypotheses! And, we do need to know a little about how the data was collected to know how to analyze the data. In sum, we'll touch on all of these stages!

This page titled 1.3.1: The Scientific Method is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Danielle Navarro.

- 1.3: The Scientific Method by Danielle Navarro is licensed CC BY-SA 4.0.

1.4: Types of Data

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. In psychology, we are interested in people, so we might get a group of people together and measure their levels of stress (one variable), anxiety (a second variable), and their physical health (a third variable). Once we have data on these three variables, we can use statistics to understand if and how they are related. Before we do so, we need to understand the nature of our data: what they represent and where they came from.

Types of Variables

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

Example 1.4.1

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

a. What is the independent variable?
b. What are the dependent variables?

Solution

a. IV: dietary supplement: none, blueberry, strawberry, and spinach
b. DVs: memory test and motor skills test

One more together, and then one on your own:

Example 1.4.2

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

a. What is the independent variable?
b. What is the dependent variable?

Solution

a. IV: Supplements: beta-carotene or placebo
b. DV: Cancer occurrence

Your turn!

Exercise 1.4.1

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

1. What is the independent variable?
2. What is the dependent variable?

Answer available at the end of this section.

Levels of an Independent Variable

If an experiment compares an experimental treatment with a control treatment, then the independent variable (type of treatment) has two levels: experimental and control. If an experiment were comparing five types of diets, then the independent variable (type of diet) would have 5 levels. In general, the number of levels of an independent variable is the number of experimental conditions.

Qualitative and Quantitative Variables

An important distinction between variables is between qualitative variables and quantitative variables. Qualitative variables are those that express a qualitative attribute such as hair color, eye color, religion, favorite movie, gender, and so on. The values of a qualitative variable do not imply a numerical ordering. Values of the variable "religion" differ qualitatively; no ordering of religions is implied. Qualitative variables are sometimes referred to as categorical variables. Quantitative variables are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and seconds.

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

Some variables appear quantitative because they are numbers, but are actually qualitative because the numbers don't mean anything. For example, I could code a sample's birth location as California equals 1, any other state equals 2, and outside of the United States equal 3. This does not mean that birth location is a quantitative variable! We will discuss these difference in more detail in a later section.

Discrete and Continuous Variables

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

Levels of Measurement

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

Although procedures for measurement differ in many ways, they can be classified using a few fundamental categories. In a given category, all of the procedures share some properties that are important for you to know about. The categories are called "scale

types," or just "scales," and are described in this section.

Nominal scales

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

Ordinal scales

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

On the other hand, ordinal scales fail to capture important information that will be present in the other scales we examine. In particular, the difference between two levels of an ordinal scale cannot be assumed to be the same as the difference between two other levels. In our satisfaction scale, for example, the difference between the responses "very dissatisfied" and "somewhat dissatisfied" is probably not equivalent to the difference between "somewhat dissatisfied" and "somewhat satisfied." Nothing in our measurement procedure allows us to determine whether the two differences reflect the same difference in psychological satisfaction. Statisticians express this point by saying that the differences between adjacent scale values do not necessarily represent equal intervals on the underlying scale giving rise to the measurements. (In our case, the underlying scale is the true feeling of satisfaction, which we are trying to measure.)

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

Interval scales

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

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

Ratio scales

The ratio scale of measurement is the most informative scale. It is an interval scale with the additional property that its zero

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

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

Another example of a ratio scale is the amount of money you have in your pocket right now (25 cents, 55 cents, etc.). Money is measured on a ratio scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have zero money, this implies the absence of money. Since money has a true zero point, it makes sense to say that someone with 50 cents has twice as much money as someone with 25 cents (or that Bill Gates has a million times more money than you do).

Let's practice!

Exercise 1.4.2

For each of the following, determine the level of measurement:

1. T-shirt size
2. Time taken to run 100 meter race
3. First, second, and third place in 100 meter race
4. Birthplace
5. Temperature in Celsius

Answer available at the end of this section.

What level of measurement is used for psychological variables?

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

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

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

Table 1.4.1: Sample Data

Participant	Easy Item 1	Easy Item 2	Easy Item 3	Easy Item 4	Easy Item 5	Difficult Item 1	Difficult Item 2	Difficult Item 3	Difficult Item 4	Difficult Item 5	Total Score
A	0	0	1	1	0	0	0	0	0	0	2
B	1	0	1	1	0	0	0	0	0	0	3
C	1	1	1	1	1	1	1	0	0	0	7

Participant	Easy Item 1	Easy Item 2	Easy Item 3	Easy Item 4	Easy Item 5	Difficult Item 1	Difficult Item 2	Difficult Item 3	Difficult Item 4	Difficult Item 5	Total Score
D	1	1	1	1	1	0	1	1	0	1	8

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

Consequences of Scale of Measurement

Why are we so interested in the type of scale that measures a dependent variable? The crux of the matter is the relationship between the variable's level of measurement and the statistics that can be meaningfully computed with that variable. For example, consider a hypothetical study in which 5 children are asked to choose their favorite color from blue, red, yellow, green, and purple. The researcher codes the results as follows:

Table 1.4.2*: Favorite Color Data Code*

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

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

Table 1.4.3*: Favorite Color Data*

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

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

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

Answers

Exercise 1.4.1

1. IV: Brightness of brake lights
2. DV: Time to hit brakes

Exercise 1.4.2

1. Ordinal
2. Ratio
3. Ordinal
4. Nominal
5. Interval

This page titled 1.4: Types of Data 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.

1.4.1: IV and DV

Okay, we've got more terminology before moving away from variables. All variables can be defined by their Scale of Measurement. Variables in research can also be described by whether the experimenter thinks that they are the cause of a behavior (IV), or the effect (DV). The IV is the variable that you use to do the explaining and the DV is the variable being explained. It's important to keep the two roles "thing doing the explaining" and "thing being explained" distinct. So let's be clear about this now.

Definition: Independent Variable (IV)

The variable that the researcher thinks is the *cause* of the effect (the DV). The IV is sometimes also called a "predictor" or "predicting variable".

A true IV is created by the experimenter, but sometimes we measure something that we think is the cause and call it an "IV." If we just measure the groups, then we can't be sure that the IV is causing the DV, so it's better to create the groups through random assignment. IVs are often qualitative/nominal; for most of this textbook, the IV levels are the groups that we are comparing.

Note

Look up "independent variable" online. What is that definition? How is it similar or different from the one provided? Which makes more sense to you?

In the Scientific Method example ((Table 1.4.1.2)), the IV levels were students who re-organized their notes, and students who only re-read their notes.

Definition: Dependent Variable (DV)

The variable that you think is the *effect* (the thing that the IV changes). The DV is the outcome variable, the thing that you want to improve.

DVs are always measured. DVs can be qualitative, but for most of this textbook the DV will be quantitative because we will be comparing means.

Note

Look up "dependent variable" online. What is that definition? How is it similar or different from the one provided? Which makes more sense to you?

The logic behind these names goes like this: if there really is a relationship between the variables that we're looking at, then we can say that DV *depends on* the IV. If we have designed our study well, then the IV isn't dependent on anything else.

Practice

Although the terms IV and DV are misleading, they are still the standard phrasing so that's what we'll work with in the following examples. In the Scientific Method example, the DV was the points earned on the first paper. The DV is what we want to improve, and the IV is a group that we think will do better on the DV plus at least one comparison group (sometimes called a control group).

Exercise 1.4.1.1

1. SAT score & type of high school (public vs. private)
    1. Q: Which variable is the IV? (groups or levels)?
    2. Q: Which variable is the DV?
2. Salary & type of job obtained
    1. Q: Which variable is the IV? (groups or levels)?
    2. Q: Which variable is the DV?
3. Fail rate & brand of computer
    1. Q: Which variable is the IV? (groups or levels)?
    2. Q: Which variable is the DV?
4. Exam score & method of course delivery (online vs. hybrid vs. face-to-face)
    1. Q: Which variable is the IV? (groups or levels)?
    2. Q: Which variable is the DV?
5. Type of phone and how long the battery lasts
    1. Q: Which variable is the IV? (groups or levels)?
    2. Q: Which variable is the DV?

Answer available at the end of this section.

Answers

Exercise 1.4.1.1

1. SAT score & type of high school (public vs. private)
    1. Q: Which variable is the IV? (groups or levels)? type of high school
    2. Q: Which variable is the DV? SAT score
2. Salary & type of job obtained
    1. Q: Which variable is the IV? (groups or levels)? type of job obtained
    2. Q: Which variable is the DV? Salary
3. Fail rate & brand of computer
    1. Q: Which variable is the IV? (groups or levels)? brand of computer
    2. Q: Which variable is the DV? Fail rate
4. Exam score & method of course delivery (online vs. hybrid vs. face-to-face)
    1. Q: Which variable is the IV? (groups or levels)? course delivery
    2. Q: Which variable is the DV? Exam score
5. Type of phone and how long the battery lasts
    1. Q: Which variable is the IV? (groups or levels)? type of phone
    2. Q: Which variable is the DV? how long the battery lasts

This page titled 1.4.1: IV and DV is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

1.4.2: Types of Variables

Some researchers call the first two scales of measurement (Ratio Scale and Interval Scale) “quantitative” because they measure things numerically, and call the last scale of measurement (Nominal Scale) “qualitative” because you count the number of things that have that quality.

MooMoo Math on Types of Variables

The MooMooMath YouTube series did a short segment on these two types of variables.

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online



It turns out that there are a LOT of videos online about statistics! Use them any time you are confused! My only caution is that some videos use slightly different formulas than in this textbook, and some use software that will not be discussed here, so make sure that the information in the video matches what your professor is showing you.

Ordinal scales are sort of in-between these two types. In terms of types of variables, ordinal scales should probably be considered qualitative variables because they aren't quite the ratio scale or the interval scale. However, in real life in statistical analyses, ordinal variables are often treated as quantitative variables.

Quantitative Variables

Quantitative variables are measured with some sort of scale that uses numbers. For example, height can be measures in the number of inches for everyone. Halfway between 1 inch and two inches has a *meaning*. Anything that you can measure with a number and finding a mean makes sense is a quantitative variable. If a decimal makes sense, then the variable is quantitative.

Definition: Quantitative Variable

Type of variable that is measured with some sort of scale that uses numbers that measure something.

Quantitative variables are usually continuous.

Note

Look up “quantitative variable” online. What is that definition? How is it similar or different from the one provided? Which makes more sense to you?

Let's look at an example.

Example 1.4.2.1

The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Is the weight of the backpacks a quantitative variable?

Solution

Yes, the weights are quantitative data because weight is a numerical variable that is measured.

Qualitative Variables

Qualitative variables, which are the nominal Scale of Measurement, have different values to represent different *categories* or kinds. Qualitative/nominal variables name or label different categories of objects.

Definition: Qualitative Variable

Type of variable that has different values to represent different categories or kinds This is the same as the nominal scale of measurement.

If your variable name, it's probably qualitative (especially if it was counted rather than measured).

Note

Look up “qualitative variable” online. What is that definition? How is it similar or different from the one provided? Which makes more sense to you?

Something is either an apple or an orange, halfway between an apple and an orange doesn’t mean anything. Qualitative variables are counted, and the counts are used in statistical analyses. The name or label of a qualitative variable can be a number, but the number doesn’t mean anything. The Exercise below is an example of this.

Exercise 1.4.2.1

Let’s say I collected data and coded it:

- Women = 1
- Men = 2,
- Non-binary = 3

Does it make any sense to add these numbers? To find the “mean” of gender?

Answer available at the end of this section.

Qualitative variables are discrete.

Deciding: Quant or Qual?

Use the following to practice identifying whether variables are quantitative (measured with numbers) or qualitative (categories).

Exercise 1.4.2.2

1. City
2. Gender
3. Weight
4. Type of degree
5. College major
6. Percent correct on Exam 1.
7. Score on a depression scale (between 0 and 10)
8. How long it takes you to blink after a puff of air hits your eye.
9. What is another example of a quantitative variable?
10. What is another example of a qualitative variable?

Answer available at the end of this section.

Okay, that probably makes it seem like it's easy to know whether your variable is qualitative or quantitative.

Exercise 1.4.2.3 shows that variables can be defined in different ways.

Exercise 1.4.2.3

You go to the supermarket and purchase three cans of soup (19 ounces) tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces Cherry Garcia ice cream and two pounds (32 ounces chocolate chip cookies).

Name data sets that are quantitative discrete, quantitative continuous, and qualitative.

Solution

One Possible Solution:

- The three cans of soup, two packages of nuts, four kinds of vegetables and two desserts are quantitative discrete data because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables and desserts are qualitative data because they are categorical.

Try to identify additional data sets in this example.

As you'll learn in the next chapter, there are types of graphs that are designed for qualitative variables and other graphs that are most appropriate for quantitative variables. Before you learn about that, why don't you check out these graphs to see if you can figure out whether the variable is qualitative or quantitative.

Exercise 1.4.2.6

When combining the academic majors that students graduated from into categories, it is clear that the social sciences and helping professions are very popular. What type of data does the pie chart in Figure 1.4.2.1 show?

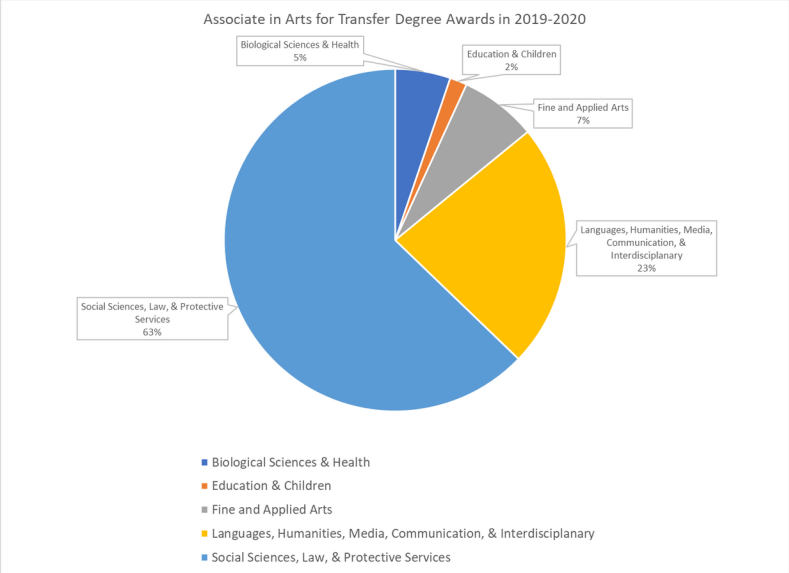


Figure 1.4.2.1- Pie Chart of Graduate's Majors (Copyright CC0; chart created by Michelle Oja via California Community Colleges Chancellor's Office DataMart)

Answer available at the end of this section.

Exercise 1.4.2.7

The Registrar keeps records of the number of credit hours students complete each semester. The data she collects are summarized in the histogram.

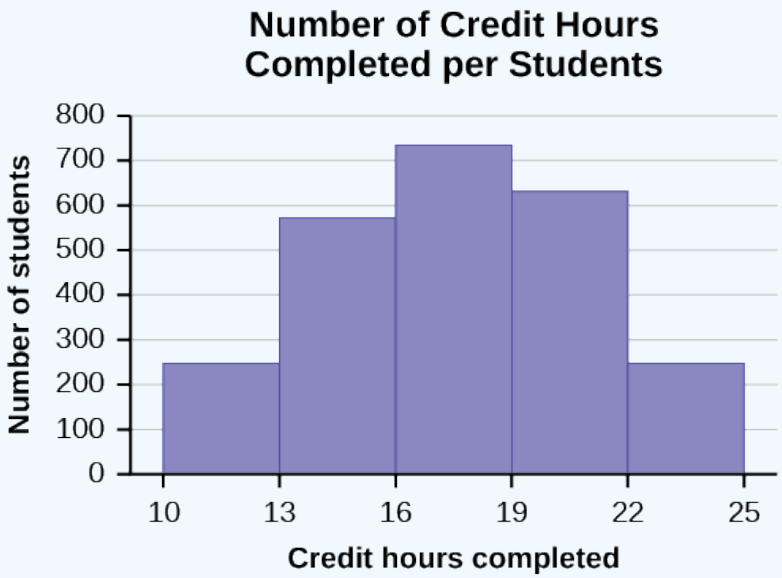


Figure 1.4.2.2- Credit Hours (CC-BY by Barbara Illowsky & Susan Dean (De Anza College) from OpenStax)

What type of data does this graph show?

Answer available at the end of this section.

Ordinal Variables

Ordinal variables are variables with an ordinal Scale of Measurement. This means that ordinal variables are variables that are in an order, but aren't exactly numbers that measure something. I often get shoe size as an example of a quantitative variable, but I really think that's it's ordinal. The numbers aren't measuring anything, but they are in an order.

Definition: Ordinal Variable

Scale of measurement in which levels have an order; a type of variable that can be put in numerical order. The variables are in ranks (first, second, third, etc.).

For another example, I could look at students' letter grade on their final exam. Obviously, students who earn A's have a higher grade than students who earn any other letter grade, but I can't be sure that these letters (qualitative) map directly onto the grade percentages (quantiative). This makes me think that letter grades are ordinal; there is an order of highest to lowest in grades, but it's not a number. Similarly, I could look at everyone who's currently enrolled at the high school that you graduated from by their grade level. I could use numbers (9 for 9th graders), which would be quantitative. I could also use the names that we give them (freshman), and names are qualitative. However, there is an *order* to these grade levels (senior is higher than freshman) so this also feels like an ordinal variable to me.

Note

Look up “ordinal variable” online. What is that definition? How is it similar or different from the one provided? Which makes more sense to you?

I wish it was more clear which type of variable was which! Don't worry too much, though, since we usually treat ordinal variables as quantitative variables in statistics.

Does the type of variable even matter?

The type of scale determines what specific statistical analysis you should use. I'm going to share a flow chart now that shows how knowing the type and number of variables (IVs and levels, and DVs) and whether they are related (dependent) or not related (independent) is how you choose which statistical analysis to choose: Decision Tree: Which Statistical Analysis? I know, that might be a little overwhelming right now! Dr. MO isn't sharing this to scare you, but to show how important knowing the type of variable will be when analyzing data statistically. Plus, it's easier to learn new material if you can connect it to something that you already know. You might want to print out the Decision Tree, then write notes on it when you learn about each type of analysis. That way, you can "hang" your new knowledge on the "tree" that you already have.

In good news, by the end of this book, you'll be familiar with all of these, and know how to compute most of them!

In bad news, statistical software will run what you ask, regardless of the measurement scale of the variable. If it’s a number, you can analyze it. When this happens with qualitative variables, the results are junk. If you say apple=1 and orange=2, it will find the average of an appleorange. Since that's not a thing, you're answer wouldn't mean anything. More reason to understand the different kinds of variables!

Answers

Exercise 1.4.2.1

No. An average gender of 1.75 (or whatever) doesn't tell us much since gender is a qualitative variable (nominal scale of measurement), so you can only count it.

Exercise 1.4.2.2

1. City: Qualitative (named, not measured)
2. Gender: Qualitative (named, not measured)
3. Weight: Quantitative (number measured in ounces, pounds, tons, etc.; decimal points make sense)
4. Type of degree: Qualitative (named, not measured)
5. College major: Qualitative (named, not measured)
6. Percent correct on Exam 1: Quantitative (number measured in percentage points; decimal points make sense)
7. Score on a depression scale (between 0 and 10): Quantitative (number measured by the scale; decimal points make sense)

8. How long it takes you to blink after a puff of air hits your eye: Quantitative (number measured in milliseconds; decimal points make sense)
9. What is another example of a quantitative variable? (Your answer should be something that was measured, not counted, and in which decimal points make sense.)
10. What is another example of a qualitative variable? (Your answer should be something that is a category or name.)

Exercise 1.4.2.6

This pie chart shows the frequency (in percentages) of categories of majors from graduates in 2019-2020. Categories of the major is a qualitative data.

Exercise 1.4.2.7

A histogram is used to display quantitative data: the numbers of credit hours completed.

This page titled 1.4.2: Types of Variables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 1.3: Data, Sampling, and Variation in Data and Sampling by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.
- 1.4.2: Qualitative versus Quantitative Variables by Michelle Oja is licensed CC BY 4.0.

1.4.3: Scales of Measurement

Look at the Figure 1.4.3.1:



Figure 1.4.3.1: 3 Beakers. (CC-BY-SA Photo by Jaeger5432. The file was originally uploaded to English Wikipedia. This file is licensed under the Creative Commons Attribution-Share Alike 2.5 Generic license.)

Example 1.4.3.1

List three ways that the beakers in Figure 1.4.3.1 can be described.

Solution

Some examples include: by color, by size (bigger, smaller), and measurement (ounces, milligrams, etc). Your descriptions probably map pretty well onto the different Scales of Measurement.

Scales of Measurement

A very useful concept for distinguishing between different types of variables is what's known as ***scales of measurement***.

Note

Scales of Measurement

- Ratio: true numerical measurement
- Interval: created numerical scale
- Ordinal: variables with an order
- Nominal: Named variable

Table 1.4.3.1- *Description & Examples of Each Scale of Measurement*

Scale of Measurement	Description of Scale of Measurement	Example from Figure 1.3.1	Another Example
Ratio Scale	A type of variables that is measured, and zero means that there is none of it.	Ounces of liquid (and you could have no liquid)	Liters of ice cream (you could have zero liters of ice cream)
Interval Scale	A type of variable that is measured, and the intervals between each measurement are equal (but zero does not mean the absence of the measured item).	[There are no variables with an interval scale in the image.]	A satisfaction scale in which -5 means "very unsatisfied", 0 means "neutral," and +5 means "very satisfied."
Ordinal Scale	A type of variables that can be put in numerical order, but the difference between any pair is not the same as the difference between any other pair. The variables are in ranks (first, second, third, etc.).	You could order these from biggest to smallest; the difference between the biggest and the middle might be much larger than the difference between the middle and the smallest.	The medals earned in the Olympics; the difference between the person who wins the gold medal and the person who wins the silver medal might be much larger than the difference between the person who wins the silver medal compared to the person who wins the bronze medal.
Nominal Scale	Type of variable that has quality or name, but not a number that means something. "Nom" means "name," so it's just a name, not anything that is measured. This variable is counted, rather than measured.	Color of the liquids	Your major in college.

Things that are measured use some sort of tool (like a ruler, weight scale, or survey scale), which differs from things that can only be counted.

If you're still a little lost (which is fine, these types of variables are confusing and a little squishy), keep reading. If you're good, skip down to the last section in this page, then make sure to read the section on Qualitative and Quantitative variables!

Ratio Scale

The fourth and final type of variable to consider is a ratio scale variable, in which zero really means zero, and it's okay to multiply and divide. A good example of a ratio scale variable is response time (RT). In a lot of tasks it's very common to record the amount of time somebody takes to solve a problem or answer a question, because it's an indicator of how difficult the task is. Suppose that Amar takes 2.3 seconds to respond to a question, whereas Becky takes 3.1 seconds. As with an interval scale variable, addition and subtraction are both meaningful here. Becky really did take 3.1 - 2.3 = 0.8 seconds longer than Amar did. However, notice that multiplication and division also make sense here too: Becky took 3.1 / 2.3 = 1.35 times as long as Amar did to answer the question. And the reason why you can do this is that, for a ratio scale variable such as RT, "zero seconds" really does mean "no time at all".

Interval Scale

In contrast to nominal and ordinal scale variables, ***interval scale*** and ratio scale variables are variables for which the numerical value is genuinely meaningful. In the case of interval scale variables, the *differences* between the numbers are interpretable, but the variable doesn't have a "natural" zero value. A good example of an interval scale variable is measuring temperature in degrees celsius. For instance, if it was 15° yesterday and 18◦ today, then the 3° difference between the two is genuinely meaningful. Moreover, that 3° difference is *exactly the same* as the 3° difference between 7° and 10°. In short, addition and subtraction are meaningful for interval scale variables.[8]

However, notice that the 0° does not mean "no temperature at all": it actually means "the temperature at which fresh water

freezes", which is pretty arbitrary. As a consequence, it becomes pointless to try to multiply and divide temperatures. It is wrong to say that 20° is *twice as hot* as 10°, just as it is weird and meaningless to try to claim that 20° is negative two times as hot as -10°.

Again, lets look at a more psychological example. Suppose I'm interested in looking at how the attitudes of first-year university students have changed over time. Obviously, I'm going to want to record the year in which each student started. This is an interval scale variable. A student who started in 2003 did arrive 5 years before a student who started in 2008. However, it would be completely insane for me to divide 2008 by 2003 and say that the second student started "1.0024 times later" than the first one. That doesn't make any sense at all.

Ordinal Scale

Ordinal scale variables have a bit more structure than nominal scale variables, but not by a lot. An ordinal scale variable is one in which there is a natural, meaningful way to order the different possibilities, but you can't do anything else. The usual example given of an ordinal variable is "finishing position in a race". You *can* say that the person who finished first was faster than the person who finished second, but you *don't* know how much faster. As a consequence we know that 1st > 2nd, and we know that 2nd > 3rd, but the difference between 1st and 2nd might be much larger than the difference between 2nd and 3rd.

I accidently found an ordinal scale when trying to find an example of a nominal scale: highest degree earned. I was looking for categories, but there is definitely an order to level of education!

1. Did not finish high school
2. High school degree only
3. Some college
4. Certificate
5. Associate's degree
6. Bachelor's degree
7. Graduate degree

Since there is a natural order to education level, it would be very weird to list the options like this...

1. Did not finish high school
2. Graduate degree
3. High school degree only
4. Some college
5. Certificate
6. Bachelor's degree
7. Associate's degree

... because it seems against the natural "structure" to the question. Notice that while we *can* use the natural ordering of these items to construct sensible groupings, what we *can't* do is average them. The #1 doesn't have any meaningful numerical value, and is not really related to the #7 in any measurable way. If you can tell me what the mean of these categories might mean, I'd love to know. Because that sounds like gibberish to me!

Nominal Scale

A ***nominal scale*** variable (also referred to as a ***categorical*** variable) is one in which there is no particular relationship between the different possibilities: for these kinds of variables it doesn't make any sense to say that one of them is "bigger' or"better" than any other one, and it absolutely doesn't make any sense to average them. The classic example for this is "eye color". Eyes can be blue, green and brown, among other possibilities, but none of them is any "better" than any other one. As a result, it would feel really weird to talk about an "average eye colour". In short, nominal scale variables are those for which the only thing you can say about the different possibilities is that they are different. That's it.

Let's take a slightly closer look at this. Suppose I was doing research on how students studied. This "study" variable could have quite a few possible values,but for now, let's suppose that these four are the only possibilities, and suppose that when I ask 100 students what they did the last time that they studied, and I get this:

Table 1.4.3.2.1- Example of Nominal Variable

Method of Studying	Number of people
(1) Read the whole chapter	12
(2) Skimmed the whole chapter, read some sections	30
(3) Watched YouTube videos on class content	48
(4) Took notes on the chapter	10

So, what's the average way that students studied? The answer here is that there isn't one. It's a silly question to ask. You can say that watching YouTube is the most popular method, and taking notes is the least popular method, but that's about all. Similarly, notice that the order in which I list the options isn't very interesting. I could have chosen to display the data like this

Table 1.4.3.2.2- Example of Nominal Variable

Method of Studying	Number of people
(3) Watched YouTube videos on class content	48
(2) Skimmed the whole chapter, read some sections	30
(1) Read the whole chapter	12
(4) Took notes on the chapter	10

and it doesn't matter.

Some Complexities

Okay, I know you're going to be shocked to hear this, but ... the real world is much messier than this little classification scheme suggests. Very few variables in real life actually fall into these nice neat categories, so you need to be kind of careful not to treat the scales of measurement as if they were hard and fast rules. It doesn't work like that: they're guidelines, intended to help you think about the situations in which you should treat different variables differently. Nothing more.

So let's take a classic example, maybe *the* classic example, of a psychological measurement tool: the ***Likert scale***. The humble Likert scale is the bread and butter tool of all survey design. You yourself have filled out hundreds, maybe thousands of them, and odds are you've even used one yourself. Suppose we have a survey question that looks like this:

Which of the following best describes your opinion of the statement that "all pirates are freaking awesome" ...

and then the options presented to the participant are these:

(1) Strongly disagree
(2) Disagree
(3) Neither agree nor disagree
(4) Agree
(5) Strongly agree

This set of items is an example of a 5-point Likert scale: people are asked to choose among one of several (in this case 5) clearly ordered possibilities, generally with a verbal descriptor given in each case. However, it's not necessary that all items be explicitly described. This is a perfectly good example of a 5-point Likert scale too:

(1) Strongly disagree
(2)
(3)
(4)
(5) Strongly agree

Likert scales are very handy, if somewhat limited, tools. The question is, what kind of variable are they? They're obviously discrete, since you can't give a response of 2.5. They're obviously not nominal scale, since the items are ordered; and they're not ratio scale

either, since there's no natural zero.

But are they ordinal scale or interval scale? One argument says that we can't really prove that the difference between "strongly agree" and "agree" is of the same size as the difference between "agree" and "neither agree nor disagree". In fact, in everyday life it's pretty obvious that they're not the same at all. So this suggests that we ought to treat Likert scales as ordinal variables. On the other hand, in practice most participants do seem to take the whole "on a scale from 1 to 5" part fairly seriously, and they tend to act as if the differences between the five response options were fairly similar to one another. As a consequence, a lot of researchers treat Likert scale data as if it were interval scale. It's not interval scale, but in practice it's close enough that we usually think of it as being quasi-interval scale.

These different Scales of Measurement are important because they determine which statistical analysis to use on your data. We'll talk about choosing the correct statistical analysis when we know more specific statistical analyses, but that's why we care about what kind of variables we have so much.

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)
- Photo by Jaeger5432. The file was originally uploaded to English Wikipedia. This file is licensed under the Creative Commons Attribution-Share Alike 2.5 Generic license.

This page titled 1.4.3: Scales of Measurement is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

1.5: Populations and Samples

Because social scientists want to make people's lives better, they see what works on a small group of people, and then apply it to everyone. The small group of people is the sample, and "everyone" is the population.

Sample

Definition: Sample

People who participate in a study; the smaller group that the data is gathered from.

A sample is the small group of people that scientists test stuff on. We want at least 30 people in each group, so a study that has two groups will need about 60 people in the sample. Sampl*ing* (a verb) is how the researcher recruits participants into the study. Just because people were invited to participate in the study does not mean that they participated! In other words, a researcher may try sampling 100 people to complete a survey , but only get 10 respondents in their sample.

A random sample, sometimes called a simple random sample, is when a researcher recruits their participants in such a way that anyone in the population has an equal chance of being recruited. For example, if Dr. MO wanted to survey 1,000 students from her college, she could have a computer randomly choose 1,000 participants to invite to participate. This would make a sample that represents the population of her college better than surveying only her students or only students who used the cafeteria; these are convenience samples. The sample would be the students who actually completed the survey out of the 1,000 who were invited.

Although we won't get into it her, random *sampling* is different from random *assignment*. Random assignment is when you already have a sample, and you use some random technique (like flipping a coin, pulling names out of a hate, or using a random number generator) to assign participants into their IV condition. For example, let's say 800 of the randomly chosen 1,000 people that Dr. MO invited to fill out a survey actually completed the survey. The 800 students would be her sample. Then, she could randomly assign different survey titles (even though the questions could be the same); for example, 400 students could be randomly assigned by a computer to fill out a survey entitled "Long and Boring Student Survey" and the other 400 students could be randomly assigned a title like "Short and Fun Survey for Prizes" to see if students at her college, on average, are affected by the name of the survey that they are completing.

Population

Definition: Population

The biggest group that your sample can represent.

A population is the "everyone" that we want to apply the results to. Sometimes, "everyone" can be a pretty small group; if I measured the GPA of one of my Behavioral Statistics classes, then the sample would be the class and the population could be students in all Behavioral Statistics classes at the college. (GPA would be the DV.)

Historical Context

In an early measurement of a population (in the year 2 AD!), Chinese Emperor Wang Mang conducts the first census from which data survives (Champkin, 2014).



Figure 1.5.1: Emperor of the Xin dynasty from Waseda University Library on Wikipedia. Public Domain

A sample is a concrete thing. You can open up a data file, and there's the data from your sample. A population, on the other hand, is a more abstract idea. It refers to the set of all possible people, or all possible observations, that you want to draw conclusions about, and is generally *much* bigger than the sample. In an ideal world, the researcher would begin the study with a clear idea of what the population of interest is, since the process of designing a study and testing hypotheses about the data that it produces does depend on the population about which you want to make statements. However, that doesn't always happen in practice: usually the researcher has a fairly vague idea of what the population is and designs the study as best he/she can on that basis.

Examples

In our Scientific Method example, the sample would be the class from which we got the data from, and the population would be the biggest group that they could represent. There's often more than one possible population, but I might say all college students could be a good population for this sample.

You might have heard the phrase "random sample." This means that everyone in the population has an equal chance of being chosen to be in the sample; this almost never happens.

Exercise 1.5.1

Let's say I want to know if there's a relationship between intelligence and reading science fiction books. If I survey 100 of my Introduction to Psychology students on their intelligence and reading of science fiction:

1. Who is the Sample?
2. Who could be the population? In other words, what is the biggest group that this sample could represent?

Answer available at the end of this section.

Sometimes it's easy to state the population of interest. In most situations the situation is much less simple. In a typical a psychological experiment, determining the population of interest is a bit more complicated. Suppose Dr. Navarro ran an experiment using 100 undergraduate students as participants. Her goal, as a cognitive scientist, is to try to learn something about how the mind works. So, which of the following would count as "the population":

- All of the undergraduate psychology students her university in Australia?
- Undergraduate psychology students in general, anywhere in the world?
- Australians currently living?
- Australians of similar ages to my sample?
- Anyone currently alive?
- Any human being, past, present or future?
- Any biological organism with a sufficient degree of intelligence operating in a terrestrial environment?
- Any intelligent being?

Each of these defines a real group of mind-possessing entities, all of which might be of interest to me as a cognitive scientist, and it's not at all clear which one ought to be the true population of interest. Maybe surprisingly for you, there's no "right" answer! Although some the suggestions get a little vague, they all could potentially be a population that her sample represents. Irrespective of how the population is defined, the critical point is that the sample is a subset of the population. The goal of

researchers is to use our knowledge of the sample to draw inferences about the properties of the population. More on that in later chapters!

Exercise 1.5.1

Actual drug use is much higher than drug arrests suggest, so you might want to measure how many people use marijuana. If you send out a survey asking about their drug use to everyone with a driver's license in California, but only 30% fill it out:

1. Who is the Sample?
2. Who could be the population? In other words, what is the biggest group that this sample could represent?

Answer available at the end of this section.

This last example shows that sometimes our sample limits who we can generalize our results about, who could be our population.

In almost every situation of interest, what we have available to us as researchers is a sample of data. We might have run experiment with some number of participants; a polling company might have phoned some number of people to ask questions about voting intentions; etc. Regardless: the data set available to us is finite, and incomplete. We can't possibly get every person in the world to do our experiment; a polling company doesn't have the time or the money to ring up every voter in the country etc.

Reference

Champkin, J. (2014, January 24). *The timeline of statistics*. Significance. https://www.statsref.com/timeline.pdf

Answers

Exercise 1.5.1

Add texts here. Do not delete this text first.

1. Who is the Sample? 100 of my Introduction to Psychology students
2. Who could be the population? In other words, what is the biggest group that this sample could represent? There are many possible populations, but all Introduction to Psychology students could work, or all Introduction to Psychology students at my college might make sense, too.

Exercise 1.5.1

Add texts here. Do not delete this text first.

1. Who is the Sample? 30% of Californians with driver's licenses
2. Who could be the population? In other words, what is the biggest group that this sample could represent? There are many possible populations, but Californians who have driver's licenses might make the most sense here.

This page titled 1.5: Populations and Samples is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 10.1: Samples, Populations and Sampling by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.
- 1.5: Populations and Samples by Michelle Oja is licensed CC BY-SA 4.0.

1.6: Collecting Data- More Practice with Populations and Samples

We are usually interested in understanding a specific group of people. This group is known as the population of interest, or simply the population. The population is the collection of all people who have some characteristic in common; it can be as broad as “all people” if we have a very general research question about human psychology, or it can be extremely narrow, such as “all freshmen psychology majors at Midwestern public universities” if we have a specific group in mind.

Populations and samples

In statistics, we often rely on a sample --- that is, a small subset of a larger set of data --- to draw inferences about the larger set. The larger set is known as the population from which the sample is drawn.

Example 1.6.1

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

Solution

It is not practical to ask every single American how he or she feels about the fairness of the voting procedures. Instead, we query a relatively small number of Americans, and draw inferences about the entire country from their responses. The Americans actually queried constitute our sample of the larger population of all Americans.

A sample is typically a small subset of the population. In the case of voting attitudes, we would sample a few thousand Americans drawn from the hundreds of millions that make up the country. In choosing a sample, it is therefore crucial that it not over-represent one kind of citizen at the expense of others. For example, something would be wrong with our sample if it happened to be made up entirely of Florida residents. If the sample held only Floridians, it could not be used to infer the attitudes of other Americans. The same problem would arise if the sample were comprised only of Republicans. Inferences from statistics are based on the assumption that sampling is representative of the population. If the sample is not representative, then the possibility of sampling bias occurs. Sampling bias means that our conclusions apply only to our sample and are not generalizable to the full population.

Example 1.6.2

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

Solution

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

To solidify your understanding of sampling bias, consider the following example. Try to identify the population and the sample, and then reflect on whether the sample is likely to yield the information desired.

Example 1.6.3

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

Solution

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

Example 1.6.4

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

Solution

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

Simple Random Sampling

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

Example 1.6.5

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

Solution

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

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

Sample size matters

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

More complex sampling

Sometimes it is not feasible to build a sample using simple random sampling. To see the problem, consider the fact that both Dallas and Houston are competing to be hosts of the 2012 Olympics. Imagine that you are hired to assess whether most Texans prefer Houston to Dallas as the host, or the reverse. Given the impracticality of obtaining the opinion of every single Texan, you must construct a sample of the Texas population. But now notice how difficult it would be to proceed by simple random sampling. For example, how will you contact those individuals who don’t vote and don’t have a phone? Even among people you find in the telephone book, how can you identify those who have just relocated to California (and had no reason to inform you of their move)? What do you do about the fact that since the beginning of the study, an additional 4,212 people took up residence in the state of Texas? As you can see, it is sometimes very difficult to develop a truly random procedure. For this reason, other kinds of sampling techniques have been devised. We now discuss two of them.

Stratified Sampling

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

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

Convenience Sampling

Not all sampling methods are perfect, and sometimes that’s okay. For example, if we are beginning research into a completely unstudied area, we may sometimes take some shortcuts to quickly gather data and get a general idea of how things work before fully investing a lot of time and money into well-designed research projects with proper sampling. This is known as convenience sampling, named for its ease of use. In limited cases, such as the one just described, convenience sampling is okay because we intend to follow up with a representative sample. Unfortunately, sometimes convenience sampling is used due only to its convenience without the intent of improving on it in future work.

This page titled 1.6: Collecting Data- More Practice with Populations and 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.

- 1.4: Collecting Data by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

1.7: "Research shows that..."

You may have heard someone say, “Research shows that...” What this means is that an experimenter used the scientific method to test a Research Hypothesis, and either supported that Research Hypothesis or did not support that Research Hypothesis. However, much of what you read online is *not* supported by research using the scientific method. Instead, people who don’t know what an IV or DV is try to convince you that their ideas are correct.

As a group, scientists seem to be bizarrely fixated on running statistical tests on everything. In fact, we use statistics so often that we sometimes forget to explain to people why we do. It’s a kind of article of faith among scientists – and especially social scientists – that your findings can’t be trusted until you’ve done some stats. Undergraduate students might be forgiven for thinking that we’re all completely mad, because no one takes the time to answer one very simple question:

> *Why do you do statistics? Why don’t scientists just use common sense?*

There’s a lot of good answers to it,[1] but a really simple answer is: we don’t trust ourselves enough. Scientists, especially psychologists, worry that we’re human, and susceptible to all of the biases, temptations and frailties that humans suffer from. Much of statistics is basically a safeguard. Using “common sense” to evaluate evidence means trusting gut instincts, relying on verbal arguments and on using the raw power of human reason to come up with the right answer. Most scientists don’t think this approach is likely to work.

In fact, come to think of it, this sounds a lot like a psychological question. In this behavioral statistics textbook, it seems like a good idea to dig a little deeper here. Is it really plausible to think that this “common sense” approach is very trustworthy? Verbal arguments have to be constructed in language, and all languages have biases – some things are harder to say than others, and not necessarily because they’re false (e.g., quantum electrodynamics is a good theory, but hard to explain in words). The instincts of our “gut” aren’t designed to solve scientific problems, they’re designed to handle day to day inferences – and given that biological evolution is slower than cultural change, we should say that our gut is designed to solve the day to day problems for a different world than the one we live in. Most fundamentally, reasoning sensibly requires people to engage in “induction”, making wise guesses and going beyond the immediate evidence of the senses to make generalizations about the world. If you think that you can do that without being influenced by various distractors, well, I have a bridge in Brooklyn I’d like to sell you.

The Curse of Belief Bias

People are mostly pretty smart. Our minds are quite amazing things, and we seem to be capable of the most incredible feats of thought and reason. That doesn’t make us perfect though. And among the many things that psychologists have shown over the years is that we really do find it hard to be neutral, to evaluate evidence impartially and without being swayed by pre-existing biases.

But first, suppose that people really are perfectly able to set aside their pre-existing beliefs about what is true and what isn’t, and evaluate an argument purely on its logical merits. We’d expect 100% of people to say that the valid arguments are valid, and 0% of people to say that the invalid arguments are valid. So if you ran an experiment looking at this, you’d expect to see data like this:

Table 1.7.1- *Valid Arguments and Feelings*

Valid?	Conclusion feels true	Conclusion feels false
Argument is valid	100% say “valid”	100% say “valid”
Argument is invalid	0% say “valid”	0% say “valid”

If the data looked like this (or even a good approximation to this), we might feel safe in just trusting our gut instincts. That is, it’d be perfectly okay just to let scientists evaluate data based on their common sense, and not bother with all this murky statistics stuff. However, you have taken classes, and might know where this is going . . .

In a classic study, Evans, Barston, and Pollard (1983) ran an experiment looking at exactly this. What they found is show in Table 1.7.2; when pre-existing biases (i.e., beliefs) were in agreement with the structure of the data, everything went the way you’d hope (in bold). Not perfect, but that’s pretty good. But look what happens when our intuitive feelings about the truth of the conclusion

run against the logical structure of the argument (the not-bold percentages):

Table 1.7.2- *Beliefs Don't Match Truth*

Valid?	Conclusion feels true	Conclusion feels false
Argument is valid	**92% say "valid"**	46% say "valid"
Argument is invalid	92% say "valid"	**8% say "valid"**

Almost all people continue to believe something that isn't true, even with a valid argument against it. Also, almost half of people continue to believe that true arguments are false (if they started out not believing it).

Oh dear, that's not as good. Apparently, when we are presented with a strong argument that contradicts our pre-existing beliefs, we find it pretty hard to even perceive it to be a strong argument (people only did so 46% of the time). Even worse, when people are presented with a weak argument that agrees with our pre-existing biases, almost no-one can see that the argument is weak (people got that one wrong 92% of the time!).

Overall, people did do better than a chance at compensating for their prior biases, since about 60% of people's judgments were correct (you'd expect 50% by chance). Even so, if you were a professional "evaluator of evidence", and someone came along and offered you a magic tool that improves your chances of making the right decision from 60% to (say) 95%, you'd probably jump at it, right? Of course you would. Thankfully, we actually do have a tool that can do this. But it's not magic, it's statistics. So that's reason #1 why scientists love statistics. It's just *too easy* for us humans to continue to "believe what we want to believe"; so if we want to "believe in the data" instead, we're going to need a bit of help to keep our personal biases under control. That's what statistics does: it helps keep us honest.

Using statistics to find what works, rather than their intuition or prior beliefs, Kozyreva et al. (2020) (link to PDF) found that lateral reading was an effective strategy to avoid being duped! Lateral reading is when you look at alternative sources to fact check an idea. It's best to use trustworthy sources, and a variety of alternative sources. It's not actually lateral reading if you only review what sources that you know would already agree!

Before you "like" or share news articles or posts online, try to get to get as close to the original study as you can get, and figure out what the IV, DV, sample, and population is so that you can decide if their clam is actually supported by research. Was the IV created (not just measured)? Does the way that the DV was measured make sense for that outcome? Does the sample really represent the population that they are saying it does? Then, check with other sources; do they find something similar? Science is hard to do well, but it's a great way to learn new things about the world.

---

[1] Including the suggestion that common sense is in short supply among scientists.

Reference

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*, 295-306.

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the interne: Confronting digital challenges with cognitive tools. Psychological Science in the Pulbic Interest, 21(3), 103-156. DOI: 10.1177/1529100620946707

This page titled 1.7: "Research shows that..." is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Danielle Navarro.

- 1.2: The Cautionary Tale of Simpson's Paradox by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.

1.8: Learning (Statistics)

As we close this chapter, if you are not at least a little overwhelmed, then you probably didn't read the chapter closely enough. We went over many, many words that you know what they mean, but now they have technical definitions use in statistics and research that much more specific than how most people use the words. When learning a new field, you must learn new vocabulary. In this case, you are learning statistics for the social sciences. A disadvantage of the social sciences is that so much of the vocabulary that we use are technical definitions of words that you are familiar with.

What can you do to learn?

Learn How To Learn

This five-part video series from cognitive psychologist Dr. Chew, first mentioned in 1.3: Scientific Method, provides many suggestions for how best to learn in any class. We will be returning to some of these ideas throughout the first unit. You might want to watch one video every day for the first week of class to get in the practice of spending time and effort on your classes. Plus, the videos have a lot of information in them, so it's best not to watch them all at once.

- 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?

Spend Time Before, During, & After Class

In his videos, Dr. Chew emphasizes how important time and effort is for learning. In addition to some of the practices that he suggests for learning, you can also read this article by Zanardelli Sickler (2017) that describes what you can do before your class session or lecture video, what you should do during lecture, and what you can do after lecture to make sure that you learn and understand the material.

Another strategy is to decide what activities can be done easily enough on your phone for when you are on the go or don't have access to a bigger screen. For example, if you are reading this textbook (or any online material for your classes, really), then your phone might be a great option, especially if you are taking notes in an online document. Your phone might be a great way to preview assignments so that you can start thinking about what you'll be submitting. However, maybe taking quizzes should be reserved when you have a bigger screen and time to think. Doing practice problems are probably best when you have a table or desk to write on. Just planning ahead for what activities can be done when and how will help you organize your studying.

Practice!

- Use the Examples and Exercises* throughout the chapters as practice. Don't just read them, but try to do them on your own, then check to see how close you are to the answers.
- Utilize this support "course" or try to finish the chapters during the first few weeks of the semester: Support Course for Elementary Statistics, which is a course to prepare you to take a more math-focused statistics course (rather than social science focused like this textbook)
- Your school probably has tutors or a learning lab (math lab?), so set up *weekly meetings* to go over what you're learning in class. Don't just show up in a panic before an exam or paper is due; learning takes time.
- Work in pairs to check each others' work. But don't do exactly what your partner is doing (that looks like cheating).
- What else might work for *you*?

Phew! That was a lot! Take a break, and I'll meet you in the next chapter to learn about organizing data!

---

*Notes on Examples and Exercises

- When reviewing on LibreTexts, the answers for the Exercises (but not the Examples) will be hidden so that you can work on your own, then check your work. However, the PDF version of the textbook shows the answers for both Examples and Exercises.
- Some faculty cluster their courses into units, so some Examples and Exercises are based on a theme for these potential units. The potential units and themes are as follows:
  - Unit 1: Description theme is student success
  - Unit 2: Mean Differences theme is growth mindset
  - Unit 3: Relationships theme is nutrition
  - Unit 4: Wrap-Up theme is future careers

This page titled 1.8: Learning (Statistics) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

2: WHAT DO DATA LOOK LIKE? (GRAPHS)

2.1: Introduction to Looking at Data
2.2: Frequency Tables
2.3: APA Style Tables
2.4: Graphing Qualitative Variables- Bar Graphs
2.5: Graphing Qualitative Variables- Pie Charts
2.6: Graphing Quantitative Variables
2.7: Skew and Kurtosis
2.8: Graphing Quantitative Data- Line Graphs
2.9: Graphing Quantitative Data- Histograms
2.10: Graphing Quantitative Data- Means Plot
2.11: Graphing Quantitative Data- Boxplots
2.12: Graphing Quantitative Data- Scatterplots
2.13: Summary and Some Honesty
2.14: APA Style Charts

This page titled 2: What Do Data Look Like? (Graphs) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

2.1: Introduction to Looking at Data

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 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 available at the end of this section.

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 different ethnic groups who are taking credit courses at California community college in academic year 2022-2023 ("x" column). This data can be found on the California Community Colleges Chancellor's Office DataMart.

Table 2.2.3- California Community College Students' Ethnicities

x	f
African-American	88,167
American Indian/Alaskan Native	5,333
Asian/Filipino/PI	215,591
Hispanic	754,092
Multi-Ethnicity	70,950
Unknown	57,929
White Non-Hispanic	362,931

The ethnic categories 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 ethnic group, an alphabetical order works best. But if you'd like emphasize which ethnic categories has the most students, 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)

Answers

Exercise 2.2.1

1. One
2. Three

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: APA Style Tables

You are probably a student in the social sciences. Most fields in the social sciences use the Publication Manual of the American Psychological Association (Seventh Edition). This style guide is full of information on formatting and writing in APA Style. In this Seventh Edition, you are allowed to put tables in an appendix (which would be the last section of the paper), or near where they are being discussed in the appear.

For our purposes right now, we'll be looking at how to format the titles for tables (which have rows and columns, and are not charts or pictures). In APA Style, tables should be numbered and have a descriptive title. "Table 1" should be in bold, and on the line above the title (which should be in italics). They both should be above the table and flush-left. An example is below, and you can find more details at the best site to learn about APA Style: Purdue Online Writing Lab. Here is OWL Purdue's page on tables and figures in APA Style. Below is a Note on what the title and number should look like in APA Style.

Note

Table 1

Final Exam Scores

You might notice that this is not at all like how the table titles have been throughout this textbook so far. And there's a good reason for this! APA Style formatting is for sending your work to a publisher. Forcing everyone to format their papers the same way makes it easier on the reviewers, and makes the playing field more fair for the researchers. With a standard formatting style, no one is given preference for fancy graphics or exciting desktop publishing; instead, your work is judged on the quality of the science.

Once a work has been accepted for publishing, the formatting is then modified to fit the formatting of the journal, book, or website where it is being published. That's what is happening here! It might be confusing to see the table titles in one format in your textbook while your own work should be formatted differently. Sorry! When in doubt with formatting, go to your APA Manual or OWL Purdue's APA Style website!

This page titled 2.3: APA Style Tables 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- 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. describes.

Bar Charts

Bar charts are meant to display the frequencies of qualitative variables. Figure 2.4.1 shows the same data as the frequency table in Table 2.2.3, but with bars.

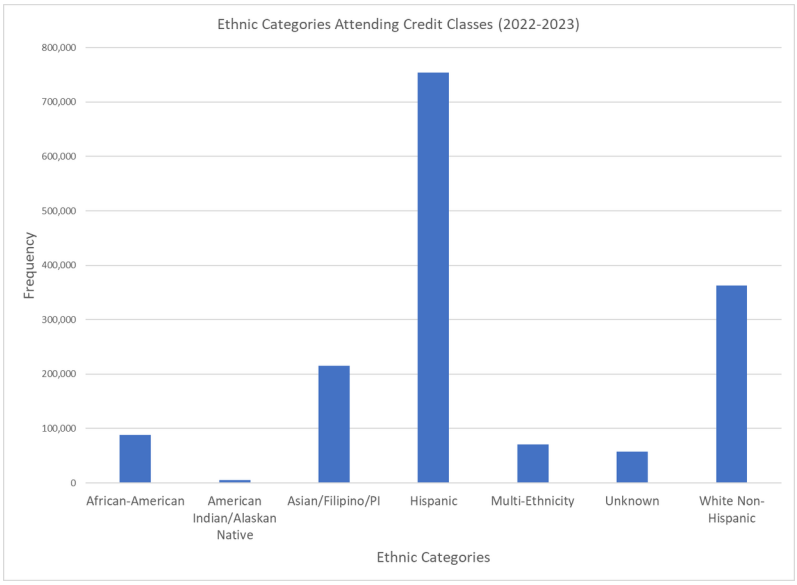


Figure 2.4.1: California Community College Students by Ethnicity (Copyright CC0; chart created by Michelle Oja via California Community Colleges Chancellor's Office DataMart)

Let's unpack Figure 2.4.1 a bit.

First, the x-axis has the same information as the "x" column in Table 2.2.3, the ethnic group that each student self-identified as to their college. 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 2.4.1 shows the same information as the "f" column in Table 2.2.3, the number of students in each ethnic category. 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 2.4.1?
    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 2.4.1?
    1. The x-axis is the one on the bottom, and it was named “Ethnic Categories”. This is the ethnic group that each student is from out of all students attending credit courses at a California community college in the 2022-2023 academic year.
3. What does the y-axis measure in Figure 2.4.1?
    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 were from each ethnic group.
4. What do you notice from Figure 2.4.1? What pops out to you?
    1. What I noticed is that most students are Hispanic, and that there are very few Native Americans.
5. What does Figure 2.4.1 make you wonder about?
    1. I wonder why so many students chose not to identify their ethnicity to their colleges; there are almost as many of those unidentified students as there are multi-ethnic students!
6. What is a catchy headline for Figure 2.4.1?
    1. Join the Tapestry of California Community College
7. How could you summarize the info in Figure 2.4.1 into one sentence?
    1. The most frequent ethnicity is Hispanic, with a large portion of White and Asian/Filipino/Pacific Islander students, too.
8. Who might want to know the information in Figure 2.4.1?
    1. California community college students? Potential California community college students? Faculty or administrators of California community colleges?

Exercise 2.4.1

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 available at the end of this section.

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.

Answers

Exercise 2.4.1

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.

This page titled 2.4: 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.

2.5: 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 number of responses in the category. Instead of showing frequencies, a pie chart shows *proportions*. Figure 2.5.1 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.

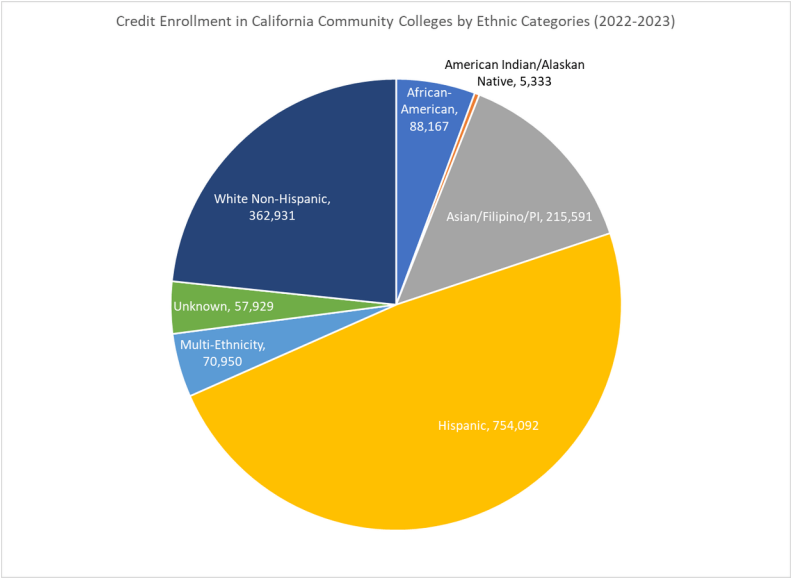


Figure 2.5.1- Pie Chart of Ethnic Categories (Copyright CC0; chart created by Michelle Oja via California Community Colleges Chancellor's Office DataMart)

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. 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 2.5.1?
    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 2.5.1?
    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 2.5.1?
    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 2.5.1? What pops out to you?
    1. The pie chart shows which ethnic groups have the most students attending California community colleges; almost half of students attending credit classes identified as Hispanic in 2022-2023. The next largest group of students are White, and the third largest group being Asian, Filipino, or Pacific Islander.
    2. What does Figure 2.5.1 make you wonder about?
        1. Why did almost 60,000 students not identify their ethnicity to their college?
5. What is a catchy headline for Figure 2.5.1?
    1. CCC are De Colores

6. How could you summarize the info in Figure 2.5.1 into one sentence?
    1. While Hispanic students are almost half of students attending credit courses at California community colleges in 2022-2023, there is also a large portion of White and Asian/Filipino/Pacific Islander students and very few Native American students.
7. Who might want to know the information in Figure 2.5.1?
    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 2.5: 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.

2.6: 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 2.6.1. 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.

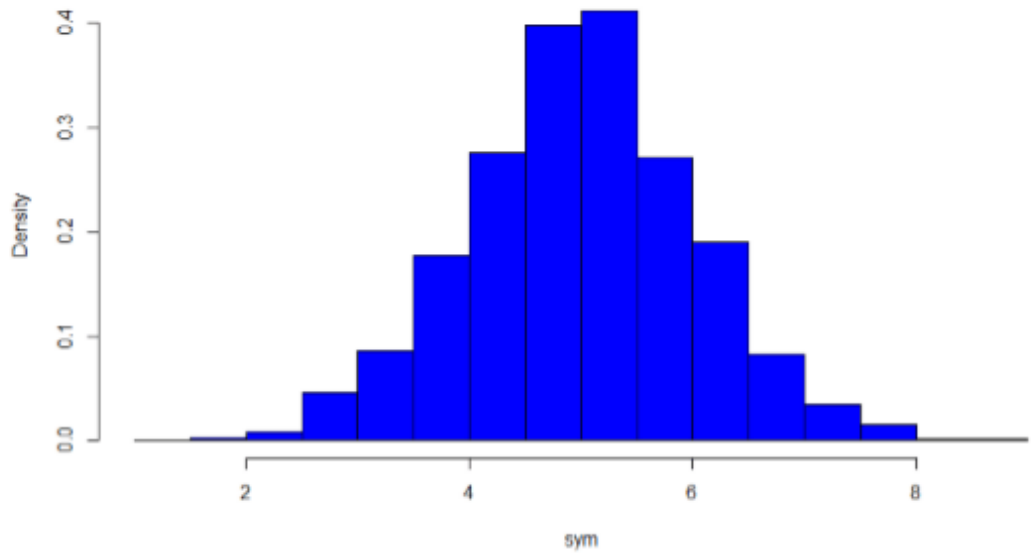


Figure 2.6.1: A symmetrical distribution in a histogram (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Symmetrical distributions can also have multiple peaks. Figure 2.6.2 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.

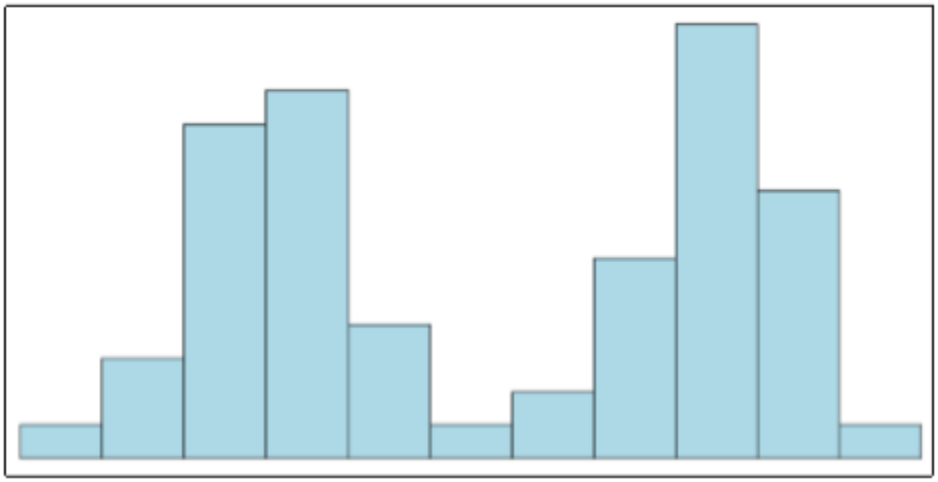


Figure 2.6.2: A bimodal distribution in a histogram (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 2.6.3 and 2.6.4 show positive (right) and negative (left) skew, respectively.

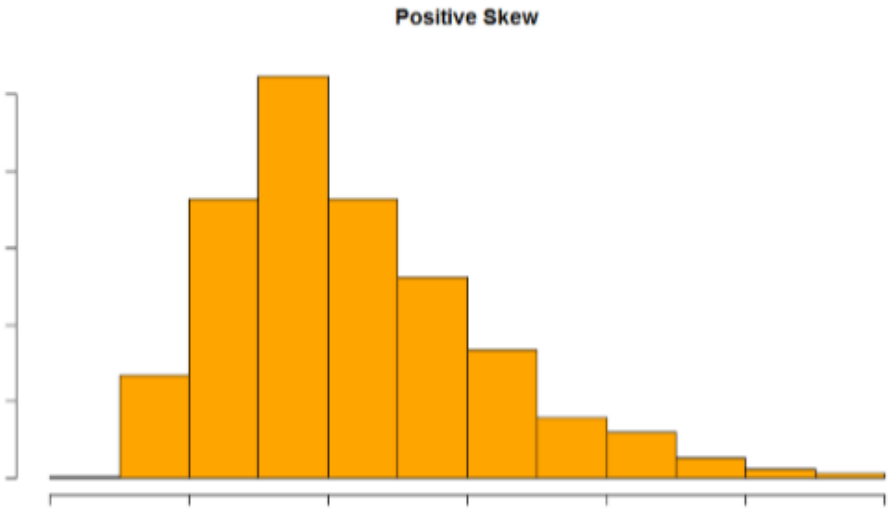


Figure 2.6.3: A positively skewed distribution in a histogram (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

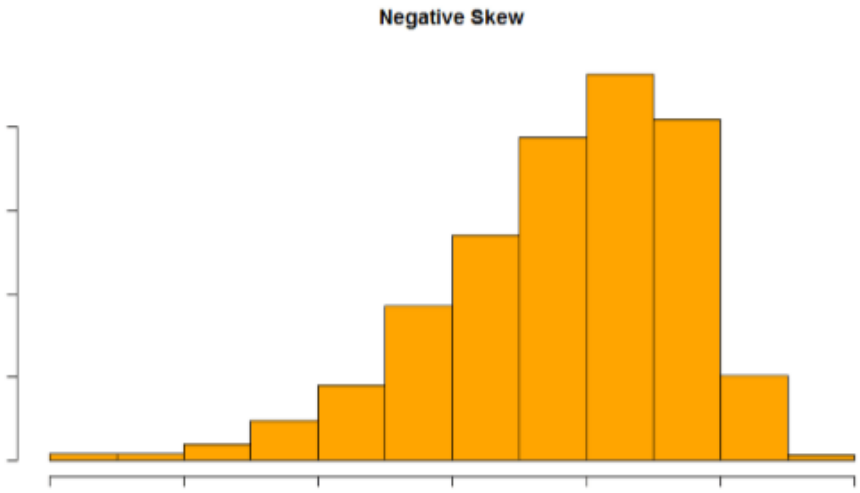


Figure 2.6.4: A negatively skewed distribution in a histogram (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 2.6: Graphing Quantitative Variables 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.

2.7: 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, and this blog from Medium (Dugar, 2018) also provides a short introduction to the words that describe the shape of a distribution: skew and kurtosis.

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 2.7.1 shows examples of a positively skewed line graph (on the right) and a negatively skewed line graph on the left.

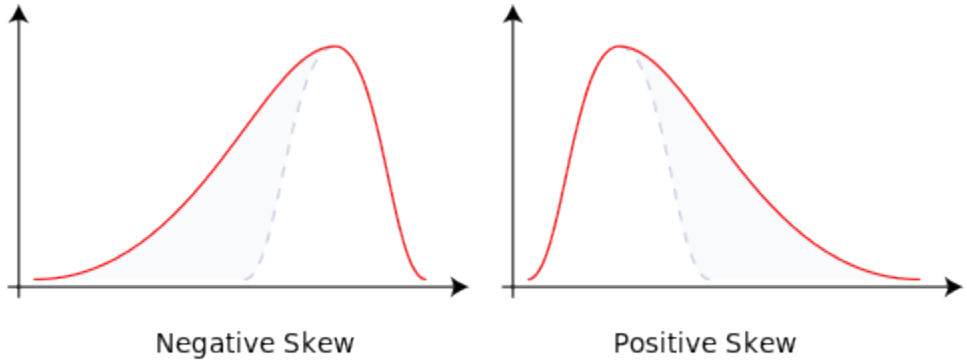


Figure 2.7.1- Diagram of Skew. (CC-BY-SA Rodolfo Hermans (Godot), CC BY-SA 3.0, via Wikimedia Commons)

Skewness is basically a measure of asymmetry, and the easiest way to explain it is by drawing some pictures. As Figure 2.7.2 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.

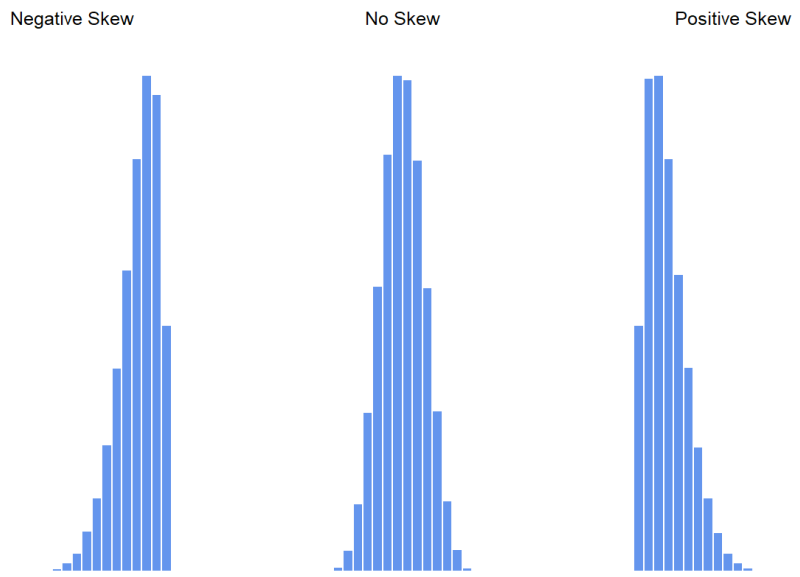


Figure 2.7.2- Diagram of Skew and No Skew (CC-BY-SA Danielle Navarro from Learning Statistics with R)

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 2.7.3, 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.

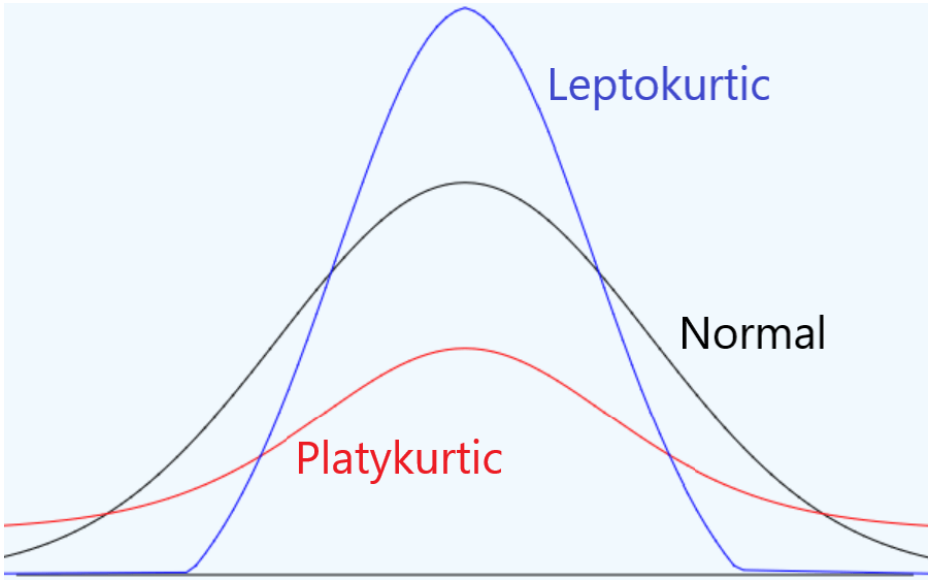


Figure 2.7.3- Diagram of Three Kurtosis Types in Line Graphs (CC-BY Larry Green)

Figure 2.7.4 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.

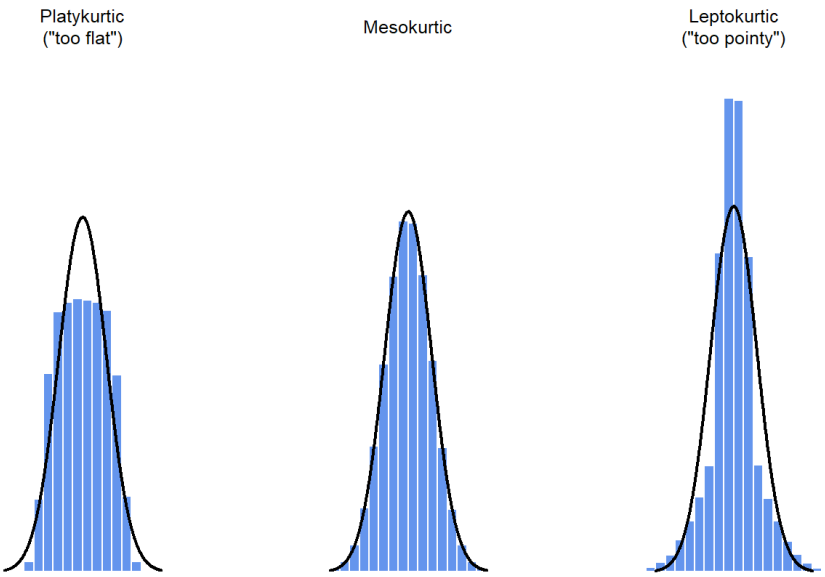


Figure 2.7.4: Diagram of Three Kurtosis Types in Histograms (CC-BY-SA Danielle Navarro from Learning Statistics with R)

The types of kurtosis are summarized in Table 2.7.1.

Table 2.7.1- *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 2.7.1 Image Attribution: Rodolfo Hermans (Godot), CC BY-SA 3.0, via Wikimedia Commons
- Figure 2.7.3 Image Attribution: CC-BY Larry Green from Lake Tahoe Community College

This page titled 2.7: Skew and Kurtosis is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Danielle Navarro.

2.8: 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 2.8.1 shows the same information but as a graph:

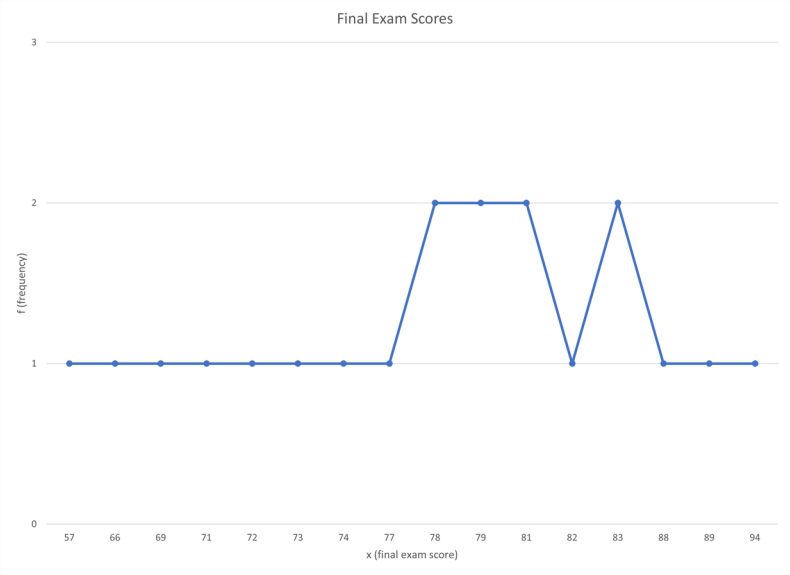


Figure 2.8.1: Line Graph of Frequency of Final Exam Scores. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

Let's unpack Figure 2.8.1, 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 2.8.1?
    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 2.8.1?
    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 2.8.1?
    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. However, the y-axis in Figure 2.8.2 shows the average percentages earned out of all points possible in the class each

semester.

4. [New question!] 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 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 2.8.1? 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 2.8.1 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 2.8.1?
    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 2.8.1 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 2.8.1?
    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 2.8.2 shows the percentage of points earn out of all points in Dr. MO's behavioral statistics courses in six consecutive semesters (starting in Fall 2013).

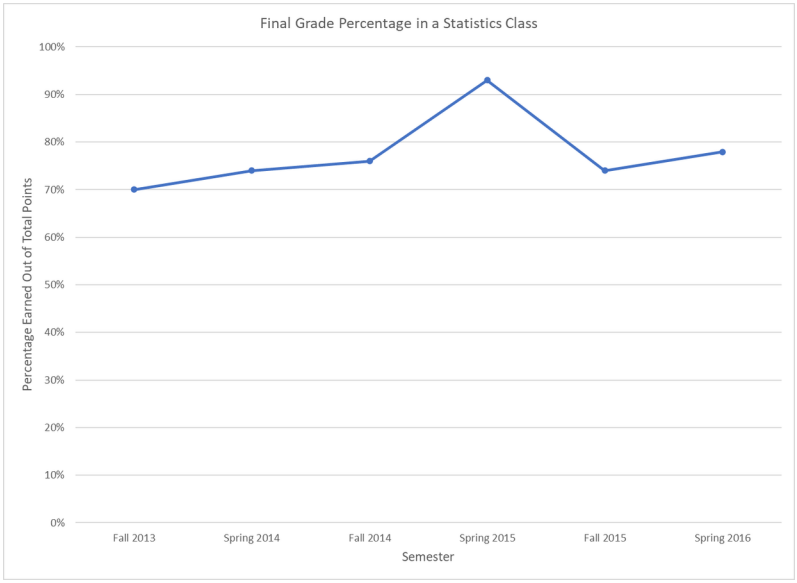


Figure 2.8.2- Line Graph of Percentages Through Time (Copyright CC-BY-SA; Michelle Oja)

Unpacking Figure 2.8.2, what do you see? First, let's compare Figure 2.8.2 to Figure 2.8.1. In Figure 2.8.1, the x-axis was each score, while in Figure 2.8.2 the x-axis is the time period. Similarly, in Figure 2.8.1, the y-axis in the frequency, the number of students who earned that score. In Figure 2.8.2, 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 2.8.2?
    - 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 2.8.2?
  - 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 2.8.2?
  - 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 2.8.2 shows the average percentages earned out of all points possible in the class each semester.
- Is Figure 2.8.2 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 2.8.2? What pops out to you?
-
  - Spring 2015 was a great year for students!
- What does Figure 2.8.2 make you wonder about?
  - What happened in Spring 2015?! And why didn't it keep happening?
- What is a catchy headline for Figure 2.8.2?
  - Average is Passing in Dr. MO's Behavioral Statistics Class
- How could you summarize the info in Figure 2.8.2 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 2.8.2?
  - 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 2.8: 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.

2.9: 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 2.9.1

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

Answer available at the end of this section.

Answer available at the end of this section.

Exercise 2.9.1

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

Answer available at the end of this section.

Figure 2.9.1 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.

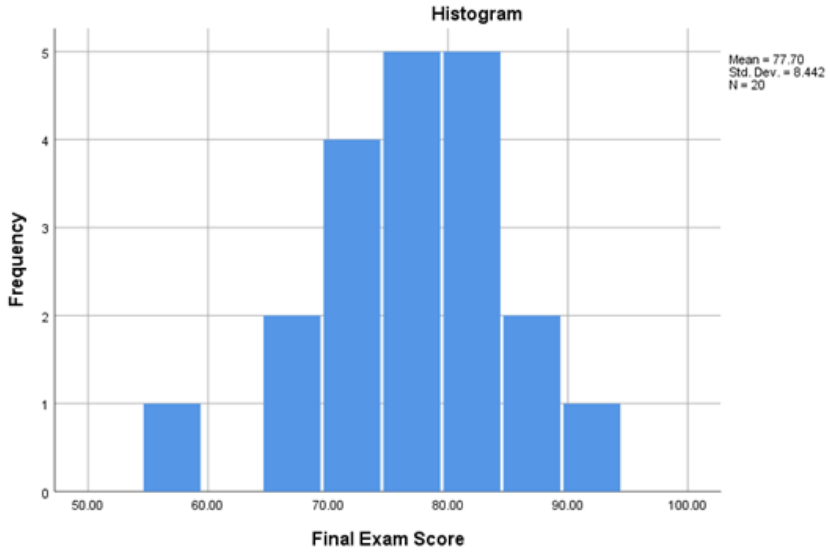


Figure 2.9.1: Histogram of Frequency of Final Exam Scores. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

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 2.9.1 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 2.9.1, 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.

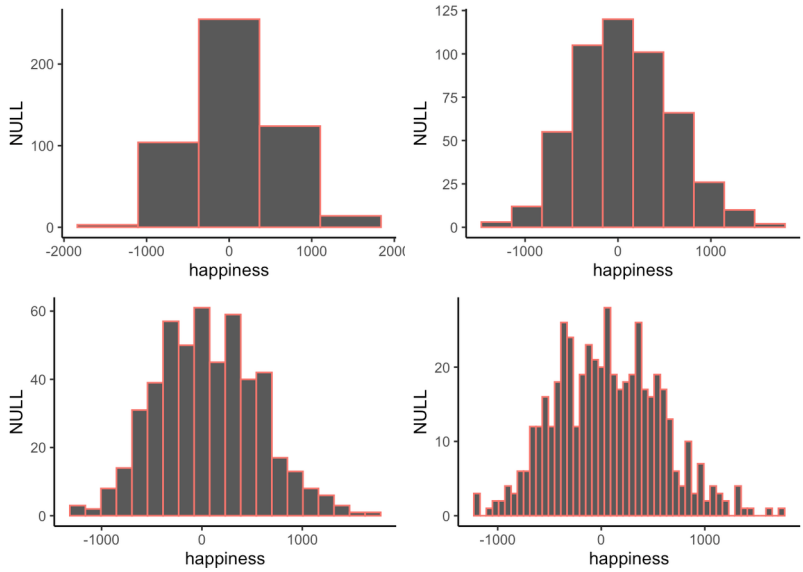


Figure 2.9.2: Four histograms of the same data using different bin widths. (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

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.

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 2.9.1?
    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 2.9.1?
    1. The x-axis on histograms show categories based on each score. In Figure 2.9.1, the software created 10-point categories for each Final Exam Score that a student earned.
3. What does the y-axis measure in Figure 2.9.1?
    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 2.9.1, that means that the y-axis shows how many students earned scores on the Final Exam in each 10-point range.

4. Is Figure 2.9.1 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 2.9.1 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 2.9.1? 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 2.9.1 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 2.9.1?
    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 2.9.1 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 2.9.1?
    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

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).

Contributors and Attributions

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

Answers

Exercise 2.9.1

Exercise 2.9.1

Line graph (sometimes called a frequency polygon)

Exercise 2.9.1

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.

This page titled 2.9: 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/.

2.10: Graphing Quantitative Data- Means Plot

New type of graph alert! Means plots are not like the other types of graphs that we've been learning. Because means are quantitative, a means plot is a type of graph showing quantitative data. This could be confusing, though, because the groups are qualitative variables.

Exercise 2.10.1

If a means plot shows the means of different groups in a graph, how are they different from the other graphs that we've learned about?

Answer available at the end of this section.

Figure 2.10.1 shows a means plot of GPA for Duke University students by their major. Notice that this is not showing the frequency of academic majors, nor is it showig the frequency of individual GPAs. Instead, it is showing the mean GPA for the three categories of academic majors.

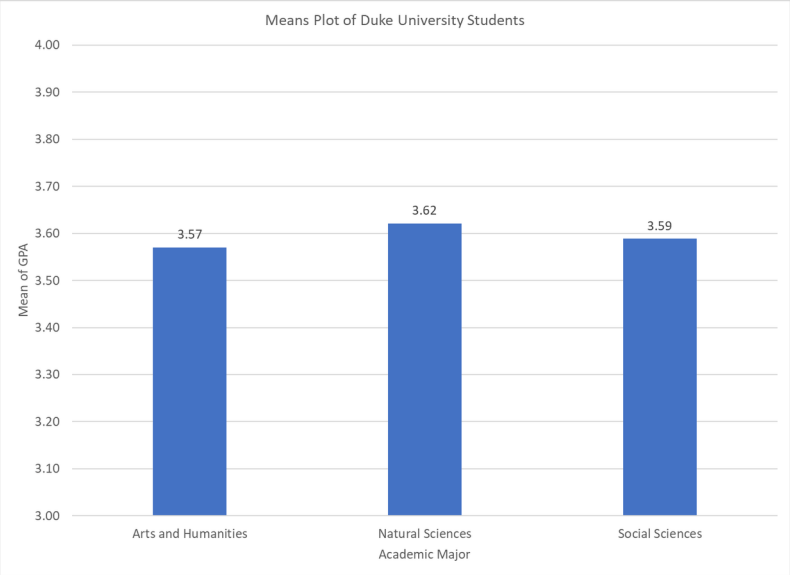


Figure 2.10.1: Means Plot of Average GPA by Academic Majors. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

Like with a frequency bar chart for qualitative variables, the x-axis show the qualitative variable's categories. In Figure 2.10.1, that is academic major. While these are in alphabetical order (A for Arts, N for Natural, S for Social), they could be ordered in any way that you'd like since the categories are qualitative variables and don't have a natural order. Different from a frequency bar chart, the y-axis of a means plot shows the average of the quantitative variable. For Figure 2.10.1, that is the mean GPA for each category of academic major.

Sometimes you'll see the means plotted like a line graph, as shown in Figure 2.10.2. This is the exact same data as in Figure 2.10.1 , just displayed as dots that are connected by a line.

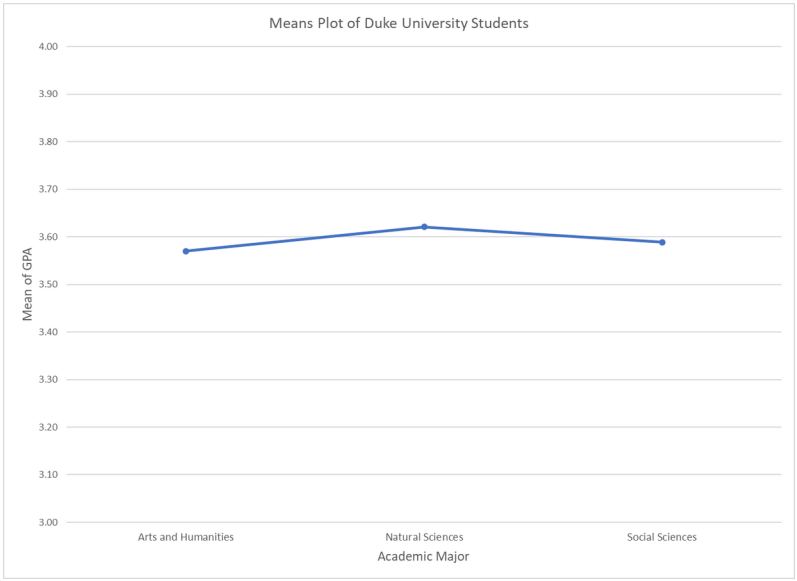


Figure 2.10.2: Means Plot of Average GPA by Academic Majors. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

Even though Figure 2.10.2 looks like a line graph or frequency polygon, it does not make sense to try to describe the shape of the distribution (kurtosis) or skew since this isn't a distribution of all of the data. However, we can still interpret what the graph is telling us. Both Figure 2.10.1 and Figure 2.10.2 show that the average GPA is pretty similar between the types of majors, but that the average GPA of Natural Science majors might be a little higher.

Means Plot Interpretation

Most, but not all, of the questions that we've been answering to help interpret graphs make sense for a means plot since we are only looking at group means, not frequencies or the shape of a distribution.

1. What kind of graph is Figure 2.10.1 or Figure 2.10.2?
    1. Both of these graphs are means plots, which show the average score on a quantitative variable for qualitative groups.
2. What does the x-axis measure in Figure 2.10.1 or Figure 2.10.2?
3. 1. The x-axis on means plots show categories of the qualitative variable. In these charts, the qualitative categories are based on academic major.
4. What does the y-axis measure in Figure 2.10.1 or Figure 2.10.2?
    1. The y-axis is the axis that goes up and down. Unlike most charts discussed in this textbook, the y-axis in means plots is the average score on the quantitative variable. For our data, this is GPA.
5. ~~Is Figure 2.10.1 or or Figure 2.10.2 skewed? If so positively or negatively? If not, is the graph tall/narrow, medium/normal, or wide/flat?~~
    1. This questions doesn't make sense for means plots since the x-axis categories could be plotted in any order.
6. What do you notice from Figure 2.10.1? What pops out to you?
    1. I notice that the means look pretty similar based on the scale of the y-axis (from a GPA of 3.00 to 4.00), but that Natural Sciences does look a little higher than the other two groups of majors.
7. What does Figure 2.10.1 or Figure 2.10.2 make you wonder about?
    1. This means plot makes me wonder why Natural Sciences have a little higher GPA. Are the instructors better in those fields? Do the students have more study skills?
8. What is a catchy headline for Figure 2.10.1 or Figure 2.10.2?
    1. Major isn't Major (because academic major doesn't affect GPA)
9. How could you summarize the info in Figure 2.10.1 or Figure 2.10.2into one sentence?
    1. GPA is pretty similar between academic majors, although Natural Science majors might have slightly higher average GPAs.
10. Who might want to know the information in Figure 2.10.1 or Figure 2.10.2?
    1. I imagine that Duke University students would be interested in this information, as well as Duke University administrators. As a professor, Dr. MO would also be interested in this! She might work even harder to help her students to try to have the Social Sciences have the highest average GPA.

Summary

Means plots aren't complex but they feel tricky because they look like bar charts or line graphs, but show very different information (means, instead of frequencies). We won't talk about them for awhile, but keep them in mind when we start comparing multiple means in Unit 3 (More Mean Differences).

Figure 2.10.1: The surprisingly logical minds of babies TED Talk

This TED Talk (Shulz, 2015) shows means plots in actions, and previews how we'll use probability in analyzing our data. It's about babies!

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Figure 2.10.1: Laura Shulz's TED Talk about research on how babies think. (CC-BY-NC-ND TED Talks)

We have a few more types of graphs for quantitative variables to learn about. Let's go!

Contributors

- Dr. MO (Taft College)

Answers

Exercise 2.10.1

Your answer should mention that most of the graphs that we've been learning are frequency graphs, meaning that they show the frequency of each score. However, a means plot only shows the means of different groups.

This page titled 2.10: Graphing Quantitative Data- Means Plot 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/.

2.11: Graphing Quantitative Data- Boxplots

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 2.11.1.

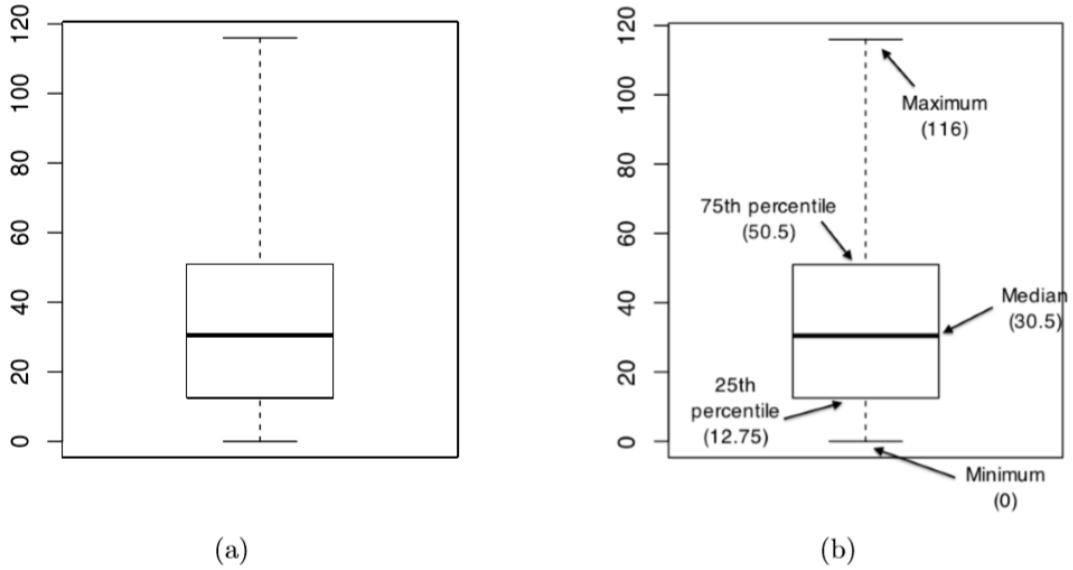


Figure 2.11.1- A basic boxplot (panel a), plus the same plot with annotations added to explain what aspect of the data set each part of the boxplot corresponds to (panel b). (CC-BY-SA- Danielle Navarro from Learning Statistics with R)

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 2.11.2, 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).

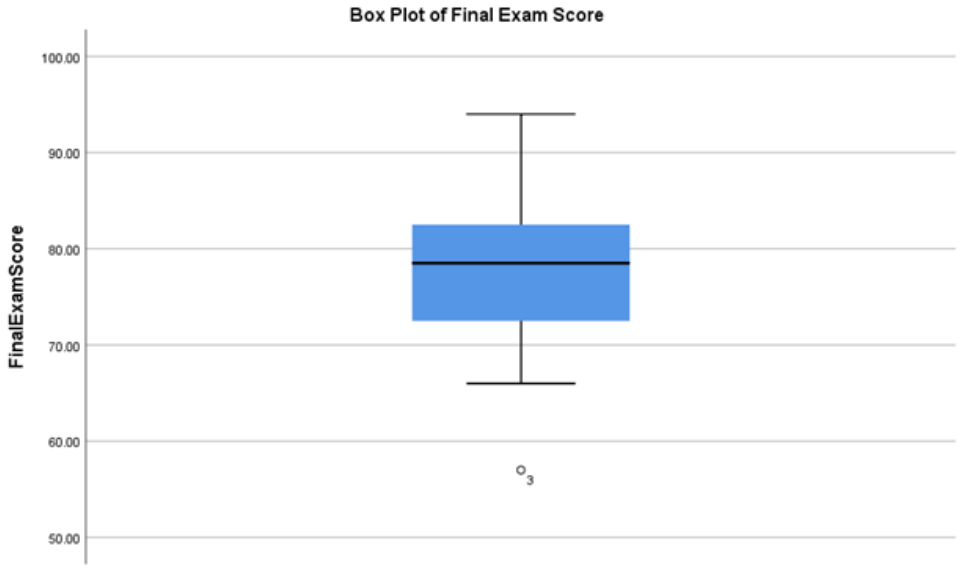


Figure 2.11.2- Boxplot of Frequency of Final Exam Scores. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

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 2.11.2.

1. What kind of graph is Figure Figure 2.11.2?
    1. This box plot.
2. What does the x-axis measure in Figure 2.11.2?
    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 2.11.2?
    1. The y-axis is the axis that goes up and down. In Figure 2.11.2,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 2.11.2? 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 2.11.2? 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 2.11.2 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 2.11.2?
    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 2.11.2 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 2.11.2?
    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.

Historical Context

According to Wikipedia, Mary Eleanor Hunt Spear (March 4, 1897 – January 22, 1986) was an American data visualization specialist and author, known in part for devising an early version of the box plot.



Figure 2.11.1: Mary Eleanor Hunt Spear from Wikipedia, Fair Use

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

2.12: 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 two 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 2.12.1 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?

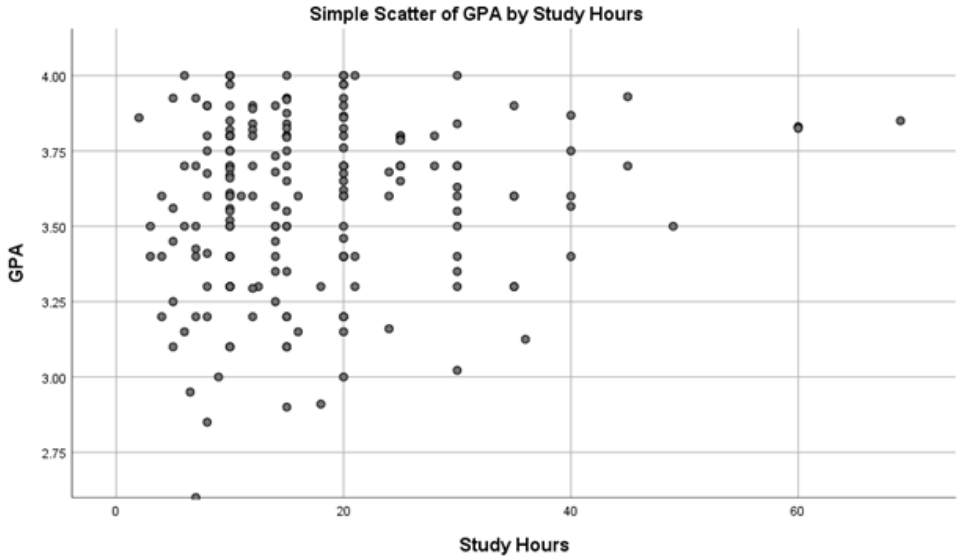


Figure 2.12.1: Scatterplot of Study Hours and GPA. (Copyright CC-SA, chart by Dr. MO via data provided by OpenIntro.org)

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 2.12.1, 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 2.12.1?
    1. This is a scatterplot. Scatterplots are used to plot two quantitative variables for each person. In Figure 2.12.1, the two variables are Study Hours and GPA.
2. What does the x-axis measure in Figure 2.12.1?
    1. The x-axis in Figure 2.12.1 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 2.12.1?

    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 2.12.1 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 2.12.1 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 2.12.1? 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 2.12.1 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 2.12.1?
    1. Study Less and Still Get a 4.0!
8. How could you summarize the info in Figure 2.12.1 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 2.12.1?
    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 2.12.1

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 available at the end of this section.

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 2.12.2 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.

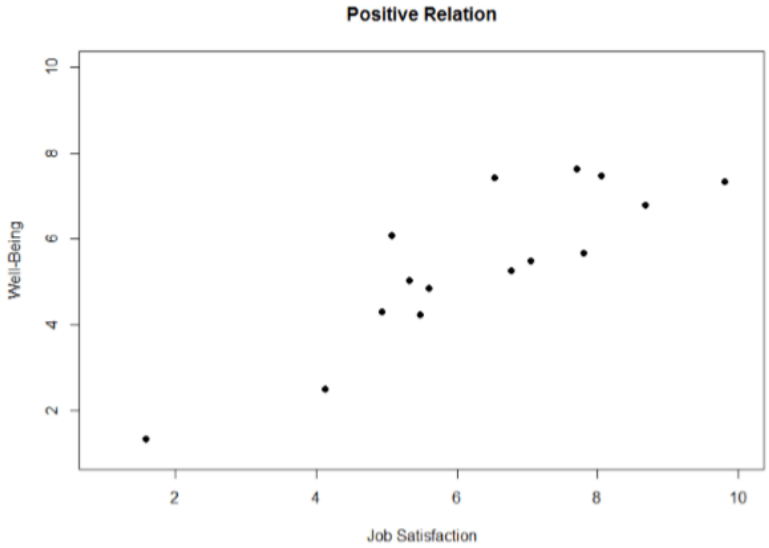


Figure 2.12.2- Scatterplot of Job Satifaction & Well-Being (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Figure 2.12.2 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 2.12.3 shows a negative relation between Job Satisfacation on the x-axis and Burnout on the y-axis. As we can see from Figure 2.12.3, 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 2.12.2, 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.

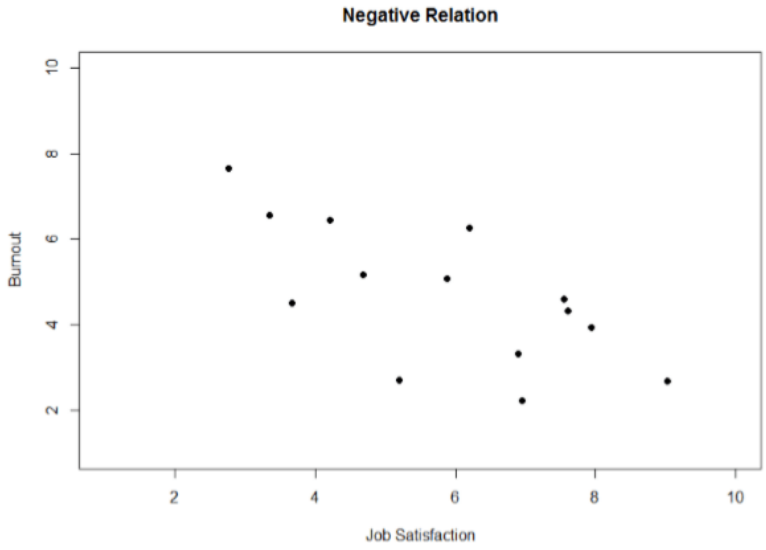


Figure 2.12.3- Scatterplot of Job Satisfaction and Burnout (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 2.12.4 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.

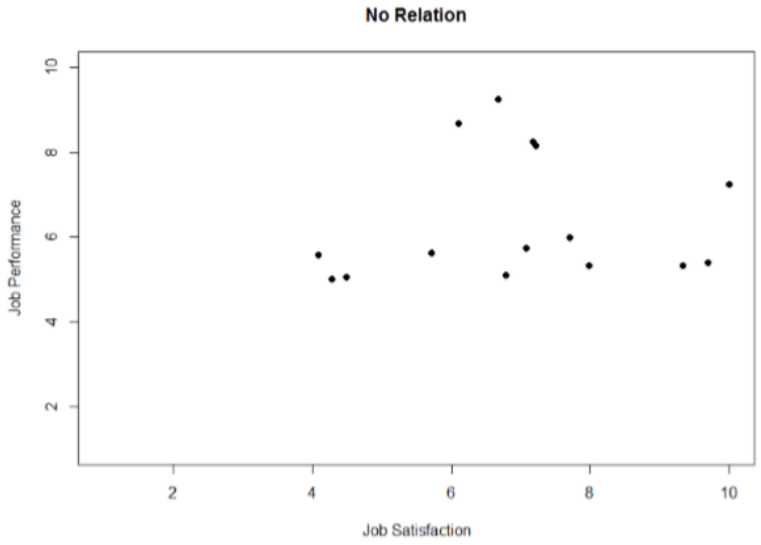


Figure 2.12.4- Scatterplot of Job Satisfaction and Job Performance (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Curvilinear Relationships

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

A linear relation is what we saw in Figure 2.12.2 and Figure 2.12.3. 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 2.12.5 and 2.12.6.

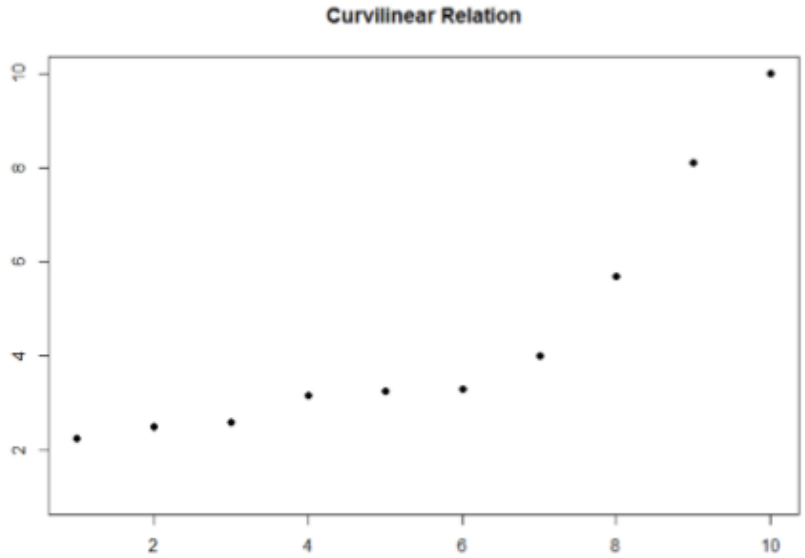


Figure 2.12.5: Curvilinear Scatterplot with an Elbow (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

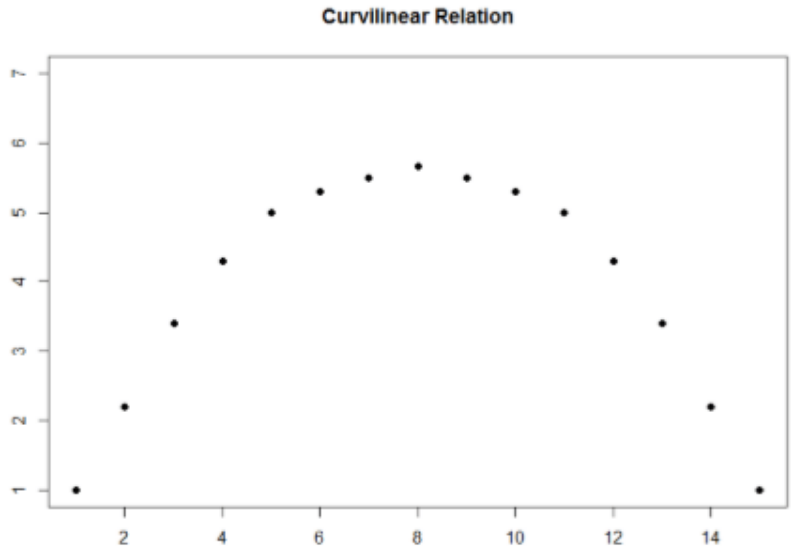


Figure 2.12.6: Curvilinear Scatterplot with an Inverted-U (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics).

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)

Answers

Exercise 2.12.1

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?

This page titled 2.12: Graphing Quantitative Data- Scatterplots is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

2.13: 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 2.13.1 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.

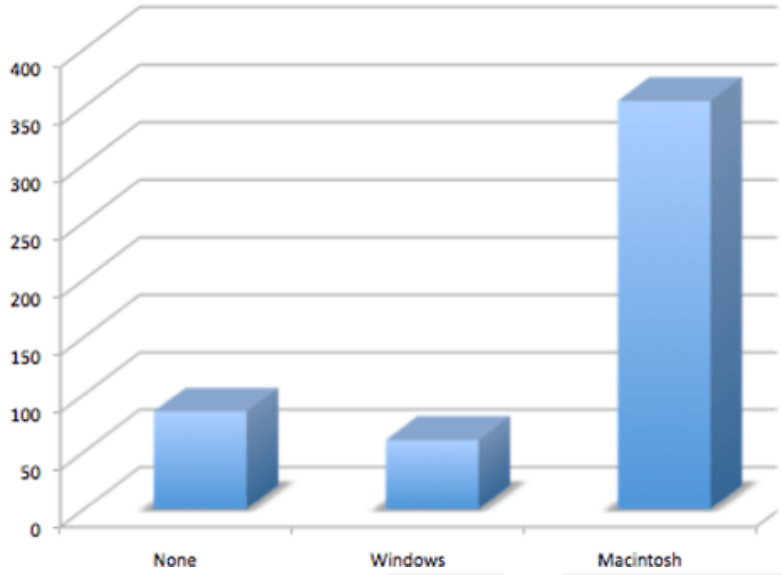


Figure 2.13.1: A three-dimensional bar chart (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure 2.13.2 presents iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure 2.13.2 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 2.13.2 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 \({2{.}13{.}\eatSpaces{2\)! 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.

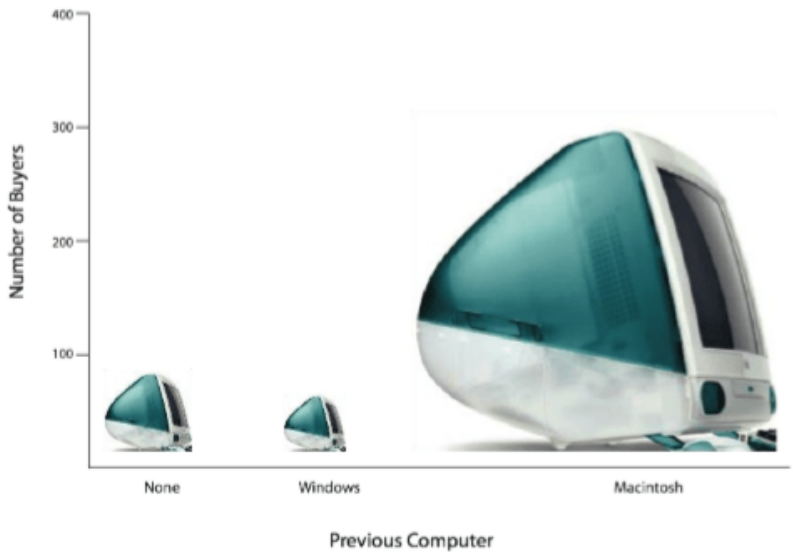


Figure Extra close brace or missing open brace : Example with a lie factor greater than 8. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 2.13.3 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%.

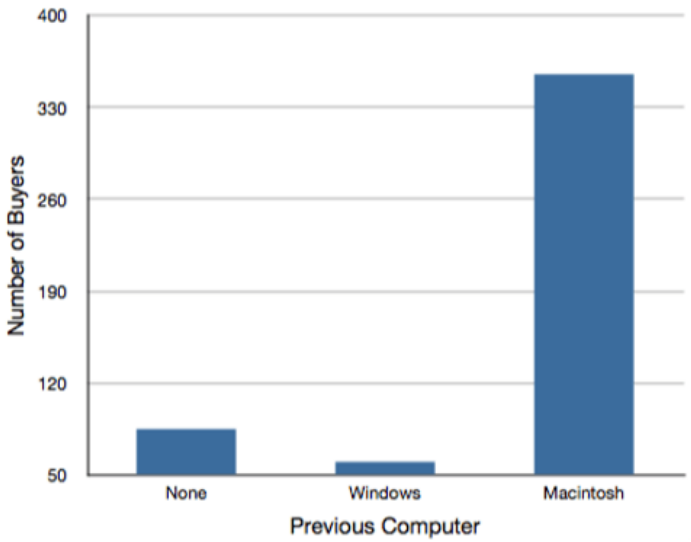


Figure 2.13.3: Example with a baseline of 50. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 2.13.1

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 2.13.1

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 2.13: Summary and Some Honesty 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.

2.14: APA Style Charts

APA Style, as described in the Publication Manual of the American Psychological Association (Seventh Edition), has specific ways to format the title of tables and graphs. We already talked about tables, now it's time to talk about graphs.

Graphs and pictures are considered figures (not tables). In APA Style, figure labels should be numbered and have a descriptive title. “Figure 1” should be in bold, and on the line above the title (which should be in italics). They both should be above the figure, flush-left. An example is below, and you can find more details at the best site to learn about APA Style: Purdue Online Writing Lab. Here is a link to OWL Purdue’s page on tables and figures in APA Style.

Example:

Note

Figure 1

Final Exam Scores

Figures should be near where they are being discussed in the appear or in an appendix (which would be the last section of the paper). Finally, make sure to always label your *axes* with a title and with the numbers/categories, not just the figure title.

That's it for graphs! I'll see you in the next chapter to talk about how to describe these distributions of data with just a few numbers!

This page titled 2.14: APA Style Charts is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

3: MEASURES OF CENTRAL TENDENCY

What can we learn about the center of a distribution from just one or a few numbers?

3.1: Introduction to Measures of Central Tendency

3.2: Math Refresher

3.2.1: Decimals

3.3: Mode

3.4: Median

3.5: Mean

3.6: Interpreting All Three

3.7: Summary of Measures of Central Tendency

This page titled 3: Measures of Central Tendency is shared under a license and was authored, remixed, and/or curated by Michelle Oja.

3.1: Introduction to Measures of Central Tendency

We've seen that we can get a sense of data by plotting frequency data on a graph. These frequency charts show us what the whole distribution of scores looks like, approximately how big and small the scores are, and how similar and different the scores are from another. In addition to summarizing distributions of scores with graphs, we can summarize a whole set of scores using just a few numbers.

From many numbers to one

Measures of central have one important summary goal: to reduce a pile of numbers to a single number (or a few) that can be used to describe that distribution of scores. We already know that looking at lots of numbers is hopeless. If your boss ever asks the frightening question, "hey, what are all these numbers like?". You can find a number that represents the center of the distribution of scores to answer your boss's questions. Each of the following sections will look at a different way to define the center of a distribution: the mode, the median, and then the mean. Each of these are most useful in different situations, but they can be used together to predict what the frequency distribution could be shaped like (if you couldn't see a frequency line graph or histogram).

But before we get into learning about these measures of central tendency, let's first go back to our earlier math classes to remember what the symbols mean and other mathematical operations.

This page titled 3.1: Introduction to Measures of Central Tendency 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.

- 2.4: Measures of Central Tendency (Sameness) by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

3.2: 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, and the next section will discuss rounding decimal points and scientific notation. You learned much of this in junior high or high school, but that was a long time ago for most of us!

Order of Operations

Example 3.2.1

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.

Historical Context

The earliest evidence of written mathematics dates back to the ancient Sumerians, who built the earliest civilization in Mesopotamia, from 3000 BCE. From around 2500 BCE onward, the Sumerians wrote multiplication tables on clay tablets and dealt with geometrical exercises and division problems. Knowledge of Babylonian mathematics is derived from more than 400 clay tablets that were inscribed whilst the clay was moist, and baked hard in an oven or by the heat of the sun; some of these appear to be graded homework (Wikipedia)!



Figure 3.2.1: Babylonian mathematical tablet Plimpton 322 by author unknown from University of British Columbia. Public Domain

Notation & Symbols

Table 3.2.1 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 3.2.1- 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 3.2.2:

Table 3.2.2- 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

$$(\sum X)^2 = 4.6^2 + 5.1^2 + 4.9^2 + 4.4^2$$
$$= 21.16 + 26.01 + 24.01 + 19.36 = 90.54$$

Notice that:

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

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 3.2.2- Table of Scores and Products

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

Practice with Equations

Let's practice some simple equations to see how the order of operations and the symbols actually work.

If you were given this formula:

$$\frac{\sum X}{N}$$

You would need to figure out what

$$\sum$$

means, and then decide if you would do what

$$\sum$$

is telling you to do first or after the division.

Since

$$\sum$$

means to sum, you would add up each and very number in your set. PEMDAS says that you'd do Division before Addition, but there's nothing to divide until you add everything up! Let's practice with the numbers in each column of Table 3.2.2.

$$\frac{\sum X}{N}$$

For the X column,

1. Adding them all up is: $\sum = 1 + 2 + 3 = 6$
2. Then divide by 3 since there are three numbers in the X column.

$$\frac{\sum = 1 + 2 + 3 = 6}{3} = \frac{6}{3} = 2$$

Example 3.2.2

How would we apply this information on the numbers in the Y column?

Solution

$$\frac{\sum = 3+2+7 = 12}{3} = \frac{12}{3} = 4$$

You can try the last column on your own!

Exercise 3.2.1

How would we apply this information on the numbers in the XY column?

Answer available at the end of this section.

The next section will refresh you on rounding decimals and scientific notation. This section is adapted from the Support Course for Elementary Statistics, if you'd like more details or help in other areas!

Answers

Exercise 3.2.1

$$\frac{\sum = 3+4+21 = 28}{3} = \frac{28}{3} = 9.3\overline{3}$$

The line over the second 3 is telling you that that number keeps repeating indefinitely. Check out the next section to learn how to round decimals numbers.

This page titled 3.2: Math Refresher is shared under a CC BY-NC-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.

3.2.1: Decimals

This section is adapted from the Support Course for Elementary Statistics. That support course textbook is another resource to help refresh your math memory. In most statistics applications that you will encounter, the numbers will not come out evenly, and you will need to round the decimal.

In this section, we will go over how to round decimals to the nearest whole number, nearest tenth, nearest hundredth, etc. For example, suppose that you used a calculator to find the probability that a randomly selected day in July will have a high temperature of over 90 degrees. Your calculator gives the answer: 0.4987230156. This is far too many digits for practical use, so it makes sense to round to just a few digits. By the end of this section you will be able to perform the rounding that is necessary to make unmanageable numbers manageable.

We will also look at how to read scientific notation. According to Dr. Green, a very common error that statistics students make is not noticing that the calculator is giving an answer in scientific notation.

Brief Review of Decimal Language

Consider the decimal number: 62.5739. There is a defined way to refer to each of the digits.

- The digit 6 is in the "Tens Place"
- The digit 2 is in the "Ones Place"
- The digit 5 is in the "Tenths Place"
- The digit 7 is in the "Hundredths Place"
- The digit 3 is in the "Thousandths Place"
- The digit 9 is in the "Ten-thousandths Place"
- We also say that 62 is the "Whole Number" part.



Note

It is reported that the mean number of classes that college students take each semester is 3.2541. Then the digit in the *hundredths place is 5*.

Keeping this example in mind will help you when you are asked to round to a specific place value. In this course, we will almost always use the hundredths or thousandths decimal places, meaning that we'll include two or three numbers after the decimal point. In real life, you rarely need to include any decimal points, however. Sometimes, that kind of precision is just distracting.

Rules of Rounding

When asked to round to a specified place value, the answer will erase all the digits after the specified digit. The process to deal with the other digits is best shown by examples.

Example 3.2.1.1: Case 1 - The Test Digit is Less Than 5

Let's say that the your professor spent 3.741 hours preparing for this week's class. What would we report if we wanted to round the hours spent to the nearest tenth?

Solution



Since the test digit (4) is less than 5, we just erase everything to the right of the tenths digit, 7. The answer is: The time that Dr. MO spent preparing for this week's class was 3.7 hours.

Example 3.2.1.2: Case 2 - The Test Digit is 5 or Greater

Most community college students do not attend "full-time," meaning 12 or more units. Imagine that, on average, your class had registered for 8.53792 units this semester. What would that be rounded to the nearest hundredth?

Solution



Since the test digit (6) is 5 or greater, we add one to the hundredths digit and erase everything to the right of the hundredths digit, 3. Thus the 3 becomes a 4. The answer is: My class registered for an average of 8.54 units this semester.

Let's try one on your own.

Exercise 3.2.1.1

The average grade percentage in Dr. MO's last behavioral statistics course was 93.8922%.

1. What would that look like with no decimal points?
2. What would that look like with one decimal point?
3. What would that look like with two decimal points?

Hint: Your answers should be a complete sentences, not just a number.

Answer available at the end of this section.

FMI, Dr. Green has provided a video (link to YouTube video on rounding) that shows an example of rounding and how you could use the names of the different decimal spaces.

How Much Rounding?

Rounding is used in most areas of statistics, since the calculator or computer will produce numerical answers with far more digits than are useful. Dr. MO wants most answers rounded to the hundredths, but ask your professor how many decimals they would like you to keep. In real life, choose the smallest number of decimals to keep so that no important information is lost. For example suppose you conducted a sample to find the proportion of college students who receive financial aid and the calculator presented 0.568429314. You could turn this into a percent at 56.8429314%. Keeping this many decimal places is not useful, and is confusing. If you wanted to present this finding to the student government, you might want to round to the nearest whole number (including no numbers after the decimal). In this case the ones digit is 6 and the test digit is 8. Since 8 ≥ 5, you add 1 to the ones digit. You can tell the student government that 57% of all college students receive financial aid.

Example 3.2.1.3

Suppose that you found out that the probability that a randomly selected person with who has misused prescription opioids will transition to heroin is 0.04998713. Round this number to four decimal places.

Solution

The first four decimal places are 0.0499 and the test digit is 8. Since 8 ≥ 5, we would like to add 1 to the fourth digit. Since this is a 9, we go to the next digit to the left. This is also a 9, so we go to the next one which is a 4. We can think of adding 0499 + 1 = 0500. Thus the answer is 0.0500. Note that we keep the last two 0's after the 5 to emphasize that this is accurate to the fourth decimal place.

Rounding and Equations

Many times, we have to do arithmetic on numbers with several decimal places and want the answer rounded to a smaller number of decimal places. One question you might ask is should you round before you perform the arithmetic or after. For the most accurate result, you should always round **after** you preform the arithmetic if possible.

Example 3.2.1.4

Suppose you pick three cards from a 52 card deck with replacement and want to find the probability of the event, A, that none of the three cards will be a 2 through 7 of hearts. This probability is:

$$P(A) = (0.8846)^3$$

Round the answer to 2 decimal places.

Solution

Note that we have to first perform the arithmetic. With a computer or calculator we get:

$$0.8846^3 = 0.69221467973$$

Now we round to two decimal places. Notice that the hundredths digit is a 9 and the test digit is a 2. Thus the 9 remains unchanged and everything to the right of the 9 goes away. the result is

$$P(A) \approx 0.69$$

If we mistakenly rounded 0.8846 to two decimal places (0.88) and then cubed the answer we would have gotten 0.68 which is not the correct answer.

To watch Dr. Green go through this process, check out Dr. Green's example of rounding on YouTube.

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Figure 3.2.1.1: Dr. Green's Example of Rounding. (CC-BY via YouTube)

Scientific Notation

Sometimes there are SO MANY numbers after the decimal point that your calculator will show you the answer in a code. That code, is scientific notation. The standard way of writing a number in scientific notation is writing the number as a product of a number greater than or equal 1 but less than 10 followed by a power of 10. This sounds pretty complicated, but you use the little number in superscript (above the 10) to count the number of numbers to the right to figure out where to put the decimal point. For example:

$$602,000,000,000,000,000,000,000 = 6.02 \times 10^{23}$$

The main purpose of scientific notation is to allow us to write very large numbers or numbers very close to 0 without having to use so many digits. Most calculators and computers use a different notation for scientific notation, most likely because the superscript is difficult to render on a screen. For example, with a calculator:

$$0.00000032 = 3.2E-7$$

Notice that to arrive at 3.2, the decimal needed to be moved 7 places to the right.

Example 3.2.1.5

A calculator displays:

$$2.0541E6$$

Write this number in decimal form.

Solution

Notice that the number following E is 6. This means move the decimal over 6 places to the right. The first 4 moves is natural, but for the last 2 moves, there are no numbers to move the decimal place past. We can always add extra zeros after the last number to the right of the decimal place:

$$2.0541E6 = 2.054100E6$$

Now we can move the decimal place to the right 6 places to get

$$2.0541E6 = 2.054100E6 = 2,054,100$$

Example 3.2.1.6

If you use a calculator or computer to find the probability of flipping a coin 27 times and getting all heads, then it will display:

$$7.45E-9$$

Write this number in decimal form.

Solution

Many students will forget to look for the "E" and just write that the probability is 7.45, but probabilities can never be bigger than 1. You can not have a 745% chance of it occurring. Notice that the number following E is –9. Since the power is negative, this means move the decimal to the left, and in particular 9 places to the left. There is only one digit to the left of the decimal place, so we need to insert 8 zeros:

$$7.45E-9 = 000000007.45E-9$$

Now we can move the decimal place to the right 9 places to the left to get

$$7.45E-9 = 000000007.45E-9 = 0.00000000745$$

Now that you've unlocked some of those memories about how to do math, we'll start to learn about how one number can describe a whole group of numbers. Next up, our first measure of central tendency: mode!

Answers

Exercise 3.2.1.1

Your answers could be something like:

1. No decimals: The average grade percentage in Dr. MO's behavioral statistics course was 94%.
2. One decimal point: The average grade percentage in Dr. MO's behavioral statistics course was 93.9%.
3. Two decimal points: The average grade percentage of Dr. MO's behavioral statistics course was 93.89%

Which of these makes more sense to you? They are all correct, just some provide more detail than others. But it's up to you to decide if that detail is needed.

This page titled 3.2.1: Decimals is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Larry Green.

- Decimals: Rounding and Scientific Notation by Larry Green is licensed CC BY 4.0.
- 1.1: On the Psychology of Statistics by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.

3.3: Mode

The **mode** is the most frequently occurring number in your measurement. That is it. How do you find it? You have to count the number of times each number appears in your measure, then whichever one occurs the most, is the mode.

Table 3.3.1- *Mode of One*

Scores
1
1
1
2
3
4
5

The mode of the above set is 1, which occurs three times. Every other number only occurs once.

OK fine. What happens here:

Table 3.3.2- *Bi-Modal*

Scores
1
1
1
2
2
2
3
4
5
6

Hmm, now 1 and 2 both occur three times each. What do we do?

Example 3.3.1

What do we do when there is more than one mode?

Solution

When there are only a couple modes, then we list them. When there are many modes, we don't have to list out each mode.

For Table 3.3.2, the mode would be both scores of 1 and scores of 2. This is called bi-modal. When we have more than two modes, it is called multi-modal, and we don't have to list out each mode. Only list what is useful to know.

Why is the mode a measure of central tendency? Well, when we ask, “what are my numbers like”, we can say, “most of the numbers are like a 1 (or whatever the mode is)”.

Is the mode a good measure of central tendency? That depends on your numbers. For example, consider these numbers

Table 3.3.1- *Mode of One Again*

Scores
1
1
2
3
4
5
6
7
8
9

Here, the mode is 1 again, because there are two 1s, and all of the other numbers occur once. But, are most of the numbers like a 1? No, they are mostly not 1s.

“Argh, so should I or should I not use the mode? I thought this class was supposed to tell me what to do?”. There is no telling you what to do. Every time you use a tool in statistics you have to think about what you are doing and justify why what you are doing makes sense. Sorry. (Not sorry).

When might a mode be useful?

Historical Context

Attackers besieging Plataea in the Peloponnesian war in 431 BCE calculated the height of the wall by counting the number of bricks. The count was repeated several times by different soldiers. The most frequent value (the mode) was taken to be the most likely. Multiplying it by the height of one brick allowed them to calculate the length of the ladders needed to scale the walls (Champkin, 2014).



Figure 3.3.1:"*Eight bookes of the Peloponnesian Warre written by Thucydides the sonne of Olorus. Interpreted with faith and diligence immediately out of the Greeke by Thomas Hobbes secretary to ye late Earle of Deuonshire*" (Houghton Library). Public Domain

Reference

Champkin, J. (2014, January 24). *The timeline of statistics*. Significance. https://www.statsref.com/timeline.pdf

This page titled 3.3: Mode 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.

- 2.4: Measures of Central Tendency (Sameness) by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

3.4: Median

The **median** is the exact middle of the data. After all, we are asking about central tendency, so why not go to the center of the data and see where we are. What do you mean middle of the data? Let's look at these numbers:

Table 3.4.1- *List of Numbers*

Scores
1
5
4
3
6
7
9

Umm, OK. So, three is in the middle? Isn't that kind of arbitrary. Yes. Before we can compute the median, we need to order the numbers from smallest to largest.

Table 3.4.2- *List of Numbers in Order*

Scores
1
3
4
5
6
7
9

Now, five is in the middle. And, by middle we mean in the middle. There are three numbers to the higher than five, and three numbers lower than five. So, five is definitely in the middle.

OK fine, but what happens when there aren't an even number of numbers? Then the middle will be missing right? Let's see:

Table 3.4.3- *List of Six Scores*

Scores
1
2
3
4
5
6

There is no number between 3 and 4 in the data, the middle is empty. In this case, we compute the median by figuring out the

number in between 3 and 4. So, the median would be 3.5:

$$\frac{3+4}{2} = 3.5$$

This calculation is a mean, which we'll talk more about soon, and for the rest of the semseter!

Is the median a good measure of central tendency? Sure, it is often very useful. One property of the median is that it stays in the middle even when some of the other numbers get really weird. For example, consider these numbers:

Scores
1
2
3
4
4
4
5
6
6
6
7
7
1000

Most of these numbers are smallish, but the 1000 is a big old weird number, very different from the rest. The median is still 5, because it is in the middle of these ordered numbers. We can also see that five is pretty similar to most of the numbers (except for 1000). So, the median does a pretty good job of representing most of the numbers in the set, and it does so even if one or two of the numbers are very different from the others.

Finally, outlier is a term will we use to describe numbers that appear in data that are very different from the rest. 1000 is an outlier, because it lies way out there on the number line compared to the other numbers. What to do with outliers is another topic we discuss sometimes throughout this course.

This page titled 3.4: Median 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.

- 2.4: Measures of Central Tendency (Sameness) by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

3.5: Mean

The mean is also called the average. And, we're guessing you might already now what the average of a bunch of numbers is? It's the sum of the numbers, divided by the number of number right? How do we express that idea in a formula? Just like this:

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

In which $\overline{X}$ (read as "x-bar") is the *mean of the sample*. Sometimes Dr. MO adds parentheses to show that the scores (all of the X's) are summed before anything is divided because PEMDAS says that you calculate anything in parentheses first.

$$\overline{X} = \frac{(\sum X)}{N}$$

"That looks like Greek to me". Yup. The $\sum$ symbol is Greek, called sigma, and it stands for the operation of summing. We sum up *all* of the numbers, then divide the sum by N, which is the total number of numbers. Sometimes you will see $\overline{X}$ to refer to the mean of all of the numbers.

Historical Context

In 450 BCE, Hippias of Elis uses the average value of the length of a king's reign (the mean) to work out the date of the first Olympic Games, some 300 years before his time (Champkin, 2014).



Figure 3.5.1: Drawing of ancient Olympia, from Pierer's Universal Lexicon (1891). US Public Domain

In plain English, the formula looks like:

$$\text{Mean} = \frac{\text{Sum of my numbers}}{\text{Count of my numbers}}$$

"Well, why didn't you just say that?". We just did!

Let's compute the mean for these five numbers:

Table 3.5.1-*Five Scores*

Scores
2
3
6
7
9

Add 'em up: $\sum = 27 = (2 + 3 + 6 + 7 + 9)$

Divide 'em by the number of scores: $\frac{27}{5} = 5.4$

Or, to put the numbers in the formula, it looks like this:

$$\overline{X} = \frac{\sum X}{N} = \frac{3+7+9+2+6}{5} = \frac{27}{5} = 5.4$$

Is the mean a good measure of central tendency? By now, you should know: it depends.

What does the mean mean?

It is not enough to know the formula for the mean, or to be able to use the formula to compute a mean for a set of numbers. We believe in your ability to add and divide numbers. What you really need to know is what the mean really "means". This requires that you know what the mean does, and not just how to do it. Puzzled? Let's explain.

Can you answer this question: What happens when you divide a sum of numbers by the number of numbers? What are the consequences of doing this? What is the formula doing? What kind of properties does the result give us? FYI, the answer is not that we compute the mean.

OK, so what happens when you divide any number by another number? Of course, the key word here is divide. We literally carve the number up top in the numerator into pieces. How many times do we split the top number? That depends on the bottom number in the denominator. Watch:

$$\frac{12}{3} = 4$$

So, we know the answer is 4. But, what is really going on here is that we are slicing and dicing up 12 aren't we. Yes, and we slicing 12 into three parts. It turns out the size of those three parts is 4. So, now we are thinking of 12 as three different pieces $12 = 4 + 4 + 4$. I know this will be obvious, but what kind of properties do our pieces have? You mean the fours? Yup. Well, obviously they are all fours. Yes. The pieces are all the same size. They are all equal. So, division equalizes the numerator by the denominator...

"Umm, I think I learned this in elementary school, what does this have to do with the mean?". The number on top of the formula for the mean is just another numerator being divided by a denominator isn't it. In this case, the numerator is a sum of all the values in your data. What if it was the sum of all of the 500 happiness ratings? The sum of all of them would just be a single number adding up all the different ratings. If we split the sum up into equal parts representing one part for each person's happiness what would we get? We would get 500 identical and equal numbers for each person. It would be like taking all of the happiness in the world, then dividing it up equally, then to be fair, giving back the same equal amount of happiness to everyone in the world. This would make some people more happy than they were before, and some people less happy right. Of course, that's because it would be equalizing the distribution of happiness for everybody. This process of equalization by dividing something into equal parts is what the mean does. See, it's more than just a formula. It's an idea. This is just the beginning of thinking about these kinds of ideas. We will come back to this idea about the mean, and other ideas, in later chapters.

> Pro tip: The mean is the one and only number that can take the place of every number in the data, such that when you add up all the equal parts, you get back the original sum of the data.

Reference

Champkin, J. (2014, January 24). *The timeline of statistics*. Significance. https://www.statsref.com/timeline.pdf

---

This page titled 3.5: Mean is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 2.4: Measures of Central Tendency (Sameness) by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

3.6: Interpreting All Three

In a symmetrical, unimodal distribution, the mode, median, and mean are all the same number.

Note

What does symmetrical mean?

What does unimodal mean?

This means that the more symmetrical a distribution is, the closer the median, mode, and mean will be.

By knowing how similar the median, mode, and mean are, you can guesstimate at the symmetricalness of the distribution.

Example 3.6.1

What's skew? Positive skew? Negative skew?

Solution

Find these terms in the glossary, or the parts of this chapter when skew was discussed.

Without having all of the data or seeing the distribution, we can make predictions about what the distribution might look like because:

- The mean is pulled in the direction of the extreme scores (or tail of the skew),
- The mode is the highest point in the skew,
- The median is between the mean and mode.

In particular, we can predict that:

- If the mean is a lot bigger than the median, that the distribution will have a positive skew.
  - Mean > Median = positive skew
- If the mean is a lot smaller than the median, that the distribution will have a negative skew.
  - Mean < Median = negative skew

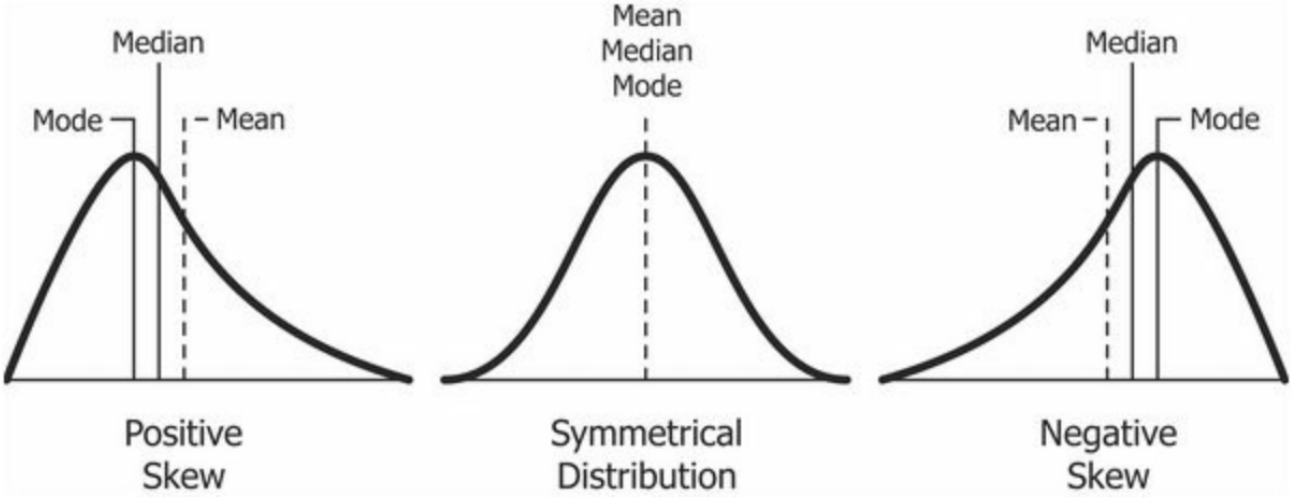


Figure 3.6.1: Mean, Median, Mode, and Skew (Copyright CC-BY-SA Diva Jain/Diva Dugar, Source)

Practice

We'll use the data in Table 3.6.1 to practice calculating and interpreting the measures of central tendency. The first column (labeled "Student") just identifies the student by an identification number. These identification numbers may not be meaningful data or they could be student names, but this column also indicates that all information in that row (left to right) is information about that same student. We will first practice calculating and interpreting the measures of central tendency together on the Final Grade Percentage column of this is real data from a section of Dr. MO's behavioral statistics class, and then you can practice calculating and interpreting the measures of central tendency yourself for the Number of Absences variable. But before we begin, why can't we interpret the three measures of central tendency for the column showing each students major?

Example 3.6.2

Calculate the measures of central tendency for student's major in Table 3.6.1.

Hint: Some of this question is a trick question?

1. Mean
2. Median
3. Mode

Solution

1. Mean: While a spreadsheet will show the sum of the student's major as zero, you should recognize that you can't add together qualitative data. Thus, you could do the math to calculate a mean, but it wouldn't make sense.

$$\overline{X} = \frac{\sum X}{N} = \frac{0}{9} = 0$$

2. Median: Similarly, to find the middle number you must have numbers in numerical order. Since Major is a qualitative variable, we can put it in alphabetical order but no numerical order.
3. Mode: We do have a mode! The most frequent answer is Education, so the mode of these student's major is Education. I did ask Excel to find the mode, though, and it said that it was "not available." Again, it's important to understand the kind of variable you have and the limitations of the software you are using!

Reviewing Example 3.6.2, you should notice that we really only can calculate one measure of central tendency for the variable of student major in Table 3.6.1 since student major is a qualitative variable. So, why can't we interpret the three measures of central tendency for the column showing each students major? Because we can only find one measure of central tendency (mode) for qualitative variables.

Table 3.6.1: Prior Class Data

Student	Major	Final Grade Percentage	Number of Absences
1	Business	83.6	5
2	Business	73.4	4
3	Criminal Justice	82.9	8
4	Criminal Justice	72.4	9
5	Education	86.0	2
6	Education	83.9	7
7	Education	83.0	4
8	Psychology	36.0	25
9	Psychology	36.0	17
$\Sigma =$	$\Sigma = 0$	$\Sigma = 637.20$	$\Sigma = 81.00$

Practice: Final Grade Percentage

Okay, let's move on to a variable that we *can* calculate all three measures of central tendency, and then see what they are trying to tell us.

Example 3.6.2

To interpret the scenario and measures of central tendency for Final Grade Percentage provided in Table 3.6.1, answer the following questions.

1. Who is the sample? Include as much information as you have (such as number in the sample, course, and professor).
2. Who could be the population? In other words, what's the biggest group that this sample could represent?
3. What is the mean, median, and mode for Final Grade Percentage?
    1. Hint: For the median, create a frequency table.
4. If we couldn't see all of the data in Table 3.6.1 but knowing the measures of central tendency (mean, median, & mode), do you think that the frequency distribution will be positively skewed, negatively skewed, or not skewed? In other words what would we guess that the frequency line graph for the Final Grade Percentage will look like based on these measures of central tendency?

Solution

1. The sample is nine students from a section of Dr. MO's behavioral statistics course.
2. The population could be all students in behavioral statistics courses. The population could also be all students from the same college, or all students in Dr. MO's behavioral statistics courses through time. There's not one right answer for who could be the population, it just has to make sense.
3. For Final Grade Percentage:
    1. Mean:

$$\overline{X} = \frac{\sum X}{N} = \frac{637.20}{9} = 70.80$$

    2. Median: The exact middle score is 82.90
        1. You have to put the scores in numerical order to find the one that is in the middle.
    3. Mode: The most frequent score is 36.00.

4. Because the mode is so small, I would predict that the shape of the distribution is positively skewed. The mean responds to extreme scores, so usually the mean is closer to the tail of the distribution than the median. When the mean is smaller than the median, you would predict a a negative skew (see Figure 3.6.1). With these competing predictions for skew, there's no way to predict the skew or shape of the distribution. This often happens with small distributions.

Despite the measures of central tendency not being useful to predict the shape of the graph (probably due to our small sample size), you can see in Figure 3.6.2 that the distribution is positively skewed. And before you get worried that the mode is such a low final grade, notice that the rest of the grades are passing. There were two students who stopped attending class during the first month, and those were the only two students who did not pass the course! When you have a small sample, a couple extreme scores can really change the whole interpretation.

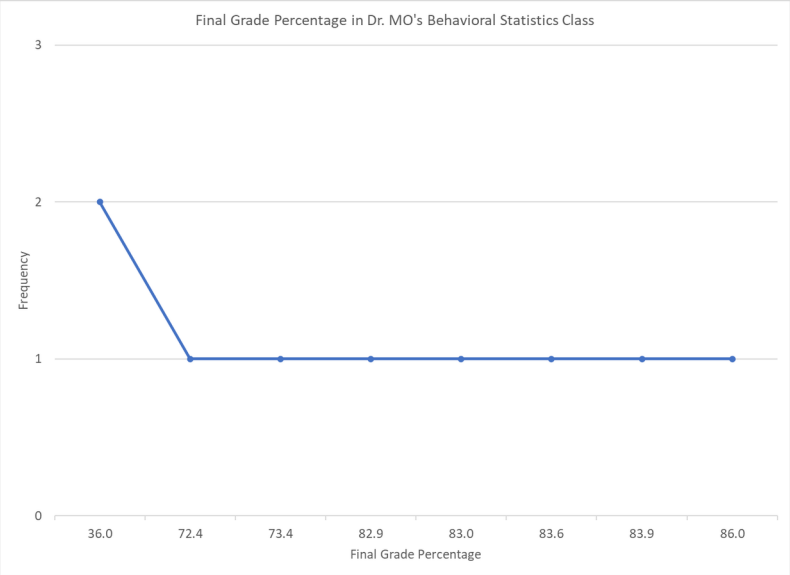


Figure 3.6.2: Line Graph of Frequency of Final Grade Percentages. (Copyright CC-SA, chart by Dr. MO via class data)

Practice: Absences

It's your turn now!

Exercise 3.6.1

To interpret the scenario and measures of central tendency for Absences provided in Table 3.6.1, answer the following questions.

1. Who is the sample? Include as much information as you have (such as number in the sample, course, and professor).
    1. Hint: This is the same as above since it's the same people in the sample.
2. Who could be the population? In other words, what's the biggest group that this sample could represent?
    1. 1. Hint: This could the same as above since it's the same people in the sample
3. What is the mean, median, and mode for Final Grade Percentage?
    1. Hint: For the median, create a frequency table.
4. If we couldn't see all of the data in Table 3.6.1 but knowing the measures of central tendency (mean, median, & mode), do you think that the frequency distribution will be positively skewed, negatively skewed, or not skewed? In other words what would we guess that the frequency line graph for the Final Grade Percentage will look like based on these measures of central tendency?

Answer available at the end of this section.

Look at the distribution in Figure 3.6.3 to see if your prediction was correct:

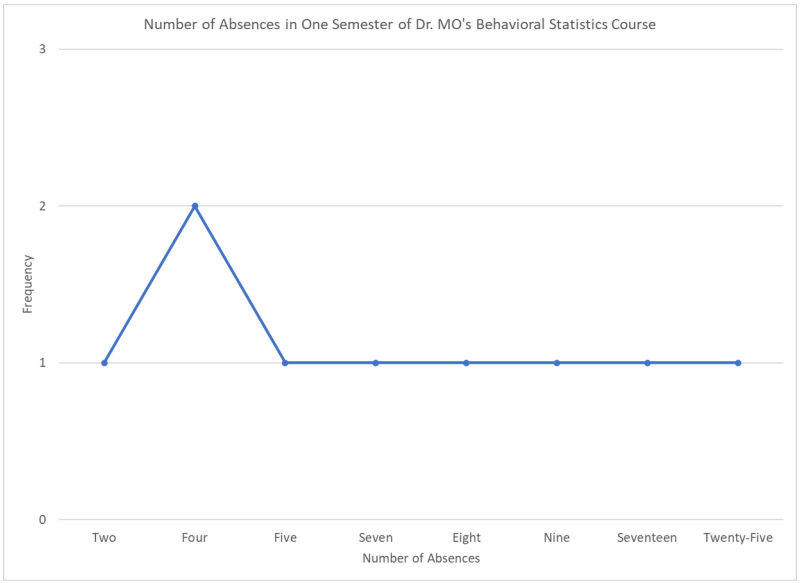


Figure 3.6.3: Line Graph of Frequency of Absences. (Copyright CC-SA, chart by Dr. MO via class data)

While we are not analyzing whether absences are related to final grades, Table 3.6.1 seems to suggest that the two students who missed a lot of classes (making our average number of absences really high!) are also the two students who did not pass the course. In sum, attend your classes!

Conclusion

Although you've learned about the measures of central tendency before, this section shows you how we can use these three numbers together to understand what a whole distribution of data might tell us.

In the next chapter, we'll look at measures of variability that, combined with the measures of central tendency, will tell us a LOT about a distribution of data.

Answers

Exercise 3.6.1

1. The sample is nine students from a section of Dr. MO's behavioral statistics course.
2. The population could be all students in behavioral statistics courses. The population could also be all students from the same college, or all students in Dr. MO's behavioral statistics courses through time. There's not one right answer for who could be the population, it just has to make sense.
3. For Absences:
    1. Mean:

$$\overline{X} = \frac{\sum X}{N} = \frac{81}{9} = 9$$

    2. Median: The exact middle score is 7.00
        1. You have to put the scores in numerical order to find the one that is in the middle.
    3. Mode: The most frequent score is 4.
4. Because the mode is smaller than the median and mean, and the mean seems be pulled higher by extreme scores, I would predict that the shape of the distribution is positively skewed.

This page titled 3.6: Interpreting All Three is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 3.4: Interpreting All Three Measures of Central Tendency by Michelle Oja is licensed CC BY-SA 4.0.

3.7: Summary of 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 "$\overline{\mathrm{X}}$" (pronounced "X-bar") is used for the mean of a sample. The formula for μ is shown below:

$$\mu = \frac{\sum \mathrm{X}}{N}$$

where $\Sigma \mathbf{X}$ is the sum of all the numbers in the population and N is the number of numbers in the population.

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

$$\overline{\mathrm{X}} = \frac{\sum \mathrm{X}}{N}$$

where $\Sigma \mathbf{X}$ 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 $\overline{\mathrm{X}}$).

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 $3.7.1$ 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 $3.7.1$: Number of touchdown passes. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 $3.7.1$, 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 3.7.1, 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 3.7.1 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 3.7.1: 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 3.7.2 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 3.7.2: 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

Value	Absolute Deviation from Median	Absolute Deviation from Mean	Squared Deviation from Median	Squared Deviation from Mean
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 3.7.2 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.

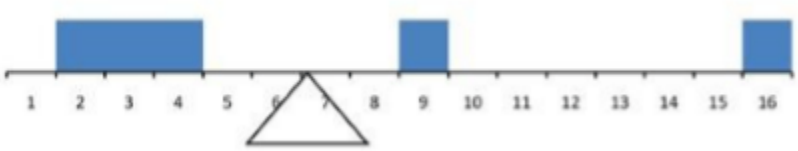


Figure 3.7.2: The distribution balances at the mean of 6.8 and not at the median of 4.0. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 3.7.3 shows the distribution of 642 scores on an introductory psychology test. Notice this distribution has a slight positive skew.

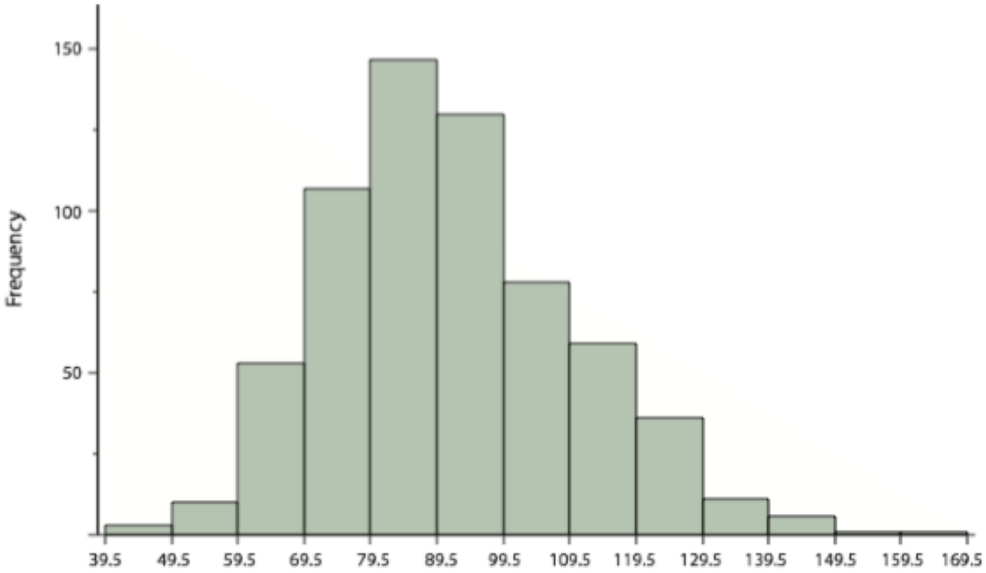


Figure 3.7.3: A distribution with a positive skew. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Measures of central tendency are shown in Table 3.7.3. 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 3.7.3: 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 3.7.4 has a much more pronounced skew than the distribution in Figure 3.7.3.

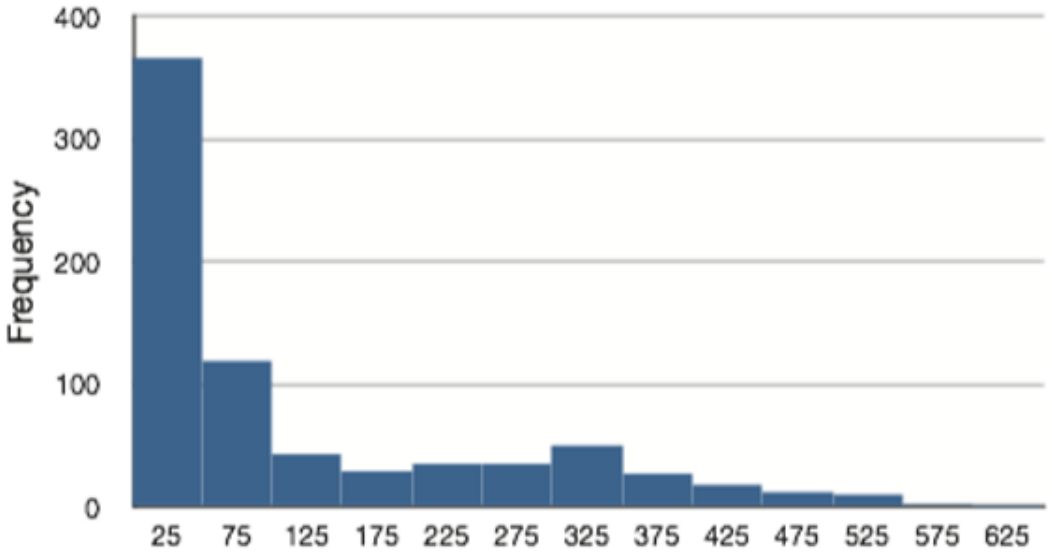


Figure 3.7.4: A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars). (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Table 3.7.4 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 3.7.4: Measures of central tendency for baseball salaries (in thousands of dollars).

Measure	Value
Mode	250
Median	500
Mean	1,183

This page titled 3.7: Summary of 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.

4: MEASURES OF VARIABILITY

4.1: Introduction to Measures of Variability

4.2: Introduction to Standard Deviation Calculations

4.3: Practice SD on Wait Times

4.4: Interpreting Standard Deviations

4.5: Fractions, Decimals, and Percents, Oh My!

4.6: Putting All of the Measures Together

This page titled 4: Measures of Variability is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

4.1: Introduction to Measures of Variability

Variability refers to how “spread out” a group of scores is. To see what we mean by spread out, consider graphs in Figure 4.1.1. These graphs represent the GPA and the hours that the student studied (the GPA_Study_Hours file from www.OpenIntro.org/data website). You can see that the distributions are quite different. Specifically, the scores on Figure 4.1.1 are more densely packed and those on Figure 4.1.2 are more spread out. The differences among students in the hours that they studies were much greater than the differences among the same students on their GPA. These histograms show that some scores are closer to the mean than other scores, and that some distributions have more scores that are closer to the mean and other distributions have scores that tend to be farther from the mean. Sometimes, knowing one, or all, of the measures of central tendency is not enough.

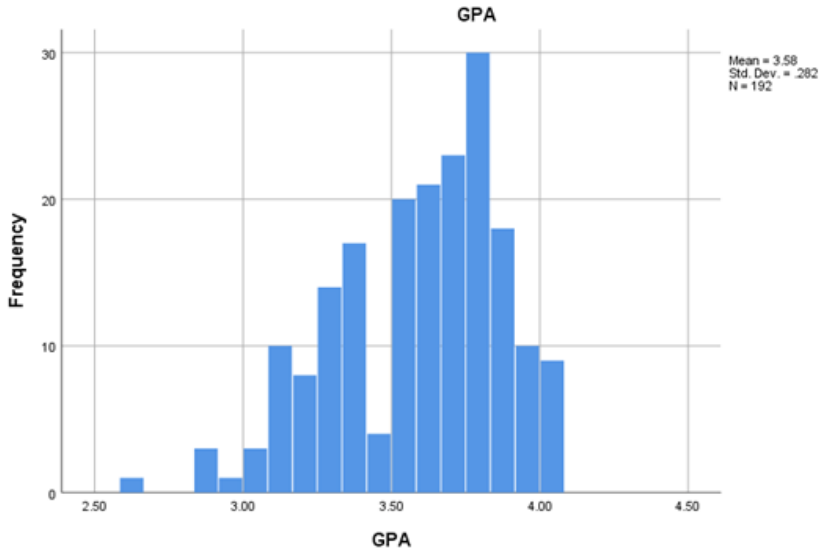


Figure 4.1.1: Histogram of GPA (CC-BY-SA Michelle Oja via data from OpenIntro.org)

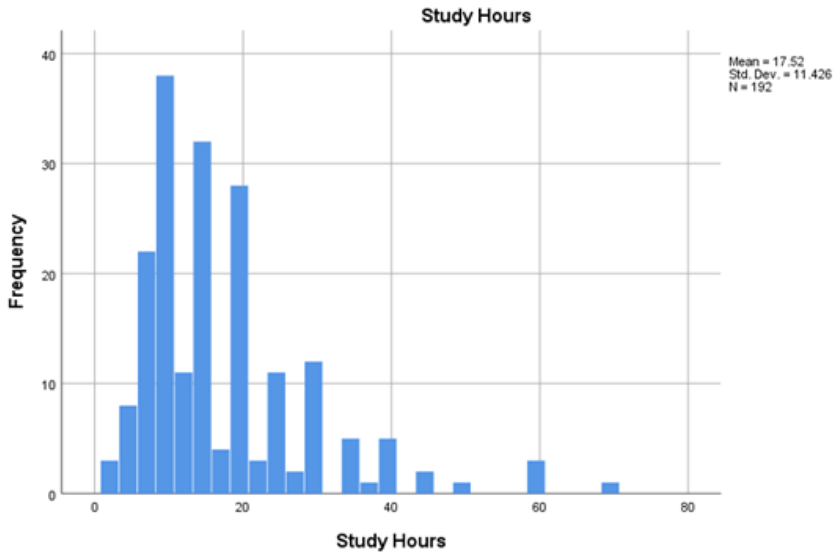


Figure 4.1.2: Histogram of Study Hours (CC-BY-SA Michelle Oja via data from OpenIntro.org)

Figure 4.1.3 shows sets of two normal distributions. Notice that some distributions are tall and narrow, and some are wider and flatter. The tall and narrow distribution has small variability between each score and the mean, and the wide and flat distribution has more variability between each score and the mean. The mean is a better description of the distribution of the data when it is tall and narrow since most scores aren't that different from the mean.

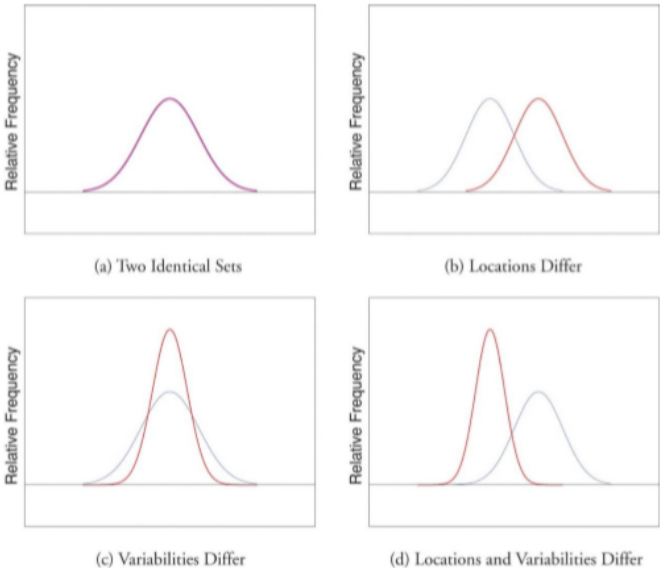


Figure 4.1.3: Differences between two datasets. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 section we will discuss measures of the variability of a distribution. We will focus on two frequently used measures of variability: range and standard deviation. We'll start with range now, then move on to standard deviation next.

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 4.1.1 and Figure 4.1.2. 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.

Contributors

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

This page titled 4.1: Introduction to Measures of Variability is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

4.2: Introduction to Standard Deviation Calculations

Standard Deviation

While range tells you some information about the variability of a distribution, what if we could provide a standard way to think about this variability? That's the standard deviation! The standard deviation tells us how different each score is from the mean, on average.

Historical Context

In 1894, Karl Pearson introduces the term "standard deviation". If errors are normally distributed, 68% of samples will lie within one standard deviation of the mean (Champkin, 2014). Unfortunately, Pearson was influential in the eugenics movement (Encyclopedi.com; Wikipedia). In 1925, he founded the journal Annals of Eugenics; this journal was renamed Annals of Human Genetics in 1954 (Gupta et al., 2015). Sometimes the historical context shows reprehensible behavior.



Figure 4.2.1: Pearson in 1910 by unknown author - National Portrait Gallery, London. Public Domain

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 a fake Quiz 1 are shown in Table 4.2.1. The mean score is 7.0:

$$\frac{\Sigma X}{N} = \frac{140}{20} = 7$$

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 4.2.1 is formatted, as this is the format you will use to calculate 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. THIS IS VERY IMPORTANT. When you mis-calculate an equation, it is often because you did some simple math (adding or subtracting) incorrectly. It is very useful when equations have these self-checking points in them, so I encourage you to use them. 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 4.2.1: 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
$\Sigma = 140$	$\Sigma = 0$	$\Sigma = 30$

The calculations in Table 4.2.1 can be done by hand, but it is also very easy to set up the data in any spreadsheet program and learn the simple commands to make the spreadsheet do the simple math. As long as you tell it what to do with the correct numbers, then your results will be correct. You can also use the memory function in graphing calculators to save the the data set, and run some of the more common mathematical functions. Using spreadsheets and your graphing calculator to do the math also saves problems with rounding since the devices keep all of the decimals so you only have to round your final result. This statistics textbook will not go into explanations on how to use software (like spreadsheets, calculators, or more sophisticated statistical programs), but much that is easily accessible (like spreadsheets on Excel or Google) are relatively easy to learn to use

Variance (of a Sample)

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:

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

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 $4.2.1$ 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.

But we won't do much with variance of a population. Instead, we'll focus on variance of a sample. The sample statistic used to estimate the variance is s^2 ("s-squared"):

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

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 $4.2.1$ and treating those scores as a sample, we can estimate the sample variance as:

\[s^{2}=\dfrac{30}{20-1}=1.58 \nonumber \]

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.

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

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

The sample standard deviation from Table $4.2.1$ is:

$$s = \sqrt{\frac{30}{20 - 1}} = \sqrt{\frac{30}{19}} = \sqrt{1.58} = 1.26$$

We'll practice calculating standard deviations, then interpreting what the numbers mean. Because in behavioral statistics, it's not about the numbers. We never end with a number, we end with a conclusion (which can be as simple as a sentence, or can be several paragraphs). Social scientists want to know what the numbers *mean* because we use statistical analyses to answer real questions.

Reference

Champkin, J. (2014, January 24). *The timeline of statistics*. Significance. https://www.statsref.com/timeline.pdf

Gupta, V.K., Mandal, B.N., & Parsad, R. (2015). History of statistics on timeline. Society of Statistics, Computer and Applications. https://www.ssca.org.in/media/History_of_Statistics_on_Timeline.pdf

Contributors

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

This page titled 4.2: Introduction to Standard Deviation Calculations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

4.3: Practice SD on Wait Times

You may or may not understand the importance of calculating and understanding the variation of your data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The standard deviation is a number that measures how far data values are from their mean.

The Standard Deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

Answering Questions

There are a couple common kinds of questions that standard deviations can answer, in addition being foundational for later statistical analyses. First, a standard deviation helps understand the shape of a distribution. Second, a standard deviation can show if a score is extreme.

Describing the Shape of a Distribution

The standard deviation provides a measure of the overall variation in a data set.

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. Distributions with small standard deviations have a tall and narrow line graph. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation. Distributions with large standard deviations may have a wide and flat lin graph, or they may be skewed (with the outlier(s) making the standard deviation bigger).

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes.

Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average.

Identifying Extreme Scores

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is **two minutes longer than the average** of five minutes.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.

- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because $5 + (1)(2) = 7$.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because $5 + (-2)(2) = 1$.



Figure 4.3.1- Scale from 0 to 7 (CC-BY by Barbara Illowsky & Susan Dean (De Anza College) from OpenStax)

- In general, a **value = mean + (#ofSTDEV)*(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because: $1 = 5 + (-2)(2)$. (The numbers in parentheses that touch should be multiplied)

The equation value = mean + (#ofSTDEVs)*(standard deviation) can be expressed for a sample and for a population.

- sample: $x = \overline{x} + (\text{Number of Standard Deviations}) \times (s)$
- Population: $x = \mu + (\text{Number of Standard Deviations}) \times (s)$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol $\overline{x}$ is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation

If x is a number, then the difference "x – mean" is called its deviation. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \overline{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The variance is the average of the squares of the deviations (the $x - \overline{x}$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire population and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N, the number of items in the population. If the data are from a sample rather than a population, when we calculate the average of the squared deviations, we divide by *n* – 1, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{n - 1}}$$

For the sample standard deviation, the denominator is $n - 1$, that is the sample size MINUS 1.

Practice!

Example 4.3.1

In a fifth grade class at a private school, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a sample of *n* = 20 fifth grade students. The ages are rounded to the nearest half year in Table 4.3.1, but first let's talk about the context.

1. Who was the sample? Who could this sample represent (population)?

The sample is the 20 fifth graders from a private school. The population could be all fifth graders from private schools?

2. What was measured?

Age, in years, was measured. This is the DV, the outcome variable.

Table 4.3.1- Ages of a sample of 20 fifth graders

Ages of Sample Fifth Graders
9
9.5
9.5
10
10
10
10
10.5
10.5
10.5
10.5
11
11
11
11
11
11
11.5
11.5

Ages of Sample Fifth Graders
11.5

3. What is the mean?

$$\overline{x} = \frac{(9 + 9.5 + 9.5 + 10 + 10 + 10 + 10 + 10.5 + 10.5 + 10.5 + 10.5 + 11 + 11 + 11 + 11 + 11 + 11 + 11.5 + 11.5 + 11.5)}{20} = 10.525 =$$

The average age is 10.53 years, rounded to two places.

4. What is the standard deviation?

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s.

Table 4.3.1- *Ages of One Fifth Grade Class*

Data	Deviations	*Deviations*^{2}
x	**(X – $\overline{X}$)**	**(X– $\overline{X}$)^2**
9	$9–10.525 = –1.525$	$(–1.525)^2 = (–1.525 \times –1.525) = 2.325625$
9.5	$9.5–10.525 = –1.025$	$(–1.025)^2 = (–1.025 \times –1.025) = 1.050625$
9.5	$9.5–10.525 = –1.0.25$	$(–1.025)^2 = 1.050625$
10	$10–10.525 = –0.525$	$(–0.525)^2 = (–0.525 \times –0.525) = 0.275625$
10	$10–10.525 = –0.525$	$(–0.525)^2 = 0.275625$
10	$10–10.525 = –0.525$	$(–0.525)^2 = 0.275625$
10	$10–10.525 = –0.525$	$(–0.525)^2 = 0.275625$
10.5	$10.5–10.525 = –0.025$	$(–0.025)^2 = (–0.025 \times –0.025) = 0.000625$
10.5	$10.5–10.525 = –0.025$	$(–0.025)^2 = 0.000625$
10.5	$10.5–10.525 = –0.025$	$(–0.025)^2 = 0.000625$
10.5	$10.5–10.525 = –0.025$	$(–0.025)^2 = 0.000625$
11	$11–10.525 = 0.475$	$(0.475)^2 = (0.475 \times 0.475) = 0.225625$
11	$11–10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11–10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11–10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11–10.525 = 0.475$	$(0.475)^2 = 0.225625$
11	$11–10.525 = 0.475$	$(0.475)^2 = 0.225625$
11.5	$11.5–10.525 = 0.975$	$(0.975)^2 = (0.975 \times 0.975) = 0.950625$
11.5	$11.5–10.525 = 0.975$	$(0.975)^2 = 0.950625$
11.5	$11.5–10.525 = 0.975$	$(0.975)^2 = 0.950625$
$\sum X$	**0 (basically)**	$\sum = 9.7375$

The first column in Table 4.3.1 has the data, the second column has has deviations (each score minus the mean), the third column has deviations squared. The first row is the row's title, the second row is the symbols for that column, the rest of the rows are the scores until the bottom row, which is the sum of each of the rows.

Take the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

$$\frac{9.7375}{20-1} = 0.5125$$

The **sample standard deviation** s is the square root of $\frac{SS}{df}$:

$$s = \sqrt{0.5125} = 0.715891$$

and this is rounded to two decimal places, $s = 0.72$. The standard deviation of the sample fo 20 fifth graders is 0.72 years.

Typically, you do the calculation for the standard deviation on your calculator or computer. When calculations are completed on devices, the intermediate results are not rounded so the results are more accurate. It's also darned easier. So why are spending time learning this outdated formula? So that you can see what's happening. We are finding the difference between *each score* and the mean to see how varied the distribution of data is around the center, dividing it by the sample size minus one to make it like an average, then square rooting it to get the final answer back into the units that we started with (age in years).

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**.
- For a sample: $x = \overline{x}$ + (#ofSTDEVs)(s)
- For a population: $x = \mu$ + (#ofSTDEVs)σ
- For this example, use *x* = $\overline{x}$ + (#ofSTDEVs)(s) because the data is from a sample

5. Verify the mean and standard deviation on your own.
6. Find the value that is one standard deviation above the mean. Find ($\overline{x}$ + 1s).
7. Find the value that is two standard deviations below the mean. Find ($\overline{x}$ – 2s).
8. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution

a. You should get something close to 0.72 years, but anything from 0.70 to 0.74 shows that you have the general idea.
b. $(\overline{x} + 1s) = 10.53 + (1)(0.72) = 11.25$
c. $(\overline{x} - 2s) = 10.53–(2)(0.72) = 9.09$
d.
   - $(\overline{x} - 1.5s) = 10.53–(1.5)(0.72) = 9.45$
   - $(\overline{x} + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

Notice that instead of dividing by $n = 20$, the calculation divided by $n - 1 = 20 - 1 = 19$ because the data is a sample. For the sample, we divide by the sample size minus one $(n - 1)$. The sample variance is an estimate of the population variance. After countless replications, it turns out that when the formula division by only N (the size of the sample) is used on a sample to infer the population's variance, it always under-estimates the variance of the population.

Which one has the bigger solution, the one with the smaller denominator or the larger denominator?

- $\frac{10}{2} =$
- $\frac{10}{5} =$

Smaller denominators make the resulting product **larger.** To solve our problem of using the population's variance formula on a sample under-estimating the variance, we make the denominator of our equation smaller when calculating variance for a sample. In other words, based on the mathematics that lies behind these calculations, dividing by $(n - 1)$ gives a better estimate of the population.

What does it mean?

The deviations show how spread out the data are about the mean. From Table $4.3.1$, The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value (age, in this case) is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean

(that particular student is younger than the average age of the class) . The deviation is –1.525 for the data value nine. If you add the deviations, the sum is always zero, so you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation. But the variance is a squared measure and does not have the same units as the data. No one knows what 9.7375 years squared *means*. Taking the square root solves the problem! The standard deviation measures the spread in the same units as the data.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is zero, there is no spread; that is, all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

Exercise 4.3.1

Scenario: Using one baseball professional team as a sample for all professional baseball teams, the ages of each of the players are shown in Table 4.3.2.

Table 4.3.2- One Baseball Team's Ages

Data	Deviations	*Deviations*^{2}
x	(x – $\overline{x}$)	(x – $\overline{x}$)^2
21		
21		
22		
23		
24		
24		
25		
25		
28		
29		
29		
31		
32		
33		
33		
34		
35		
36		
36		
36		
36		
38		
38		

Data	Deviations	*Deviations*^{2}
38		
40		
$\sum X$ = 767	$\sum X$ should be 0 (basically)	$\sum X$ = ?

If you get stuck after the table, don't forget that: $s = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}}$

All of your answers should be complete sentences, not just one word or one number. Behavioral statistics is about research, not math.

1. Who was the sample? Who could this sample represent (population)?
2. What was measured?
3. What is the mean? (Get in the practice of including the units of measurement when answering questions; a number is usually not a complete answer).
4. What is the standard deviation?

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

5. Find the value that is two standard deviations above the mean, and determine if there are any players that are more than two standard deviations above the mean.

Answer

1. The sample is 25 players from a professional baseball team. They were chosen to represent all professional baseball players (it says so in the scenario description!).
2. Age, in years, was measured.
3. The mean of the sample ($\overline{X}$ was 30.68 years.
4. The standard deviation was 6.09 years ($s = 6.09$), although due to rounding differences you could get something from about 6.05 to 6.12. Don't worry too much if you don't get exactly 6.09; if you are close, then you did the formula correctly!
5. The age that is two standard deviations above the mean is 42.86 years, and none of the players are older than that.

$$(\overline{x} + 2s = 30.68 + (2)(6.09) = 42.86$$

.

What standard deviation show us can seem unclear at first. Especially when you are unfamiliar (and maybe nervous) about using the formula. By graphing your data, you can get a better "feel" for what a standard deviation can show you. You will find that in symmetrical distributions, the standard deviation can be very helpful. Because numbers can be confusing, **always graph your data**.

Summary

The standard deviation can help you calculate the spread of data.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum (x - \overline{x})^2}{n - 1}}$ is the formula for calculating the standard deviation of a sample.

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 4.3: Practice SD on Wait Times is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 2.8: Measures of the Spread of the Data by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

4.4: Interpreting Standard Deviations

Figure 4.4.1 shows two symmetrical distributions. The frequency distribution on the left (red) has a mean of 40 and a standard deviation of 5; the frequency distribution on the right (blue) has a mean of 60 and a standard deviation of 10. For the tall and narrow distribution on the left, 68% of the distribution is between 45 and 55; for the shorter distribution on the right, 68% is between 50 and 70. Remember that the standard deviation shows how different each scores is, on average, from the mean. Notice that as the standard deviation gets smaller, the distribution becomes much narrower, regardless of where the center of the distribution (mean) is. The smaller standard deviation and the narrower distribution both show that the difference from the mean, on average, is smaller.

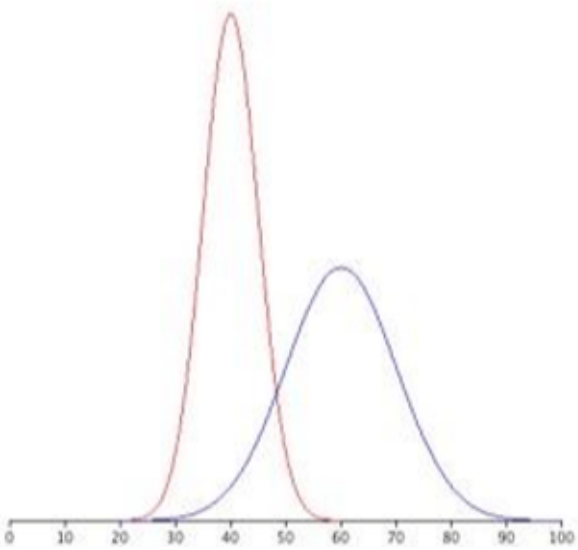


Figure 4.4.1: Symmetrical distributions with standard deviations of 5 and 10. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

The next figure, Figure 4.4.2 was in the Introduction to Measures of Variability section. In Figure Panel (a), the mean and standard deviations are both the same. You can see where the name “bell curve” comes from: it looks a bit like a bell. In Panel (b), the mean of one of the distributions of data (red, on the right) is higher than the mean of the other distribution (blue, on the left) but the standard deviations are the same as shown by these distributions have the same “width”. The only difference between them is that they’ve been shifted to the left or to the right because the means of each distribution are different; in every other respect, they’re identical.

In contrast, if we increase the standard deviation while keeping the mean constant, the peak of the distribution stays in the same place, but the distribution gets taller, as you can see in Figure 4.4.2‘s Panel (c); the means are the same (because the center of both distributions are in the same place on the x-axis), but one standard deviation is smaller (red, tall and narrow) then the other standard deviation (blue, bell shaped). The smaller standard deviation tells us that the scores, on average, are close to the mean. Notice, though, that when a distribution gets taller, the width shrinks (it gets narrower). Finally, Panel (d) shows two distributions of data in which both the means are different and the standard deviations are different (as discussed in the first Figure 4.4.1).

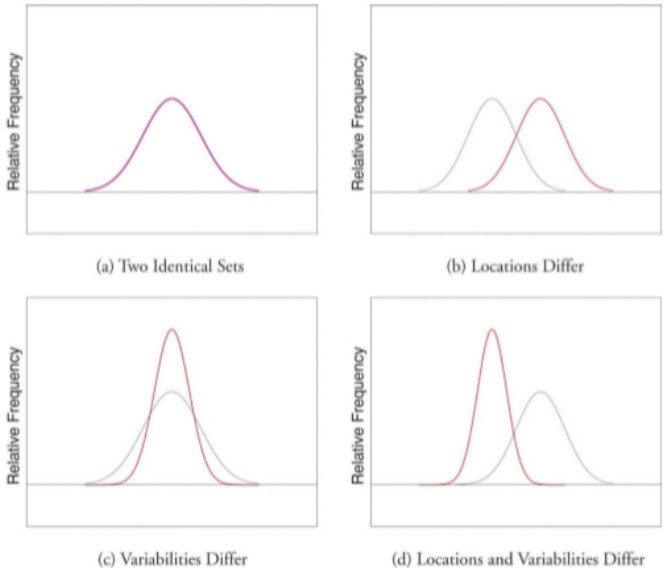


Figure 4.4.2: Differences between two datasets. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Panel (c) and Panel (d) of Figure 4.4.2 show that standard deviations can help us know the shape of a distribution's frequency line graph. In other words, standard deviations can be used to predict if a frequency distribution will be skewed or have kurtosis. See the section in the chapter on graphing that discusses Skew & Kurtosis as a reminder. A small standard deviation suggests that the shape of the distribution is leptokurtic. That's the tall and narrow (red) distribution in Panel (c) and Panel (d) of Figure 4.4.2. A medium standard deviation suggests that the shape of the distribution is normal or bell-shaped, sometimes called mesokurtic. That's the bell-shaped (blue) distribution in Panel (c) and Panel (d) of Figure 4.4.2.

Simple enough? It gets a little more complicated when you have a large standard deviation. A large standard deviation suggestions *either*:

- The shape of a distribution is skewed, *or*
- The shape of the distribution is platykurtic (wide and flat).

A few things to note with these statements. First, just knowing that you have a large standard deviation doesn't tell you which of these two options is true! However, the measures of central tendency from the previous chapter can be used to decide whether you think that a large standard suggests a skewed distribution of platykurtic distribution. If the mean is pretty different from the mode (the highest peak on the graph) and/or the median (the exact middle score) *and* your standard deviation is large, then your frequency distribution is probably skewed. Look at the skewed distribution in Figure 4.4.3 from the Skew and Kurtosis section of the chapter on graphs to see how these measures might interact.

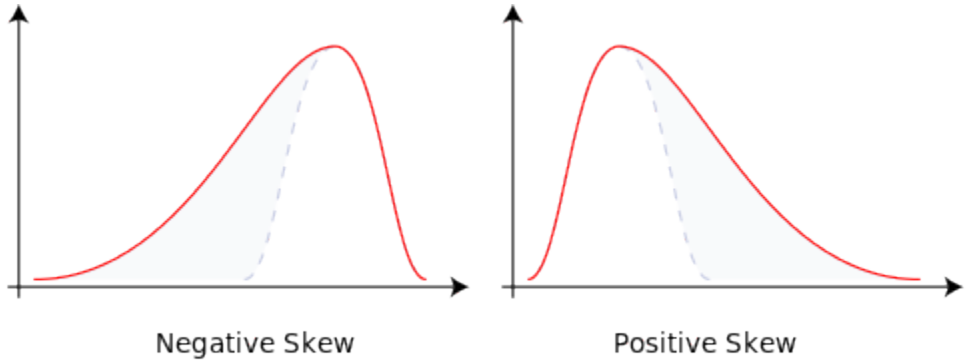


Figure 4.4.3- Diagram of Skew. (CC-BY-SA Rodolfo Hermans (Godot), CC BY-SA 3.0, via Wikimedia Commons)

When a distribution has a large standard deviation with a mean that's different from the mode and/or median, you still can't tell whether the distribution is positively or negatively skewed by just looking at the standard deviation. For example, you can't tell whether the frequency distribution is more like the negative skew on the left or the positive skew on the right of Figure 4.4.3. To determine the type of skew, you'd have to look at how the three measures of central tendency relate; the large standard deviation

is just a clue that the frequency distribution may be skewed (see Interpreting Measures of Central Tendency for a review).

Finally, if the measures of central tendency are close together and your standard deviation is large, then your frequency distribution is probably platykurtic. You can see in Figure 4.4.4 (which is the same figure in the Skew and Kurtosis section in the chapter graphs) that a large standard deviation (which means that each score is, on average, far from the mean in the center) would be platykurtic and that a small standard deviation (showing that each score, on average, is close to the mean in the center) would be leptokurtic.

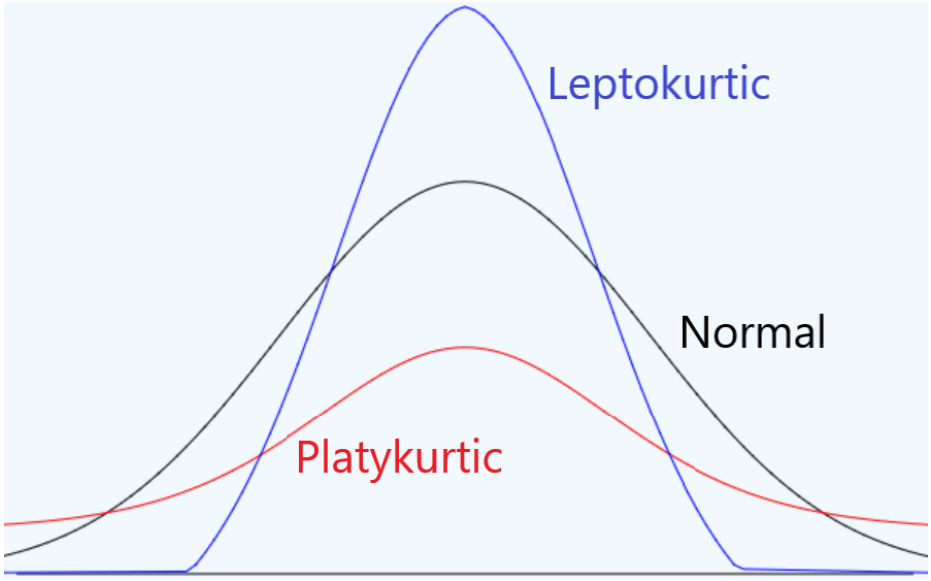


Figure 4.4.4- Diagram of Three Kurtosis Types in Line Graphs (CC-BY Larry Green)

Small, Medium, or Large?

You're probably still wrapping your head around how the measures of central tendency and the standard deviation can help understand the shape of a frequency distribution. I encourage you to create your own decision tree flow chart to follow the logic through all that. But while trying to piece together how the size of the standard deviation relates to the shape of a frequency chart, you might've wondered: how can you tell if your SD is large, medium, or small?! There's no definitive answer to that! However, Dr. MO will provide some guidelines.

Large

First, let's start with how to decide if a standard deviation is large because that is the more complicated decision. A general guideline (not a rigid rule) is that a standard deviation that is more than about a quarter or third of the mean is considered "large". The next sub-section will walk you through fractions and decimals, in case this next bit gets a little to math-y for you! The general idea is that you can divide the mean by 4 (a quarter) or 3 (a third), then compare that result to the standard deviation. If the standard deviation is *bigger* than the mean divided by 3 or 4, then the standard deviation could be considered *large*. Dr. MO hasn't seen this written anywhere, but she has decades of experience working with means and standard deviations to come up with this guideline.

If the mean is 24, then:

$$\frac{12}{4} = 6$$

$$\frac{12}{3} = 8$$

That means that the standard deviation would have to be *bigger* than 6 to 8 to be considered a large standard deviation.

Example 4.4.1

1. If the mean is 10, would a standard deviation of 1.5 be considered large? Why or why not?
2. If the mean is 50, would a standard deviation of 18 be considered large? Why or why not?

Solution

1. Since 10 divided by 4 is 2.50 and 10 divided by 3 is 3.33, a standard deviation of 1.5 is *not* bigger than either of those two, so this standard deviation would *not* be considered large. (Stay tuned for how to determine what size it might be!)
2. Since 50 divided by 4 is 12.50 and 50 divided by 3 is 16.67 and, the standard deviation of 18 would be larger than those, so we could classify it as a *large* standard deviation.

Now, you can do the math on your own to see if you can decide if the standard deviation is large or not.

Exercise 4.4.1

If the mean is 100, would a standard deviation of 20 be considered large? Why or why not?

Answer available at the end of this section.

Sometimes students try to look at just the standard deviation to use it to predict the shape of a distribution, but this is missing necessary context. The standard deviation on its own isn't all that useful; you need to know the central tendencies (especially the average scores), to decide. But as the examples above show, sometimes 18 is a large standard deviation and sometimes 20 is *not* a large standard deviation! It all depends on the data you are working with.

Before we move on, here's a reminder that finding a large standard deviation is just a piece of the puzzle. To try to predict the shape of the distribution, you'd need to compare the large standard deviation with the measures of central tendency to determine if that large standard deviation suggests a skewed distribution *or* a kurtosis that is wide and flat (platykurtic).

Small

Okay, let's figure out how to estimate if the standard deviation could be considered small. You'll see next why we'll talk about medium-sized standard deviations last! Dr. MO felt pretty confident about sharing the quarter-to-a-third general guideline to identify large standard deviations, but she's less confident in generating a guideline for identifying a small standard deviation. The best that she can come up with is if the standard deviation is less than 10% of the mean, then it's probably a small standard deviation.

If the mean is 24, then you'd multiply that by .10 (which is the decimal for 10%).

$$24 \times 0.10 = 2.40$$

That means that the standard deviation would have to be *smaller* than 2.40 to be considered a small standard deviation. If you're a little lost on how to get from a percentage to a decimal (or vice versa), you are highly encouraged to review the next section.

Example 4.4.2

1. If the mean is 10, would a standard deviation of 1.5 be considered small? Why or why not?
2. If the mean is 50, would a standard deviation of 18 be considered small? Why or why not?

Solution

1. A standard deviation of 1.5 is *not* smaller 1.00, which is 10% of 10 (10 times 0.10 is 1.00), so this standard deviation would *not* be considered small. (Stay tuned for how to determine what size it might be!)

2. Since 50 times 0.10 is 5, then a standard of 18 would also not be smaller, so this standard deviation would also *not* be considered small.

Try it yourself!

Exercise 4.4.2

If the mean is 100, would a standard deviation of 20 be considered small? Why or why not?

Answer available at the end of this section.

Exercise 4.4.3

What kurtosis would we predict with a small standard deviation?.

Answer available at the end of this section.

Medium

How do you decide if a standard deviation is medium-sized? Basically, if your standard deviation is not large and is not small, then it's medium!

Exercise 4.4.4

What kurtosis would we predict with a medium standard deviation?

Answer available at the end of this section.

Conclusion

Understanding the relationships amongst the measures of central tendencies and the standard deviation (as a measure of variability), is a major way to show that your math literacy. After the sub-section on fractions, you can practice all of this, plus refresh on using just the three measures of central tendency to predict the shape of a distribution.

Contributors

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

Answers

Exercise 4.4.1

Since 100 divided by 4 is 25.00 and 100 divided by 3 is 33.33, a standard deviation of 20 is *not* bigger than either of those two, so the standard deviation of 8 would *not* be considered large. (Stay tuned for how to determine what size it might be!)

Exercise 4.4.2

Since 10% of 100 is 10 (100 times 0.10 is 10), a standard deviation of 20 is also *not* smaller, so the standard deviation of 20 would *not* be considered small.

Exercise 4.4.3

Small standard deviations suggest that each score, on average, is similar to the mean. This would be a leptokurtic frequency distribution, which is tall and narrow.

Exercise 4.4.4

Medium standard deviations suggest a mesokurtic frequency distribution, which is shaped like a bell.

This page titled 4.4: Interpreting Standard Deviations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

4.5: Fractions, Decimals, and Percents, Oh My!

You've been working with fractions, decimals, and even percentages already in this course and in your life. Here's some specifics, though, so that you aren't guessing at what to do.

Historical Context

The Rhind papyrus, dated to c. 1650 BCE but likely a copy of an older document from the Middle Kingdom of about 2000–1800 BCE is an instruction manual for students. In addition to giving area formulas and methods for multiplication, division and working with unit fractions (Wikipedia).



Figure 4.5.1: Rhind papyrus by unknown found in ArchaeoWiki from British Museum. Public Domain

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). For example:

- Small N: Having 1 red-head in a group of 10 $= 10\%$
- Large N: Having 1 red-head in a group of 100 $= 1\%$

In this example showing the frequency of one for each group 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.

But how do you calculation a percentage?

$$(x/N) \times 100$$

In which x is the number that you're starting with (the frequency of 1) and N is the total number of scores.

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 4.5.1

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%

Let's try again:

Exercise 4.5.1

What if there were 75 students in your class, and 5 of them had earned full credit on all assignments.

1. What's "x" in this example?
2. What's "N" in this example?
3. What's the percentage of students in this new, larger class earning an A+?

Answer available at the end of this section.

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.

Transforming a Decimal to a Percent

In this section, we will convert from decimals to percents and back. We will also start with a fraction and convert it to a decimal and a percent. In statistics we are often given a number as a percent and have to do calculations on it. To do so, we must first convert it to a percent. Basically, we'll practice with fractions, decimals, and percents so that you get more comfortable with the math.

As "percent" means parts per hundred, when we want to convert a decimal to a percent, we multiply the decimal by 100. In practice, this means we move the decimal point two places to the right.

Example 4.5.2

Convert the number 0.1738 to a percent.

Solution

We move the decimal over two to the right as shown below.



We get: 17.38% for the answer.

Example 4.5.3

Convert 0.7 to a percent.

Solution

We want to move the decimal two places to the right, but there is only one digit to the right of the decimal place. The good news is that we can always add a 0 to the right of the last digit. We write:

$$0.7 = 0.70$$

Now move the decimal place two digits to the right to get 70%.

Example 4.5.4

In regression analysis (which will be discussed MUCH later), an important number that is calculated is called R-Squared. It helps us determine how helpful one variable is in predicting another variable. The computer and calculator always display it as a decimal, but it is more meaningful as a percent. Suppose that the R-Squared value that relates the amount of studying students do to prepare for a final exam and the score on the exam is: $r^2 = 0.8971$. Convert this to a percent rounded to the nearest whole number percent.

Solution

We move the decimal 0.8971 two places to the right to get 89.71%

Now round to the nearest whole number percent. Note that the digit to the left of the whole number is 7 ≥ 5. Thus we add 1 to the whole number, 89. This gives us 90%.

Transforming a Percent to a Decimal

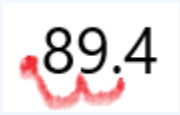
To convert a decimal to a percent, we multiply the decimal by 100 which is equivalent to moving the decimal two places to the right. Not surprisingly, to convert a percent to a decimal, we do exactly the opposite. We divide the number by 100 which is equivalent to moving the decimal two places to the left.

Example 4.5.5

Convert the percent 89.4% to a decimal.

Solution

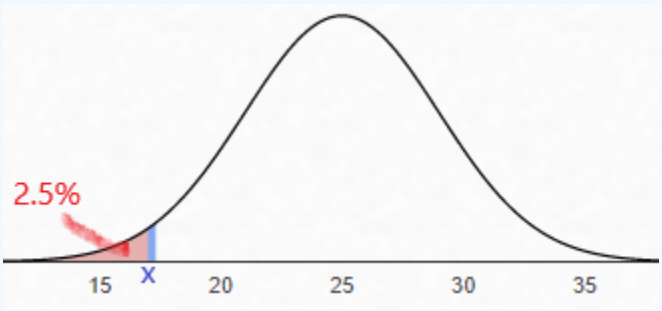
We move the decimal over two to the left as shown below.



We get: 0.894 for the answer.

Example 4.5.6

Suppose that you want to find the value of x such that 2.5% of the entire area under a frequency line graph lies to the left of x. The first step will be to convert the 2.5% to a decimal. What decimal is equivalent to 2.5%?



Solution

We want to move the decimal 2.5 two places to the left, but since there is only one digit to the left of the decimal, we add a zero first: 02.5. Now move the decimal two places to the left to get 0.025.

Converting a Fraction to a Decimal and a Percent

Often in probability it is natural to represent probabilities as fractions, but it is easier to make comparisons as decimals. Thus, we need to be able to convert fractions to decimals. To do so we just divide.

Example 4.5.7

Find the percentage of club members attended an end-of-year event if four showed up out of seven active club members.

Solution

$$(x/N) \times 100$$

$$(\frac{4}{7} \times 100) = (0.57142 \times 100) = 57.14$$

Note that we do not round up because the test number is lower than 5.

The percentage of club members who attended the end-of-year event was 57.14%.

Example 4.5.8

In statistics we need to find basic probabilities and create a table for them. Suppose that you roll two six-sided dice, what percent of the time will the sum equal to a 4? Round to the nearest whole number percent.

Solution

First, notice that there are 36 total possibilities for rolling the dice, since there are 6 faces on the first die and for each value of the first die roll, there are 6 possibilities for the second die roll. Multiplying: 6 x 6 = 36. This will be the denominator. To find the numerator, we list all the possible outcome where the sum is 4:

(1,3), (2,2), and (3,1)

There are three possible outcomes with the sum equaling a 4. Thus:

$$P(\text{sum} = 4) = 3/36$$

Now we divide:

$$\frac{3}{36} = 0.08333...$$

Next to convert this decimal to a percent, we move the decimal two places to the right to get: 8.333...%

We are asked to round to the nearest whole number percent. The digit to the right or the whole number (8) is a 3. Since 3 < 5, we can just erase everything to the left of the 8 and leave the 8 unchanged to get 8%. Thus there is an 8% chance of getting a sum of 4 if you roll two six sided dice.

- Convert Percentages to Decimals
- Relating Fractions, Decimals, and Percents

- Statistics Application of Converting Decimals to Percents

Conclusion

You won't always need to find a percentage, but keep track of this page the following resource links for when they do come back. Like all math, you lose it if you don't use it! One key to being successful is to keep track (and use!) your resources! You're usually not expected to have all of this memorized, especially at first. But you are expected to be able to find your notes and resources on how to do what you've already learned.

Resources

The following link to videos for each of these topics.

- Convert Percentages to Decimals
- Relating Fractions, Decimals, and Percents
- Statistics Application of Converting Decimals to Percents

One last note: The computer or calculator shows numbers as decimals, but for presentations, percents are friendlier. It is also much easier to compare decimals than fractions, thus converting to a decimal is helpful. For example, we often want to see if a probability is greater than 5%. A computer will display the probability as a decimal such as 0.04836. To make the comparison we will first change it to a percent and then compare it to 5%.

And now we'll move back to interpreting standard deviations!

Answers

Exercise $4.5.1$

1. The number that we're starting with ("x") is still the five students.
2. The total number of students in this new class is 75 students.
3. The formula below shows that the percentage of students earning an A+ in this larger class is 8.33%

$$\frac{5}{75} = 0.0666...$$

then

$$0.0666 \times 100 = 6.67$$

Since the test number is over 5, then we round up!

This page titled 4.5: Fractions, Decimals, and Percents, Oh My! is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 3.2: Math Refresher by Michelle Oja is licensed CC BY-NC-SA 4.0.
- Converting Between Fractions, Decimals and Percents by Larry Green is licensed CC BY 4.0.

4.6: Putting All of the Measures Together

Phew! We've covered a lot in the last few chapters, and now we're going to put it all together to calculate a standard deviation, then see if we can predict the shape of a distribution just based on the measures of central tendency and the standard deviation as the measure of variability.

Let's start with the raw data in Table 4.6.1. The data are from 9 students in a face-to-face behavioral statistics class. The scores provided are each students' score (out of 100 points) for all assignments related to research. These activities included smaller projects like worksheets identifying variables (which you could do now!), medium projects like calculating measures of central tendency and standard deviations as a measure of variability (which you could do now!), to larger projects like papers that statistically compared the means of different groups (which you'll start learning in a few chapters).

Table 4.6.1- *Research Activities Scores for Face-to-Face Class*

Research Activities Score for Face-to-Face Class
89
79
75
71
64
64
64
57
56
$\sum = 619.00$

Table 4.6.1 also includes the sum of the 9 scores in the last row because adding a bunch of scores in a calculator is hard!

Let's start calculating the measures of central tendency.

Example 4.6.1

1. What is the mode for the Research Activities points in Table 4.6.1?
2. What is the median for the Research Activities points in Table 4.6.1?
3. What is the mean for the Research Activities points in Table 4.6.1?

Solution

1. The mode is 64 points as that is the only number that repeated, so it is the most frequent score.
2. The median is also 64 points because that is the number that is in the exact middle of the distribution (four scores are higher and four scores are lower).
3. The mean is rounded to 68.78 points.

$$\overline{X} = \frac{\sum X}{n} = \frac{619.00}{9} = 68.78$$

Once you have these three numbers that all represent the center of the frequency distribution, you can use them to make predictions about the shape of the distribution.

Example 4.6.2

Based on the three measures of central tendency, what do you think the shape of the frequency distribution might be?

Solution

The mode will be the peak of the distribution, at 64 points. The median is the middle score in the whole distribution, and that is also 64 points. The mean is a little bigger at 68.78 points, which suggests that there are a few scores that are high (since the mean is higher than both the mode and the median). With a mean that's higher than the mode and median, we'd expect a positively skewed distribution in which most scores are to the left, but there are one or a few higher scores.

We can now calculate the standard deviation.

Example 4.6.3

What is the standard deviation for the Research Activities scores from Table 4.6.1?

Solution

You know how to do this!

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

Use Table 4.6.2 to calculate the numerator.

Table 4.6.2- Find the mean difference, squared mean difference, and the sum of squares in this table.

X	$X - \overline{X}$	**Sums of Squares** ($(X - \overline{X})^2$)
89	$89 - \overline{X} = 20.22$	$(X - \overline{X})^2 = 20.22^2 = 408.85$
79	$79 - \overline{X} = 10.22$	$(X - \overline{X})^2 = 10.22^2 = 104.45$

X	$X - \overline{X}$	Sums of Squares ($(X - \overline{X})^2$)
75	$75 - \overline{X} = 6.22$	$(X - \overline{X})^2 = 6.22^2 = 38.69$
71	$71 - \overline{X} = 2.22$	$(X - \overline{X})^2 = 2.22^2 = 4.93$
64	$64 - \overline{X} = -4.78$	$(X - \overline{X})^2 = -4.78^2 = 22.85$
64	$64 - \overline{X} = -4.78$	$(X - \overline{X})^2 = -4.78^2 = 22.85$
64	$64 - \overline{X} = -4.78$	$(X - \overline{X})^2 = -4.78^2 = 22.85$
57	$57 - \overline{X} = -11.78$	$(X - \overline{X})^2 = -11.78^2 = 138.77$
56	$56 - \overline{X} = -12.78$	$(X - \overline{X})^2 = -12.78^2 = 163.33$
$\Sigma = 619.00$	$\Sigma = 0$ **(basically)**	$\Sigma = 927.56$

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}} = \sqrt{\frac{927.56}{9 - 1}} = \sqrt{\frac{927.56}{8}} = \sqrt{115.94} = 10.77$$

The standard deviation is 10.77 points.

Okay, here's where it gets tricky (trickier?)! Now that we have the standard deviation (and the measures of central tendency), we can make predictions about the shape of the frequency distribution.

Example 4.6.4

1. Based on the mean, is this standard deviation small, medium or large?
2. Based on the size of the standard deviation, what shape does this suggest that the frequency distribution will be?

Solution

Let's start with deciding if the standard deviation is large. To determine that, we find out if the standard deviation of 10.77 points is larger than a quarter or a third of the mean of 68.78 points. Rounding to two decimal points, we find:

$$\frac{68.78}{4} = 17.20$$

$$\frac{68.78}{3} = 22.93$$

No, this is not a large standard deviation because 10.77 is not bigger than 17.20 or 22.93.

Our next try is to see if the standard deviation of 10.77 points is smaller than about 10% of the mean of 68.78 points. Rounding to two decimal points, we find:

$$68.78 \times 0.10 = 6.88$$

No, this is not a small standard deviation because 10.77 is not smaller than 6.88.

That means that the standard deviation is medium!

A medium standard deviation suggests a mesokurtic frequency distribution, or a bell-shaped curve.

Okay, now we have one set of numbers (measures of central tendency) suggesting a skewed distribution and the standard deviation suggesting a bell-shaped curve. Which is it? Let's see!

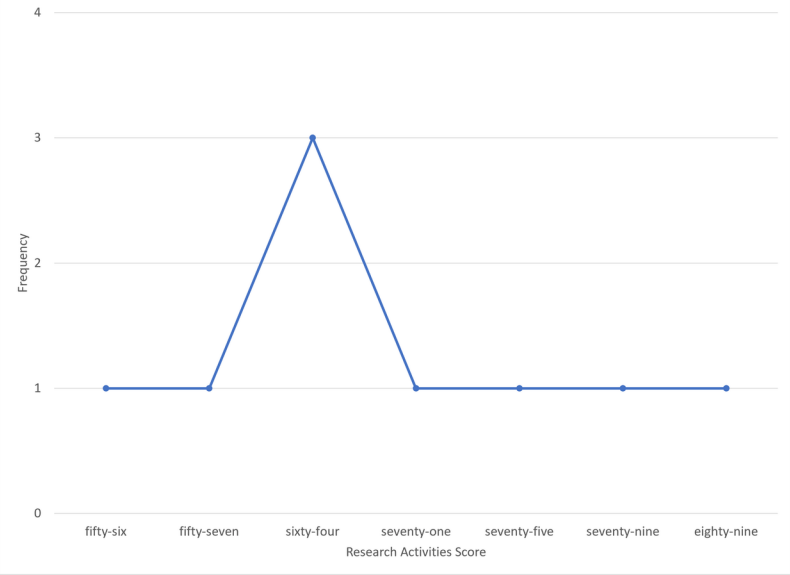


Figure 4.6.1: Frequency Line Graph of Face-to-Face Class's Research Activities Scores. (CC-BY-SA Michelle Oja via personal records)

It looks like the measures of central tendency were slightly better at predicting the shape of the graph than the standard deviation. As you'll learn in the next section on distributions, these predictions will be more accurate the more scores that we have, but we used data with a sample size of only 9 so our predictions were not robust.

Table 4.6.2 provides another set of points earned on Research Activities, but this time for an online class. The sum of these 12 scores is also provided in the last row. Are you ready? Yes, you are! Go through the following Exercises to see how well you can put it all together.

Table 4.6.2- Research Activities scores for 12 students in an online class, which sums to 967

Research Activities Score for Online Class
99
93
92
88

Research Activities Score for Online Class
88
83
79
77
76
69
67
56
$\sum = 967$

What do we start with? Yes, let's calculate the three measures of central tendency.

Exercise 4.6.1

1. What is the mode for the Research Activities points in Table 4.6.2?
2. What is the median for the Research Activities points in Table 4.6.2?
3. What is the mean for the Research Activities points in Table 4.6.2?

Answer available at the end of this section.

Once you have these three numbers that all represent the center of the frequency distribution, you can use them to make predictions about the shape of the distribution.

Exercise 4.6.2

Based on the three measures of central tendency, what do you think the shape of the frequency distribution might be?

Answer available at the end of this section.

We can now calculate the standard deviation.

Example 4.6.3

What is the standard deviation for the Research Activities scores from Table \({4{.}6{.}\eatSpaces{2} \endSpaces}\?.

Solution

You know how to do this!

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

Use Table 4.6.2 to calculate the numerator.

Table 4.6.2- Find the mean difference, squared mean difference, and the sum of squares in this table.

X	$X - \overline{X}$	**Sums of Squares** ($(X - \overline{X})^2$)
99		
93		
92		
88		
88		
83		
79		
77		
76		
69		
67		
56		

X	$X - \overline{X}$	**Sums of Squares** ($(X - \overline{X})^2$)
$\Sigma = 967.00$	$\Sigma = 0$ **(basically)**	$\Sigma = 1698.92$

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}} = \sqrt{\frac{1698.72}{12 - 1}} = \sqrt{\frac{1698.92}{11}} = \sqrt{154.45} = 12.43$$

The standard deviation is 12.43 points.

Using the standard deviation (and the measures of central tendency), make additional predictions about the shape of the frequency distribution.

Exercise 4.6.4

1. Based on the mean, is this standard deviation small, medium or large?
2. Based on the size of the standard deviation, what shape does this suggest that the frequency distribution will be?

Answer available at the end of this section.

Again, we have one set of numbers (measures of central tendency) suggesting a skewed distribution and the standard deviation suggesting a bell-shaped curve.

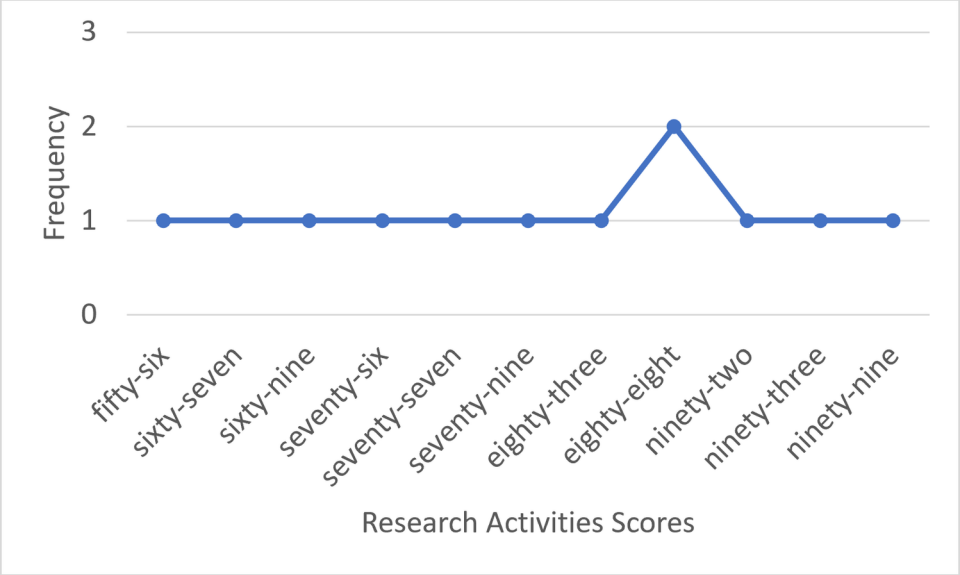


Figure 4.6.2: Frequency Line Graph of Online Class's Research Activities Scores. (CC-BY-SA Michelle Oja via personal records)

It looks like the measures of central tendency were slightly better at predicting the shape of the graph again! Figure 4.6.2 is negatively skewed.

Next, let's learn more about different kinds of distributions and theories that may explain why our predictions based on the measures of central tendencies didn't match our predictions based on the standard deviation.

Contributor

- Dr. MO (Taft College)

Answers

Exercise 4.6.1

1. The mode is 88 points as that is the only number that repeated, so it is the most frequent score.
2. The median is also 81 points because that is the number that would be in the exact middle if there was one number in the exact middle. This is the score that would be between 83 points and 79 points.
3. The mean is rounded to 80.58 points.

$$\overline{X} = \frac{\sum X}{n} = \frac{967.00}{12} = 80.58$$

Exercise 4.6.2

The mode will be the peak of the distribution, at 88 points. The median is a little lower than the mode, showing that the middle score in the whole distribution is lower than the most frequent score. The mean is a little lower still, at 80.58 points. This suggests a negatively skewed frequency distribution, especially as the mode is a much higher score.

Exercise 4.6.4

Let's start with deciding if the standard deviation is large. To determine that, we find out if the standard deviation of 12.43 points is larger than a quarter or a third of the mean of 80.58 points.

$$\frac{80.58}{4} = 20.15$$

$$\frac{80.58}{3} = 26.86$$

No, this is not a large standard deviation because 12.43 is not bigger than 20.15 or 26.86.

Our next try is to see if the standard deviation of 12.43 points is smaller than about 10% of the mean of 80.58 points.

$$80.58 \times 0.10 = 8.06$$

No, this is not a small standard deviation because 12.43 is not smaller than 8.06. That means that the standard deviation is medium!

A medium standard deviation suggests a mesokurtic frequency distribution, or a bell-shaped curve.

This page titled 4.6: Putting All of the Measures Together is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 2.7: Skewness and the Mean, Median, and Mode by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

5: DISTRIBUTIONS

5.1: Introduction to Distributions
5.2: Introduction to Probability
5.3: The Binomial Distribution
5.4: The Law of Large Numbers
5.5: Three Important Distributions Together
5.6: Sampling Distributions and the Central Limit Theorem
5.7: Putting it All Together

This page titled 5: Distributions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

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 available at the end of this section.

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)

Answers

Exercise 5.1.1

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.

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: Introduction to 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.

Historical Context

A Chinese diagram entitled "The Old Method Chart of the Seven Multiplying Squares" from 1303 shows the binomial coefficients up to the eighth power – the numbers that are fundamental to the mathematics of probability (Champkin, 2014).



Figure 5.2.1: Yang Hui's triangle appears in Jade Mirror of the Four Unknowns, a mathematical work by Zhu Shijie, from the Wikipedia page on Pascal's Triangle, Public Domain

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. 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.

Reference

Champkin, J. (2014, January 24). *The timeline of statistics*. Significance. https://www.statsref.com/timeline.pdf

Contributors

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

This page titled 5.2: Introduction to 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 Binomial Distribution

The vast majority of the content in this book relies on one of five distributions: the binomial distribution, the normal distribution, the t distribution, the F distribution, and the χ^2 ("chi-square") distribution. Let's start with the binomial distribution, since it's the simplest of the five. The normal distribution will be discussed later this chapter, and the others will be mentioned in other chapters. Our discussion of the binomial distribution will start with figuring out some probabilities of events happening.

Let's imagine a simple "experiment": in my hot little hand I'm holding 20 identical six-sided dice. On one face of each die there's a picture of a skull; the other five faces are all blank. If I proceed to roll all 20 dice, what's the probability that I'll get exactly 4 skulls?

Let's unpack that last sentence. First, in case you missed it, rolling of the 20 dice is your sample. N = 20 (20 scores in this distribution). Second, it's asking about a probability, so we'll be looking back at the last section and realizing that we should do some division. Finally, this is a question, some might even call it a Research Question.

Okay, back to the "experiment." Assuming that the dice are fair, we know that the chance of any one die coming up skulls is 1 in 6; to say this another way, the skull probability for a single die is approximately .1667, or 16.67%.

$$\frac{1}{6} = 0.166\overline{7} * 100 = 16.67$$

The probability of rolling a skull with one die is 16.67%, but I'm doing this 20 times so that's

$0.166\overline{7} * 20 = 3.33$ skulls

So, if I rolled 20 dice, I could expect about 3 of them to come up skulls. But that's not quite the Research Question is it? Let's have a look at how this is related to a binomial distribution.

Figure $5.3.1$ plots the binomial probabilities for all possible values for our dice rolling experiment, from X=0 (no skulls) all the way up to X=20 (all skulls). On the horizontal axis we have all the possible events (the number of skulls coming up when all 20 dice are rolled), and on the vertical axis we can read off the probability of each of those events. If you multiple these by 100, you'd get the probability in a percentage form. Each bar depicts the probability of one specific outcome (i.e., one possible value of X). Because this is a probability distribution, each of the probabilities must be a number between 0 and 1 (or 0% to 100%), and the heights of the all of the bars together must sum to 1 as well. Looking at Figure $5.3.1$, the probability of rolling 4 skulls out of 20 times is about 0.20 (the actual answer is 0.202036), or 20.22%. In other words, you'd expect to roll exactly 4 skulls about 20% of the times you repeated this experiment. This means that, if you rolled these 20 dice for 100 different repetitions, you'd get exactly 4 skulls in about 20 of your attempts.

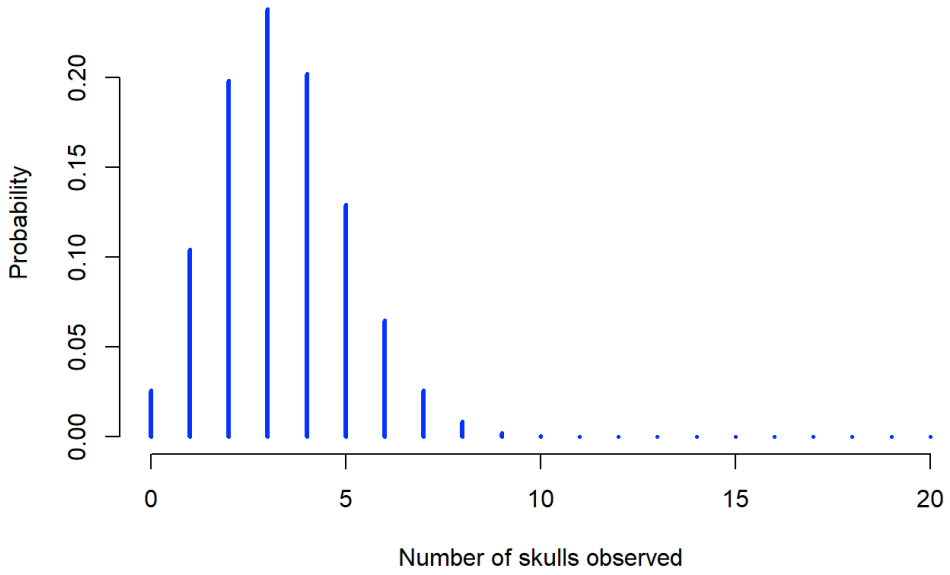


Figure 5.3.1- Probability of Rolling 4 Skulls out of 20 Dice with One Skull (CC-BY-SA Danielle Navarro from Learning Statistics with R)

Sample Size Matters

We'll be talking a lot about how sample size (N) affects distributions in this chapter, starting now!

To give you a feel for how the binomial distribution changes when we alter the probability and N, let's suppose that instead of rolling dice, I'm actually flipping coins. This time around, my experiment involves flipping a fair coin repeatedly, and the outcome that I'm interested in is the number of heads that I observe. In this scenario, the success probability is now 1/2 (one out of two options). Suppose I were to flip the coin N=20 times. In this example, I've changed the success probability (1 out of 2, instead of 1 out of 6 in the dice example), but kept the size of the experiment the same (N=20). What does this do to our binomial distribution? Well, as Figure 5.3.2 shows, the main effect of this is to shift the whole distribution higher (since there's more chance of getting a heads (one out of two options) than a 4 (one out of six options). Okay, what if we flipped a coin N=100 times? Well, in that case, we get Figure 5.3.3. The distribution stays roughly in the middle, but there's a bit more variability in the possible outcomes (meaning that there are more extreme scores); with more tosses, you are more likely to get no heads but also more likely to get all heads than if you only flipped the coin 20 times.

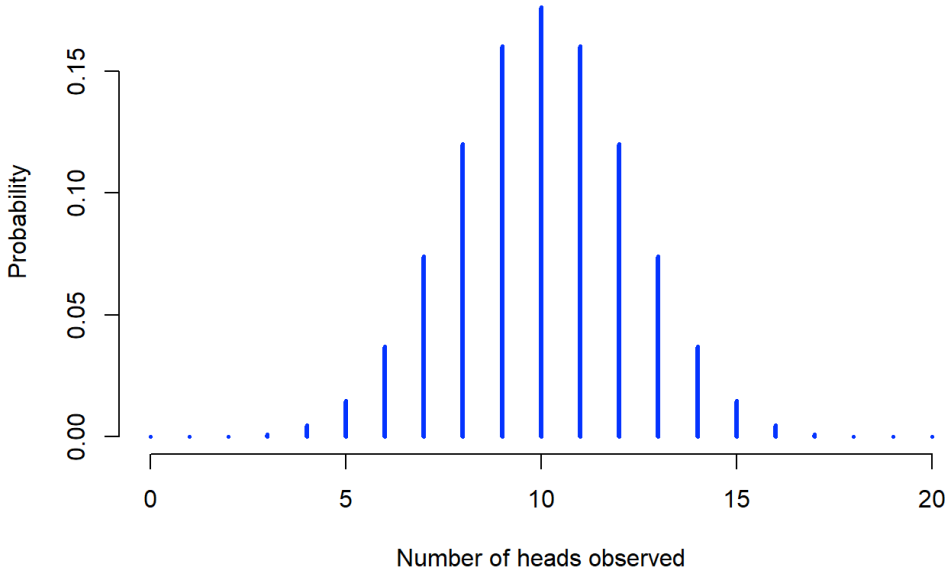


Figure 5.3.2- Probability of Heads from Flipping a Fair Coin 20 Times (CC-BY-SA Danielle Navarro from Learning Statistics with R)

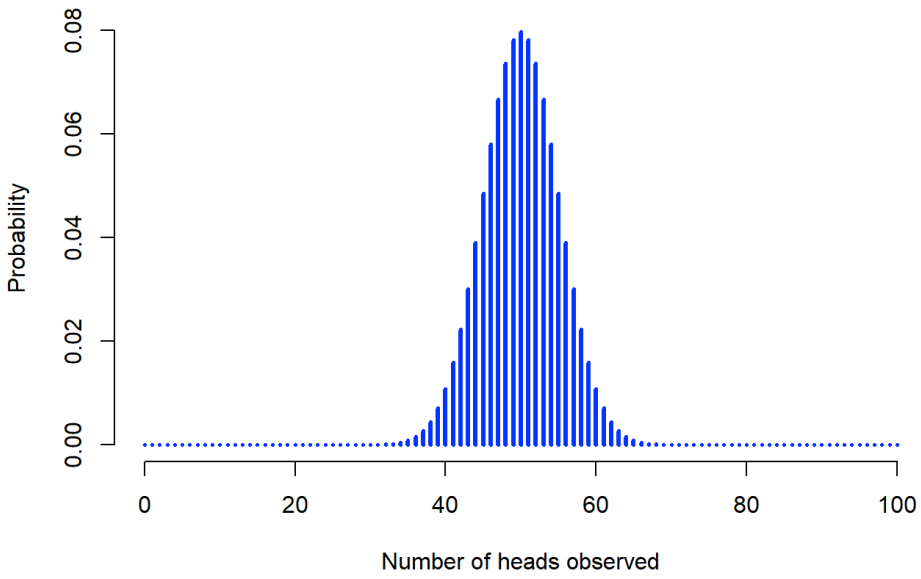


Figure 5.3.3- Probability of Heads from Flipping a Fair Coin 100 Times (CC-BY-SA Danielle Navarro from Learning Statistics with R)

And that's it on binomial distributions! We are building understanding about distributions and sample size, so nothing too earth-shattering here. One thing to note is that both the coin flip and the dice toss were measured as discrete variables. Our next distributions will cover continuous variables. The difference is that discrete variables can be one or the other (a heads or a tails, a four or not-four), while continuous variables can have gradations shown as decimal points. Although the math works out to 3.33 skulls when you throw 30 dice, you can't actually get 3.33 skulls, you can only get 3 or 4 skulls.

We're building our knowledge, so keep going!

Contributors and Attributions

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

This page titled 5.3: The Binomial Distribution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

5.4: 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 ($\overline{X} \rightarrow \mu$).

Historical Context

In 1713, Jacob Bernoulli's Ars conjectandi derives the law of large numbers – the more often you repeat an experiment, the more accurately you can predict the result (Champkin, 2014).



Figure 5.4.1: Cover Page for Ars conjectandi from Wikipedia. Public Domain

Dr. Navarro 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 ($\sqrt{\frac{\sum (X - \overline{X})^2}{N-1}}$) is different from the formula for the standard deviation of the population ($\sqrt{\frac{\sum (X - \overline{X})^2}{N}}$).

Reference

Champkin, J. (2014, January 24). *The timeline of statistics*. Significance. https://www.statsref.com/timeline.pdf

Contributors

- Danielle Navarro (University of New South Wales)

- Dr. MO (Taft College)

This page titled 5.4: 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.

5.5: Three Important Distributions Together

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.

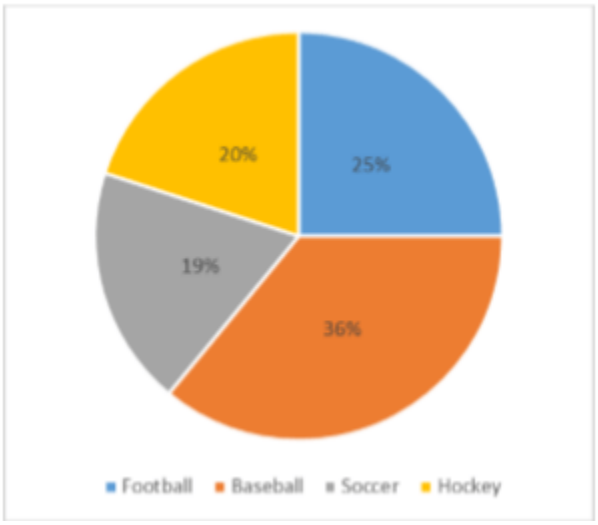


Figure 5.5.1: Favorite Sports (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Take, for example, the pie chart in Figure 5.5.1 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 even 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, and 2) the total area of the chart is equal to 100% or 1.0, which makes the size of the slices interpretable.

Normal Distributions

Historical Context

Galileo Galilei anticipated the normal distribution in his *Dialogue Concerning the Two Chief World Systems—Ptolemaic and Copernican*, which was published in 1632 (Encyclopedia.com).



Figure 5.5.1: A copy of the *Dialogo*, Florence edition, located at the Tom Slick rare book collection at Southwest Research Institute, in Texas by Nightryder84 - Own work, CC BY-SA 3.0

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 5.5.1 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 5.5.1).

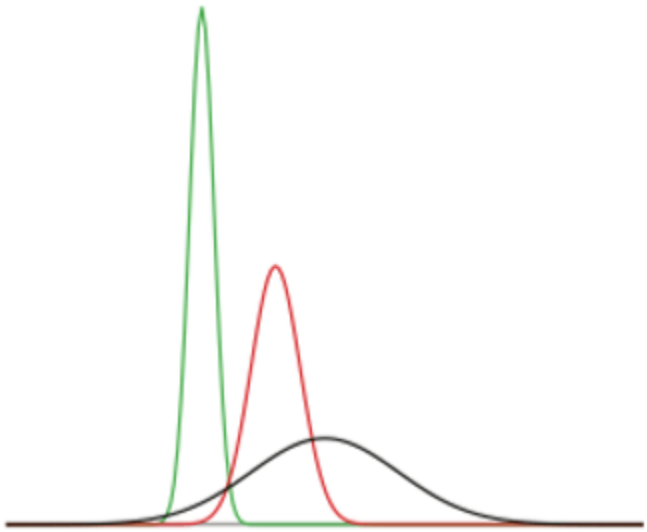


Figure 5.5.1: Symmetrical Distributions with Different Means and Standard Deviations (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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.

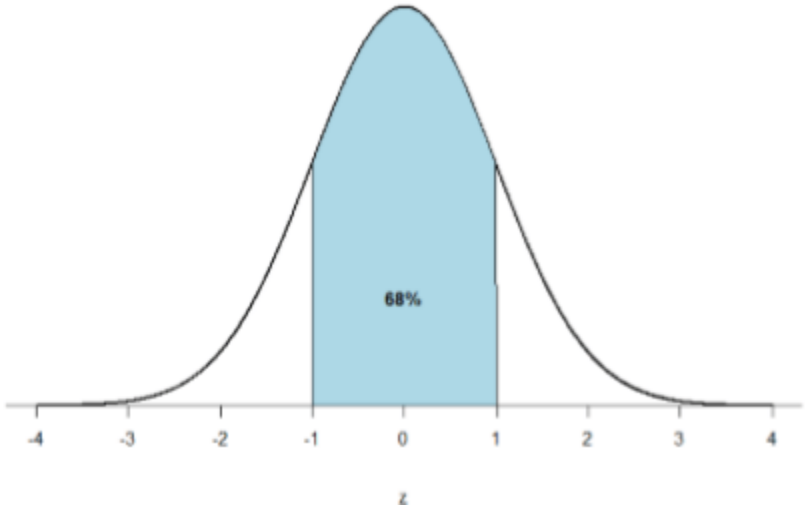


Figure 5.5.2: There is a 68% chance of selection a Score from the shaded region (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 5.5.3, and a total of 5% in the shaded areas. Since there are two shaded areas, that's about 2.5% on each side.

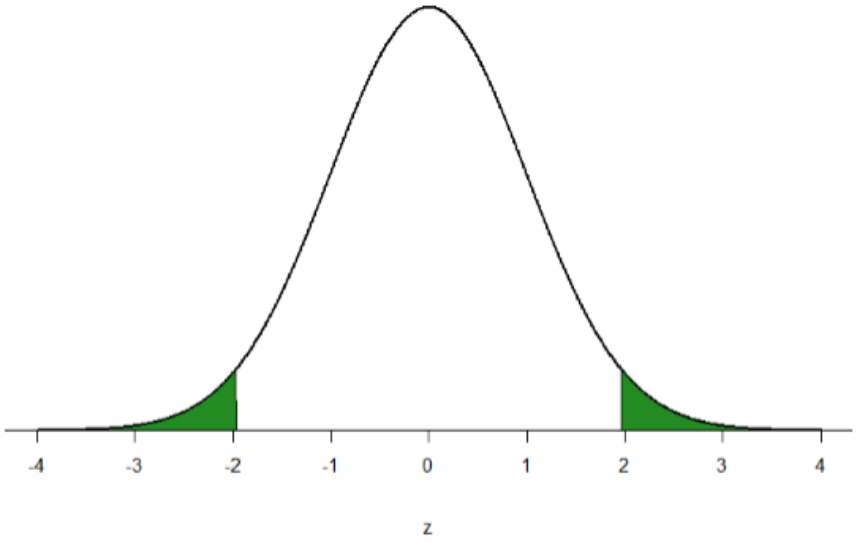


Figure 5.5.3: Area Beyond 2 Standard Deviations (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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*: We know through the Law of Large Numbers that we have enough data, then our sample will begin to become more and more "normal' (have the characteristics of a normally distributed population, including the important characteristics above).
- *Probability Distributions*: We know that we can make predictions from probability distributions. We can use probability distributions to understand the likelihood of specific events happening.
- *Standard Normal Curve*: We 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".)

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, one more piece, and then we put it all together to see why all of this information about the properties of distributions is important.

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

5.6: Sampling Distributions and the Central Limit Theorem

The law of large numbers is a very powerful tool, but it's not going to be good enough to answer all our questions. Among other things, all it gives us is a "long run guarantee". In the long run, if we were somehow able to collect an infinite amount of data, then the law of large numbers guarantees that our sample statistics will be correct. But as John Maynard Keynes famously argued in economics, a long run guarantee is of little use in real life:

> *[The] long run is a misleading guide to current affairs. In the long run we are all dead. Economists set themselves too easy, too useless a task, if in tempestuous seasons they can only tell us, that when the storm is long past, the ocean is flat again (*Keynes, 1923).

As in economics, so too in psychology and statistics. It is not enough to know that we will *eventually* arrive at the right answer when calculating the sample mean. Knowing that an infinitely large data set will tell me the exact value of the population mean is cold comfort when my *actual* data set has a sample size of N=100. In real life, then, we must know something about the behavior of the sample mean when it is calculated from a more modest data set!

Sampling Distribution of the Mean

With this in mind, let's abandon the idea that our studies will have sample sizes of 10,000, and consider a very modest experiment indeed. This time around we'll sample N=5 people and measure their IQ scores. While the IQ test has many validity issues (e.g., research showing cultural bias; see Gould, 1996), it is a good example of distribution because it is normed to a mean of 100 and a standard deviation of 15.

In a simulated study, the mean IQ in this sample turns out to be exactly 95. Not surprisingly, this is much less accurate than the previous experiment with 10,000 simulated IQ scores. Now imagine that Dr. Navarro decided to replicate the experiment. That is, she repeated the procedure as closely as possible: she could randomly sample 5 new people and measure their IQ. On a second simulation, the mean IQ in my sample is 101. When Dr. Navarro repeated the simulation 10 times, she obtained the results shown in Table 5.6.1. As you can see, the sample mean varies from one replication to the next.

Table 5.6.1- 10 Replications of Simulated IQ Scores, Each with a Sample Size of N=5 (CC-BY-SA Danielle Navarro)

Replication Time	Bindi	Coen	Koa	Kylie	Mia	Mean
Replication 1	90	82	94	99	110	**95.0**
Replication 2	78	88	111	111	117	**101.0**
Replication 3	111	122	91	98	86	**101.6**
Replication 4	98	96	119	99	107	**103.8**
Replication 5	105	113	103	103	98	**104.4**
Replication 6	81	89	93	85	114	**92.4**
Replication 7	100	93	108	98	133	**106.4**
Replication 8	107	100	105	117	85	**102.8**
Replication 9	86	119	108	73	116	**100.4**
Replication 10	95	126	112	120	76	**105.8**

Now suppose that Dr. Navarro decided to keep going in this fashion, replicating this "five IQ scores" simulation over and over again. Every time she replicate the simulation, she write down the sample mean. Over time, she would be amassing a new data set, in which every experiment generates a single data point, a single mean. What if Dr. Navarro continued like this for 10,000 replications, and then drew a histogram? Using the magical powers of statistical software, that's exactly what Dr. Navarro did, and you can see the results in Figure 5.6.1. As this picture illustrates, the average of 5 IQ scores is usually between 90 and 110. If you sample 5 people at random and calculate their *average* IQ, you'll almost certainly get a number between 80 and 120, even though there are quite a lot of individuals who have IQs above 120 or below 80. For comparison, the black line plots the population

distribution of IQ scores. But more importantly, what Figure 5.6.1 highlights is that if we replicate an experiment over and over again, what we end up with is a *distribution* of sample means! This distribution has a special name in statistics: it's called the ***sampling distribution of the mean***.

Sampling distributions are another important theoretical idea in statistics, and they're crucial for understanding the behavior of small samples. For instance, when Dr. Navarro ran the very first "five IQ scores" experiment, the sample mean turned out to be 95. If she repeats the experiment, the sampling distribution tells me that we can expect to see a sample mean anywhere between 80 and 120.

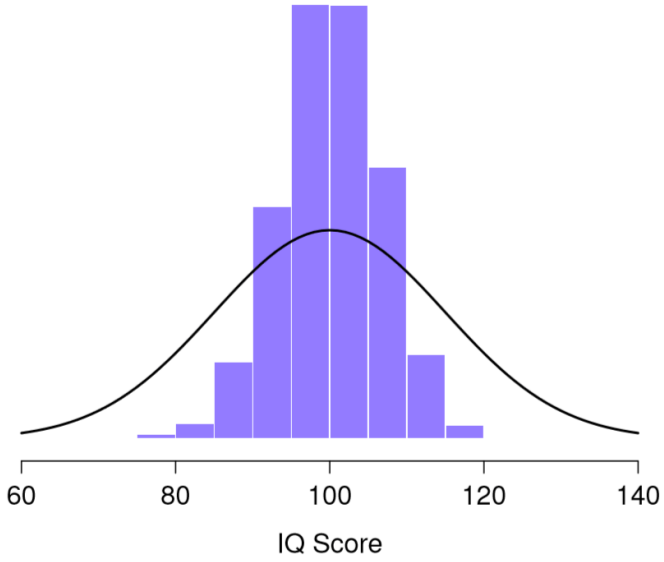


Figure 5.6.1- Distribution of Means from 10,000 Samples of 5 IQ Scores (CC-BY-SA Danielle Navarro from Learning Statistics with R)

Not Just Distribution of Sample Means

One thing to keep in mind when thinking about sampling distributions is that any sample statistic you might care to calculate has a sampling distribution. For example, suppose that each time Dr. Navarro replicated the "five IQ scores" experiment but wrote down the largest IQ score in the experiment. This would give her a data set that started out like in Figure 5.6.2:

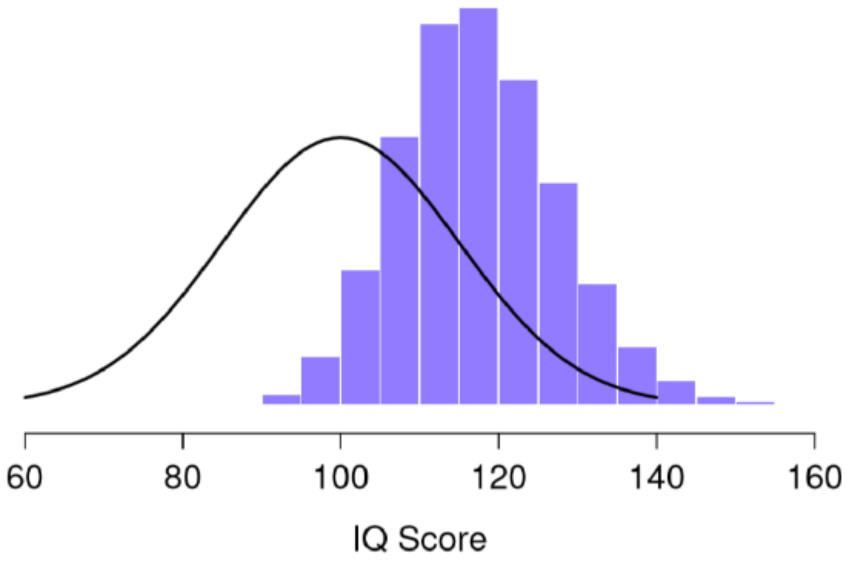


Figure 5.6.2- Distribution of Highest IQ from 10,000 Samples of 5 IQ Scores (CC-BY-SA Danielle Navarro from Learning Statistics with R)

Doing this over and over again would give Dr. Navarro a very different sampling distribution, namely the *sampling distribution of the maximum*. The sampling distribution of the maximum of 5 IQ scores is shown in Figure 5.6.2. Not surprisingly, if you pick 5 people at random and then find the person with the highest IQ score, they're going to have an above average IQ. As shown in Figure 5.6.2, most of the time you'll end up with someone whose IQ is measured in the 100 to 140 range.

Central Limit Theorem

An illustration of the how sampling distribution of the mean depends on sample size. In each panel, Dr. Navarro generated 10,000 samples of IQ data, and calculated the mean IQ observed within each of these data sets. The histograms in these plots show the distribution of these means (i.e., the sampling distribution of the mean). Each individual IQ score was drawn from a normal distribution with mean 100 and standard deviation 15, which is shown as the solid black line.

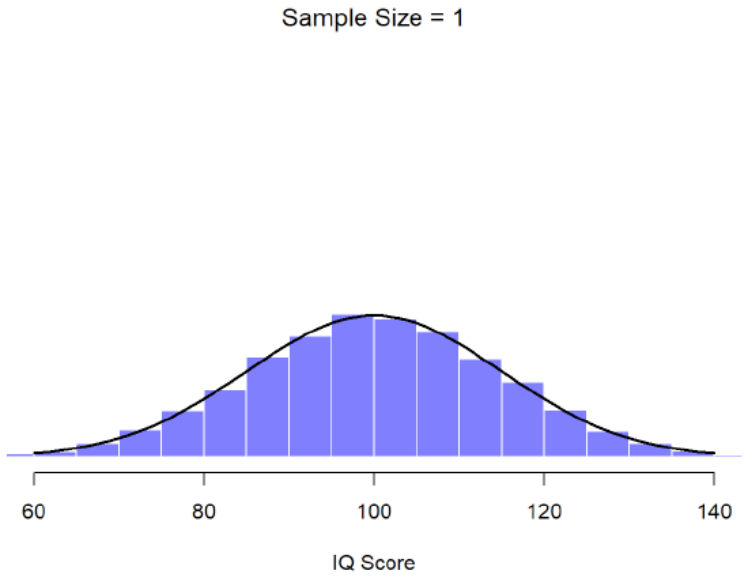


Figure 5.6.3- Distribution of Mean IQ from 10,000 Samples of 1 IQ Score (CC-BY-SA Danielle Navarro from Learning Statistics with R)

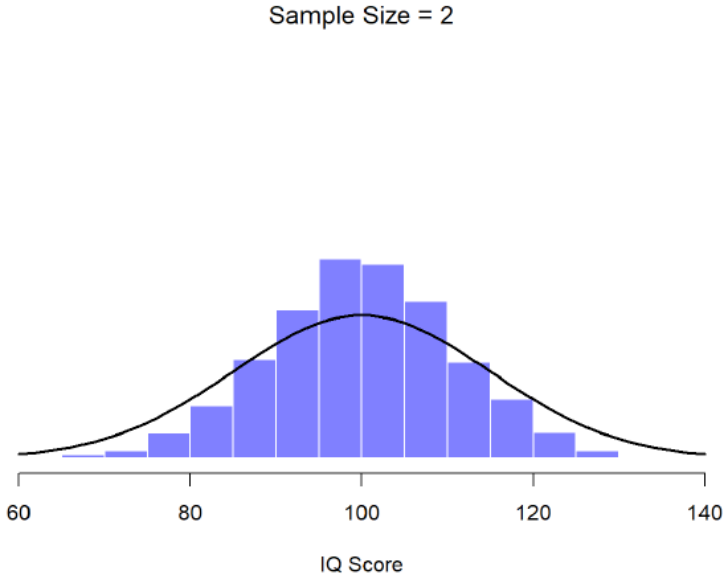


Figure 5.6.4- Distribution of Mean IQ from 10,000 Samples of 2 IQ Scores (CC-BY-SA Danielle Navarro from Learning Statistics with R)

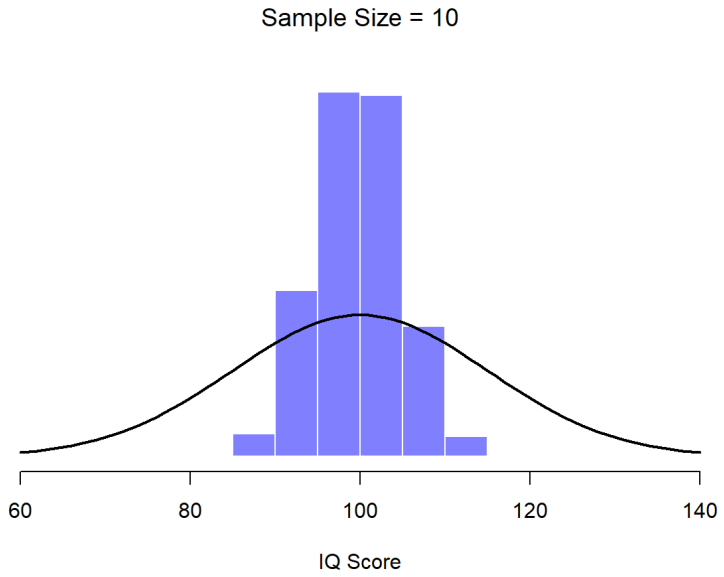


Figure 5.6.5- Distribution of Mean IQ from 10,000 Samples of 10 IQ Scores (CC-BY-SA Danielle Navarro from Learning Statistics with R)

In Figure 5.6.3, each data set contained only a single observation, so the mean of each sample is just one person's IQ score. As a consequence, the sampling distribution of the mean is of course identical to the population distribution of IQ scores because there are 10,000 individual scores. When we raise the sample size to 2, as in Figure 5.6.4, the mean of any one sample tends to be *closer* to the population mean than any one person's IQ score, and so the histogram (i.e., the sampling distribution) is a bit narrower than the population distribution. By the time we raise the sample size to 10 (Figure 5.6.5), we can see that the distribution of sample means tend to be fairly tightly clustered around the true population mean.

Sample Size Matters

At this point I hope you have a pretty good sense of what sampling distributions are, and in particular what the sampling distribution of the mean is. This section talks about how the sampling distribution of the mean changes as a function of sample size. Intuitively, you already know part of the answer: if you only have a few observations, the sample mean is likely to be quite inaccurate: if you replicate a small experiment and recalculate the mean you'll get a very different answer. In other words, the sampling distribution is quite wide. If you replicate a large experiment and recalculate the sample mean you'll probably get the same answer you got last time, so the sampling distribution will be very narrow. You can see this visually in the Figures in this section: the bigger the sample size, the narrower the sampling distribution gets. We can quantify this effect by calculating the *standard deviation of the sampling distribution, which is referred to as the* ***standard error***. The standard error of a statistic is often denoted SE, and since we're usually interested in the standard error of the sample *mean*, we often use the acronym SEM. As you can see just by looking at the picture, as the sample size N increases, the SEM decreases.

Okay, so that's one part of the story. However, there's something I've been glossing over so far. All my examples up to this point have been based on the "IQ scores" simlulations, and because IQ scores are usually from a normal distribution, I've assumed that the population distribution is shaped like that, a symmetrical, medium bell-shaped curve. What if the population distribution isn't normally distributed? What happens to the sampling distribution of the mean? The remarkable thing is this: no matter what shape your population distribution is, as N increases the sampling distribution of the mean starts to look more like a normal distribution.

To give you a sense of this, Dr. Navarro ran some simulations . To do this, she started with the "ramped" distribution shown in the histogram in Figure 5.6.6. As you can see by comparing the triangular shaped histogram to the bell curve plotted by the black line, the population distribution doesn't look very much like a normal distribution at all. Next, Dr. Navarro simulated the results of a large number of experiments. In each experiment she took N=2 samples from this "ramp" distribution, and then calculated the sample mean. Figure 5.6.7 plots the histogram of these sample means (i.e., the sampling distribution of the mean for N=2). This time, the histogram produces a $\cap$-shaped distribution (reverse-U shaped distribution): it's still not normal, but it's a lot closer to the black line than the population distribution in Figure 5.6.6. When Dr. Navarro increases the sample size to N=4, the sampling distribution of the mean is very close to normal (Figure 5.6.8, and by the time we reach a sample size of N=8 (Figure 5.6.9) it's almost perfectly symmetrical and bell-shaped!. As you can see, even though the original population distribution is non-normal, the sampling distribution of the mean becomes pretty close to normal by the time you have a sample of even 4 observations. In

other words, as long as your sample size isn't tiny, the sampling distribution of the mean will be approximately normal no matter what your population distribution looks like!

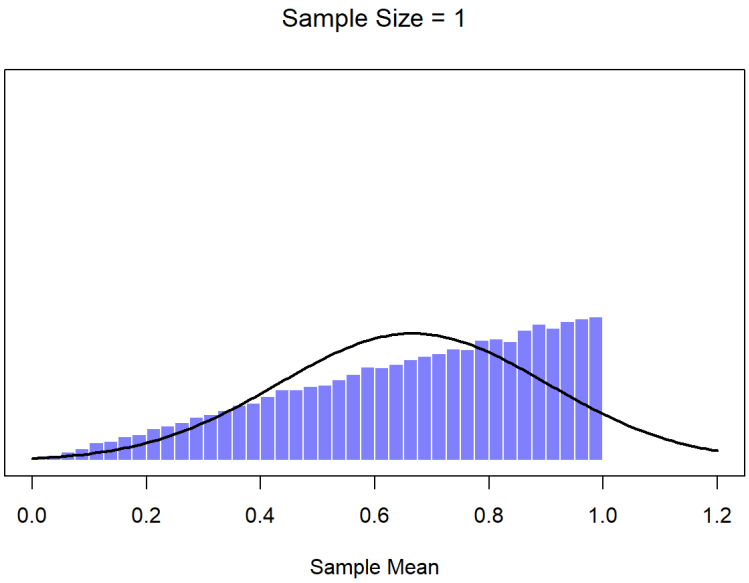


Figure 5.6.6- Ramp-Shaped Distribution 10,000 Samples of 1 Person (CC-BY-SA Danielle Navarro from Learning Statistics with R)

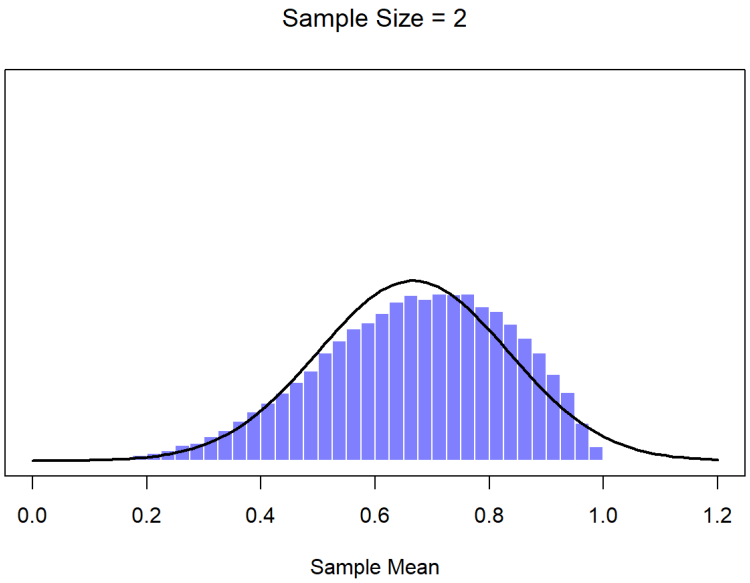


Figure 5.6.7- Distribution 10,000 Samples of 2 People (CC-BY-SA Danielle Navarro from Learning Statistics with R)

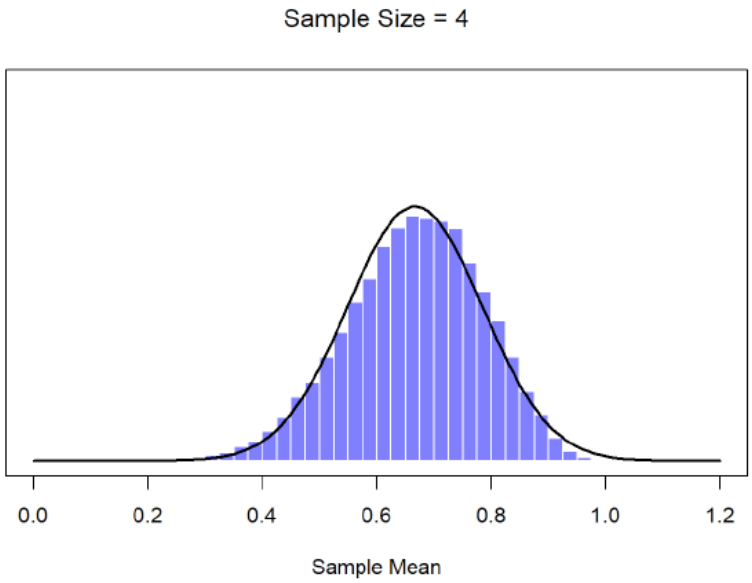


Figure 5.6.8- Distribution 10,000 Samples of 4 People (CC-BY-SA Danielle Navarro from Learning Statistics with R)

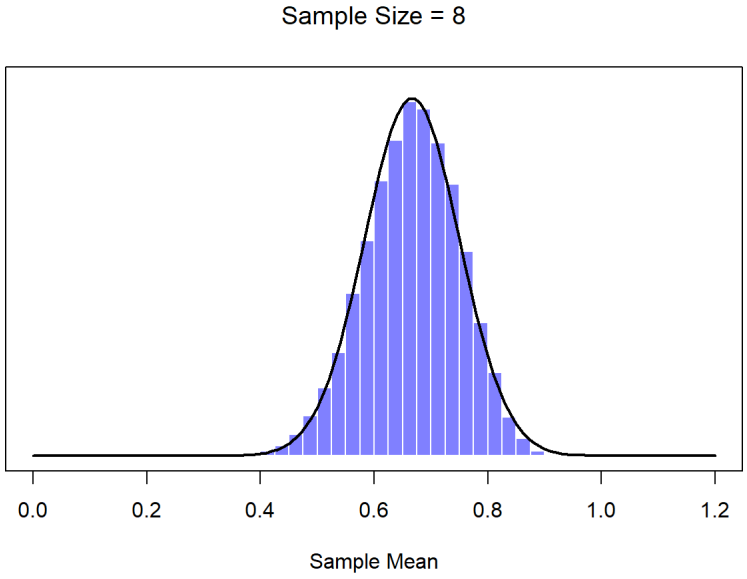


Figure 5.6.9- Distribution 10,000 Samples of 8 People (CC-BY-SA Danielle Navarro from Learning Statistics with R)

On the basis of these figures, it seems like we have evidence for all of the following claims about the sampling distribution of the mean:

- The mean of the sampling distribution will be VERY similar to the mean of the population.
- The standard deviation of the sampling distribution (i.e., the standard error) gets *smaller* (taller and narrower distribution) as the sample size *increases*.
- The shape of the sampling distribution becomes more like a normal distribution as the sample size increases.

As it happens, not only are all of these statements true, there is a very famous theorem in statistics that proves all three of them, known as the ***central limit theorem***. Among other things, the central limit theorem tells us that if the population distribution has mean μ and standard deviation σ, then the sampling distribution of the mean also has mean μ, and the standard error of the mean is

$$\text{SEM} = \frac{\sigma}{\sqrt{N}}$$

Because we divide the population standard deviation σ by the square root of the sample size N, the SEM gets smaller as the sample size increases. It also tells us that the shape of the sampling distribution becomes normal.

This result is useful for all sorts of things. It tells us why large experiments are more reliable than small ones, and because it gives us an explicit formula for the standard error it tells us *how much* more reliable a large experiment is. Pay attention to this formula for the standard error, it will come back! It tells us why the normal distribution is, well, *normal*. In real experiments, many of the things that we want to measure are actually averages of lots of different quantities (e.g., the IQ test may be an average of a large number of "specific" skills and abilities, although this is contested by cognitive theorists as well as research utilizing social identity comparisons), and when that happens, the averaged quantity should follow a normal distribution. Because of this mathematical law, the normal distribution pops up over and over again in real data.

References

Gould, S. J. (1996). The mismeasure of man (Rev. and expanded ed.). W. W. Norton.

Keynes, J. M. (1923). A tract on monetary reform. Macmillan Company.

Contributors and Attributions

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

This page titled 5.6: Sampling Distributions and the Central Limit Theorem is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Danielle Navarro.

5.7: 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

In Figure 5.7.1, 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.

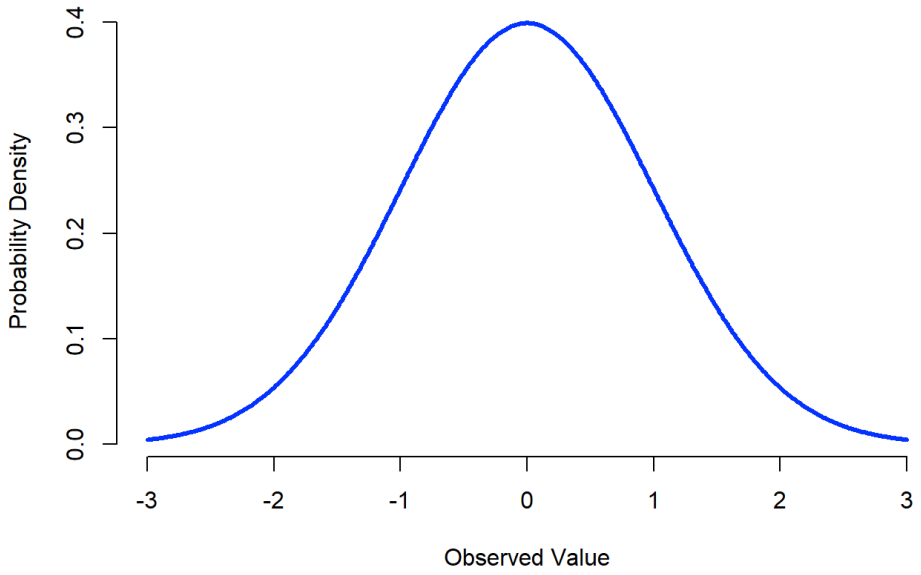


Figure 5.7.1- Standard Normal Curve (CC-BY-SA Danielle Navarro from Learning Statistics with R)

You can see where the name “bell curve” comes from in Figure 5.7.1 it looks a bit like a bell. Notice that, unlike the binomial distribution, the normal distribution in Figure 5.7.1 shows a smooth curve instead of “histogram-like” bars. This isn’t an arbitrary choice: the normal distribution is like a quantitative variable when the variable can be measured with decimals or fractions (called a continuous variable), whereas the binomial distribution is like a qualitative variable (called discrete variables). For instance, in the die rolling example, it was possible to get 3 skulls or 4 skulls, but impossible to get 3.9 skulls. Qualitative variables that are continuous don’t have this constraint. For instance, suppose we’re talking about the weather. The temperature on a pleasant spring day could be 68 degrees, 69.67 degrees, or even 72.222 degrees; temperature can be anything in between since temperature is a continuous variable, and so a normal distribution might be quite appropriate for describing Spring temperatures.

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

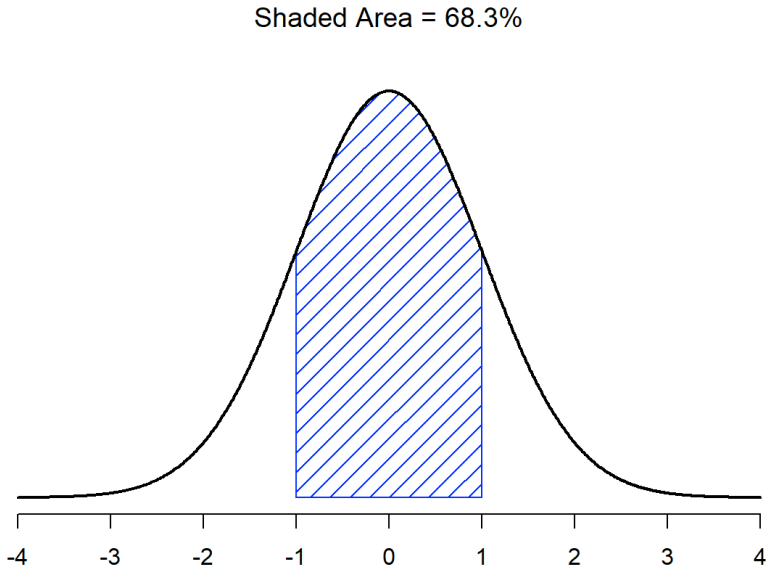


Figure 5.7.2- About 68% of Scores are One Standard Deviation Below through One Standard Deviation Above the Mean (CC-BY-SA Danielle Navarro from Learning Statistics with R)

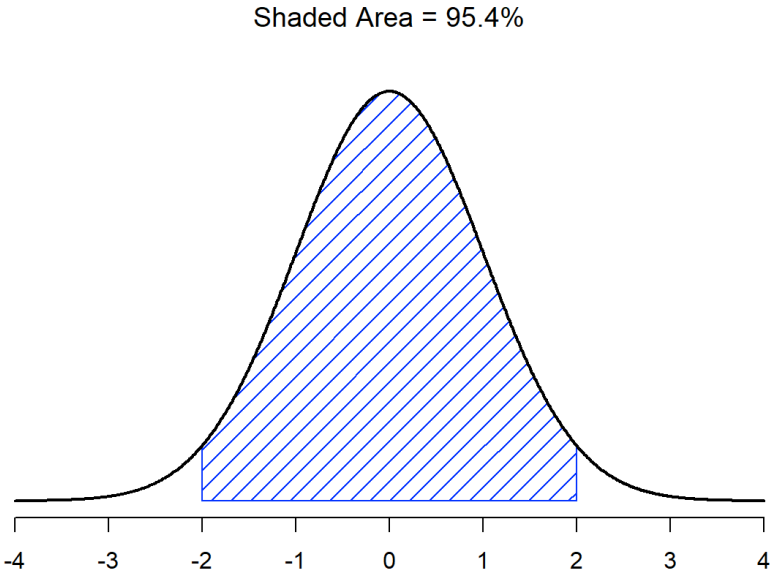


Figure 5.7.3- About 95% of Scores are Two Standard Deviations Below through Two Standard Deviations Above the Mean (CC-BY-SA Danielle Navarro from Learning Statistics with R)

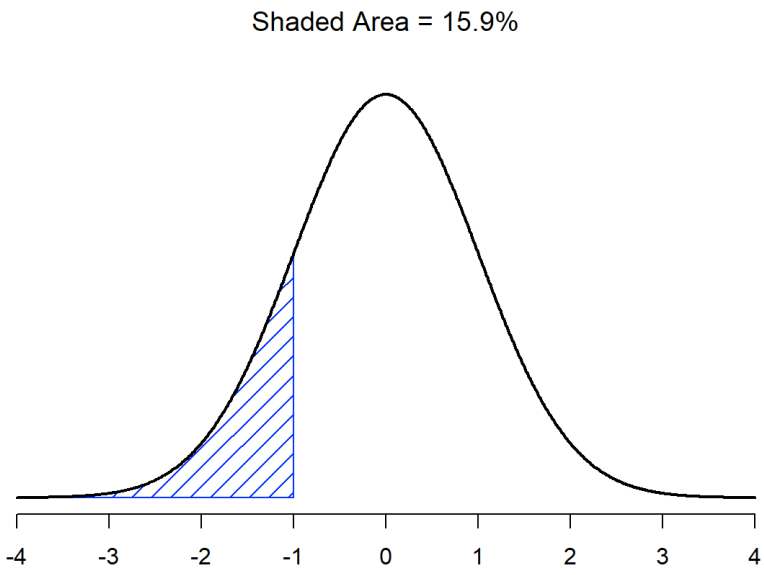


Figure 5.7.4- About 16% of Scores Should Be *Less Than* Two Standard Deviations Below the Mean (CC-BY-SA Danielle Navarro from Learning Statistics with R)

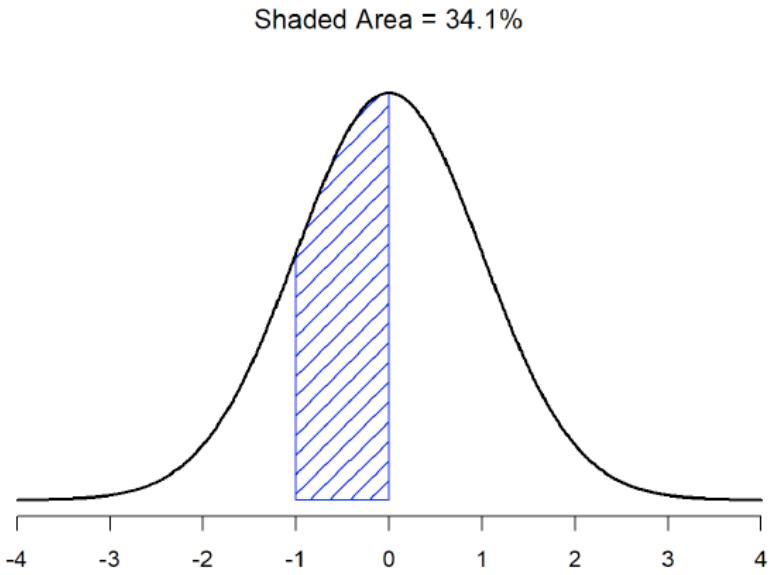


Figure 5.7.5- About 34% of Scores Should Be One Standard Deviations Below the Mean (CC-BY-SA Danielle Navarro from Learning Statistics with R)

There is a 68.3% chance (probability) that a score from any sample will fall within the shaded area of Figure 5.7.2, and a 95.4% chance that an observation will fall within two standard deviations of the mean (the shaded area in Figure 5.7.3). 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 5.7.5). 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*. Of course that also implies that there is a 50% chance that it falls above the mean.

For the interest of simplicity, let's round these percentages to whole numbers, as can be seen in (Figure 5.7.6).

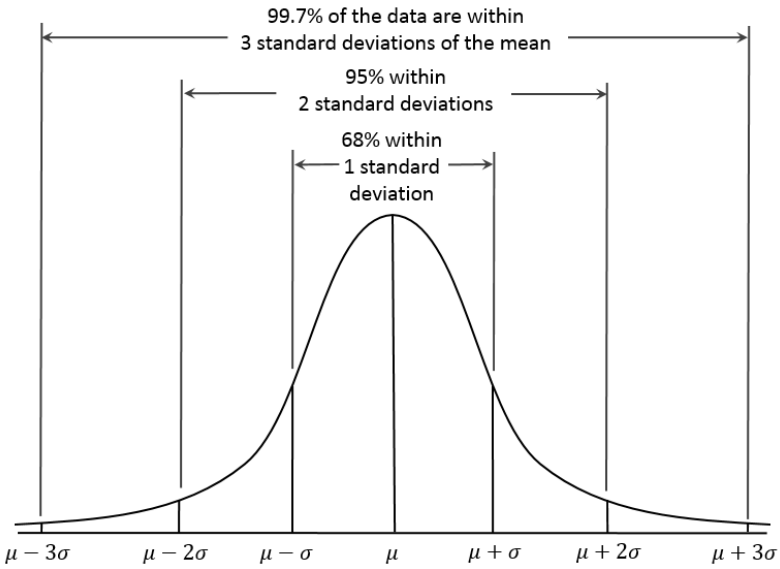


Figure 5.7.6: Empirical Rule. (CC-BY-SA; Dan Kernler from Wikipedia)

Figure 5.7.6 uses the symbols for a population; you've learned several ideas in this chapter that would explain why the Standard Normal Curve would have characteristics of the population! Let's start with the inner lines. Figure 5.7.6 shows that 68% of the scores fall within one standard deviation around the mean. On the x-axis, you can see that this 68% includes one population mean (μ) minus one population standard deviation (σ) as well as the mean plus one standard deviation. Figure 5.7.5 shows why this is the case: one standard deviation below the mean is about 34% of the whole distribution. As the Standard Normal Distribution is symmetrical, that means that one standard deviation above the means also should include about 34% of the scores. Adding those together (34% plus 34% = 68%). What Figure 5.7.6 illustrates is that we can predict the proportion of scores that should be one, two or three standard deviations around the mean for distributions that are normally distributed.

Because we know the proportion of scores that should be between the mean and each standard deviation, we can also predict the number of scores that we would expect to fall within that range for a new sample (if we knew the new sample size). Let's practice a little to see how this could be useful.

Table 5.7.1- Descriptive Statistics for Research Activity Points for Face-to-Face Class

Descriptive Statistic	Research Activities Points
N:	9
Mean:	68.78
Median:	64
Mode:	64
SD	10.78

We'll discuss a set of data that we've already covered, the Research Activities points for a face-to-face section behavioral statistics (Table 5.7.1). Everything that we've learned so far can be applied (again) to this data, so you could name the sample, such as "The

sample is nine students in a face-to-face section of behavioral statistics." Once you know the sample, you can name a possible population, such as "This sample can represent all face-to-face behavioral statistics sections." As we previously did, you can use the measures of central tendency (and standard deviation) to predict the shape of the frequency distribution.

> Based on these measures of central tendency and variability, do you think that this face-to-face section did well on their research activities, or not so much?

The research activities' measures of central tendency are pretty low (the mean, median, and mode are all a D grade), and the standard deviation is medium, so I don't think that the face-to-face class did too well on these research activities.

Let's get to some new ways that we can apply this information!

Example 5.7.1

1. What percentage of the Research Activities scores *should* be one standard deviation around the mean (one standard deviation below the mean and one standard deviation above the mean)?
2. What percentage of the Research Activities scores *should* be two standard deviations around the mean (two standard deviations below the mean and two standard deviations above the mean)?

Solution

You don't have to do math! You can just look at Figure 5.7.6.

1. Looking at Figure 5.7.6) 68% of scores should fall within one standard deviation around the mean.
2. Looking at Figure 5.7.6) 95% of scores should fall within one standard deviation around the mean.

To use the properties of the Standard Normal Curve, we first figure out what scores are one standard deviation around the mean. To find the scores that mark one standard deviation around the mean, we conduct two calculations; we first find the score that is one standard deviation below the mean (which is $\overline{X} + s$) and then finding the score that is one standard deviation above the mean ($\overline{X} + s$).

Example 5.7.2

1. What is the Research Activities score that is one standard deviation *below* the mean?
2. What is the Research Activities score that is one standard deviation *above* the mean?
3. What is the Research Activities score that is two standard deviations *below* the mean?
4. What is the Research Activities score that is two standard deviations *above* the mean?

Solution

1. The Research Activities score that is one standard deviation below the mean is 58.01 points because $\overline{X} - s = 68.75 - 10.77 = 58.01$.
2. The Research Activities score that is one standard deviation above the mean is 79.55 points because $\overline{X} + s = 68.75 + 10.77 = 79.55$.
3. The Research Activities score that is two standard deviations below the mean is 47.24 points. You can answer this in two mathematical ways:
    1. $\overline{X} - s - s = 68.75 - 10.77 - 10.77 = 47.24$
    2. $\overline{X} - 2s = 68.75 - (2 \times 10.77) = 68.75 - 21.54 = 47.24$
    3. These two formulas are mathematically identical, so choose the one that makes the most sense to you.
4. The Research Activities score that is two standard deviations above the mean is 90.32 points because
    1. $\overline{X} + s + s = 68.75 + 10.77 = 10.77 = 90.32$
    2. $\overline{X} - 2s = 68.75 + (2 \times 10.77) = 68.75 + 21.54 = 90.32$

This means that one standard deviation around the mean for this face-to-face sample is 58.01 to 79.55 points, and two standard deviations around the mean would be Research Activities scores of 47.24 points to 90.32 points. Combining that with the proportions in the Standard Normal Curve (see Figure 5.7.6), we can estimate the percentage of students who actually scored in these ranges.

In other words, we would predict (if our distribution of Research Activities scores was normally distributed) that 68% of scores for Research Activities would fall between 58.01 and 79.55, and that 95% of scores should fall between 47.24 points to 90.32 points. Just like when we used the measures of central tendency and standard deviation to predict the shape of a frequency distribution without looking at an actual frequency line graph, we can use the properties of the Standard Normal Distribution to get a sense of what the frequency line graph might look like.

We can add one more step to use this information to predict the *actual* number of students who would score one standard deviation around the mean for a *new* sample. For our example of Research Activities scores, this might be useful to estimate the number of students who would score a C or D on these assignments. Looking back at the example, we see that about 68% of scores should score between 58.01 and 79.55 points (which is almost a D to a C grade). If Dr. MO had a new class, we could figure out what 68% of that class would be:

$$\frac{68}{100} \times N$$

This converts the percentage to a decimal, then multiplies that by the total number of scores in the new sample.

Example 5.7.3

If Dr. MO used this information about the Research Activities scores from the face-to-face class of nine students, how many students should she expect to earn scores one standard deviation around the mean (near a D to a C grade) if she had a new class of 30 students?

Solution

$$\frac{68}{100} \times N = 0.68 \times 30 = 20.40$$

With a new class of 30 students, Dr. MO should expect about 20.40 students to score between 58.01 points to 79.55 points.

Unfortunately, our distribution was not normally distributed (see Figure 5.7.7), so these estimates were incorrect. We didn't do anything wrong mathematically or in our interpretation, though! It's just that small sample sizes tend to not be normally distributed (as you could predict from the Law of Large Numbers), so the Standard Normal Distribution doesn't apply well. This shows us that using the Standard Normal Curve to predict percentages is not *guaranteed* to be accurate.

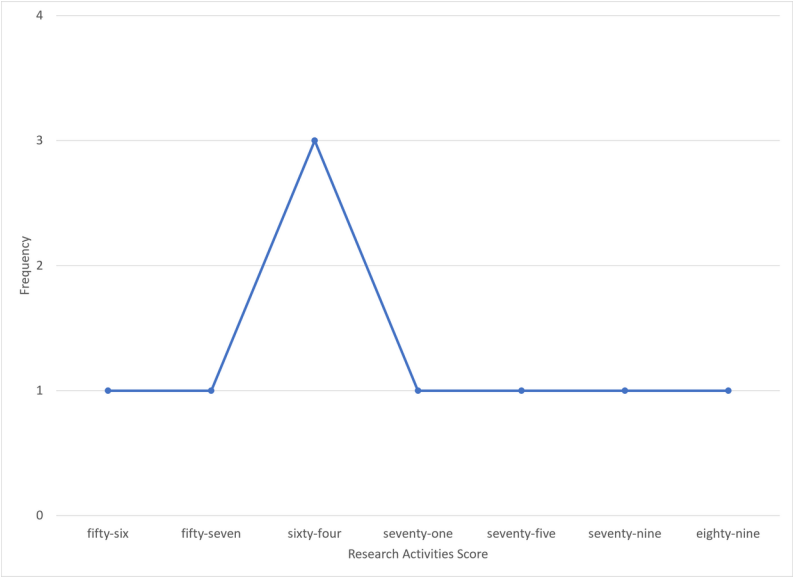


Figure 5.7.7: Frequency Line Graph of Face-to-Face Class's Research Activities Scores. (CC-BY-SA Michelle Oja via personal records)

You might also have notice that you could have pretty easily used Figure 5.7.7) (or the original raw data or frequency table) to just count scores that would be between the range of one standard deviation around the mean. For example, you could mark 58.01 and 79.55 on Figure 5.7.7), and counted the number of scores that fall within your calculations of one standard deviation below and above the mean. This is totally valid! However, most researchers are working with hundreds of scores, sometimes even millions. You wouldn't want to hand-count all of those scores! The ability to remember the proportion of scores that should fall one or two standard deviations around the mean, plus the math literacy to calculate the scores that mark these percentages, can lead to the statistical literacy to make predictions about the actual number of scores in any distribution of quantitative (continuous) data that you're shown.

Note

When might it be useful in your future career to know the expected proportion or even the actual number, of clients you could expect to "score" one standard deviation around the mean?

Let's go through the process one more time. We can use the information from Table 5.7.2, which describes the Research Activities scores for the online section that we analyzed previously.

You might want to make a flow chart or decision tree to help you follow the process before your statistical literacy is stronger.

Table 5.7.2- Descriptive Statistics for Research Activity Points for Online Class

Descriptive Statistic	Research Activities Points
N:	12
Mean:	80.58
Median:	81

Descriptive Statistic	Research Activities Points
Mode:	88
SD	12.43

Again, everything that we've learned so far can be applied to the data set described in Table 5.7.2.

Exercise 5.7.1

1. Who is the sample?
2. Who might be the population?

Answer available at the end of this section.

Exercise 5.7.2

Based on these measures of central tendency and variability, do you think that this online section did well on their research activities, or not so much?

Answer available at the end of this section.

Before we get too far, let's remind ourselves of the proportions in a Standard Normal Curve.

Exercise 5.7.3

1. What percentage of the Research Activities scores *should* be one standard deviation around the mean (one standard deviation below the mean and one standard deviation above the mean)?
2. What percentage of the Research Activities scores *should* be two standard deviations around the mean (two standard deviations below the mean and two standard deviations above the mean)?

Answer available at the end of this section.

Exercise 5.7.4

1. What is the Research Activities score that is one standard deviation *below* the mean?
2. What is the Research Activities score that is one standard deviation *above* the mean?

Answer available at the end of this section.

This means that one standard deviation around the mean for this online section is 68.15 to 93.01 points. Combining that with the proportions in the Standard Normal Curve (see Figure 5.7.6), we can estimate the percentage of students who actually scored in these ranges: we predict (if our distribution of Research Activities scores was normally distributed) that 68% of scores for Research Activities would fall between 68.15 points and 93.01 points. Without looking at an actual frequency line graph, we could get a sense that this section did pretty well since 68% got close to a C up to an A grade.

Adding one more step, let's predict the number of students who we would expect to score in this range (one standard deviation around the mean) in a new sample of online students in a behavioral statistics course.

Exercise 5.7.5

If Dr. MO used this information on the Research Activities scores from the online class, how many students should she expect to earn scores one standard deviation around the mean if she had a new class of 25 students?

Answer available at the end of this section.

Summary

The ideas in this chapter are the foundation for statistical analyses and statistical hypothesis testing. Don't worry too much if this chapter didn't make sense. You can still learn everything in the rest of this textbook even if you haven't put these pieces in the puzzle yet. Just keep working! For some people, you have to see the big picture (which come up in the chapters when we finally test research hypotheses) before these pieces about distributions fits into the picture. As you continue to work with data, your statistical literacy will improve. Eventually, if you keep analyzing data to answer questions, these big ideas will make sense! But in the meantime, you can still analyze and interpret the data that we'll be working with.

Contributors and Attributions

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

Answers

Exercise 5.7.1

1. The sample is 12 students in an online section of behavioral statistics.
2. This sample could represent all online sections of behavioral statistics; this could be the population.

Exercise 5.7.2

The research activities' measures of central tendency are pretty good (the mean, median, and mode are all a B grade), and the standard deviation is medium, so I think that the online class did pretty well on these research activities.

Exercise 5.7.3

1. Looking at Figure 5.7.6) 68% of scores should fall within one standard deviation around the mean.
2. Looking at Figure 5.7.6) 95% of scores should fall within one standard deviation around the mean.

Exercise 5.7.4

1. The Research Activities score that is one standard deviation below the mean is 68.15 points because $\overline{X} - s = 80.58 - 12.43 = 68.15$.
2. The Research Activities score that is one standard deviation above the mean is 79.55 points because $\overline{X} + s = 80.58 + 12.43 = 93.01$.

Exercise 5.7.5

$$\frac{68}{100} \times N = 0.68 \times 25 = 17.00$$

With a new class of 30 students, Dr. MO should expect about 17 students to score between 68.15 points to 93.01 points

Add texts here. Do not delete this text first.

This page titled 5.7: Putting it All Together is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6: USING Z

You've been told that learning about the properties of these distributions will help you answer questions. This chapter does just that!

6.1: Introduction to z-scores

6.2: Calculating z-scores

6.3: Practice z-score Calculation on IQ Scores

6.4: Introduction to the z table

6.5: Practice Using the z Table

6.6: Predicting Amounts

6.7: Summary of z Scores

This page titled 6: Using z is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.1: 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 6.1.1 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.

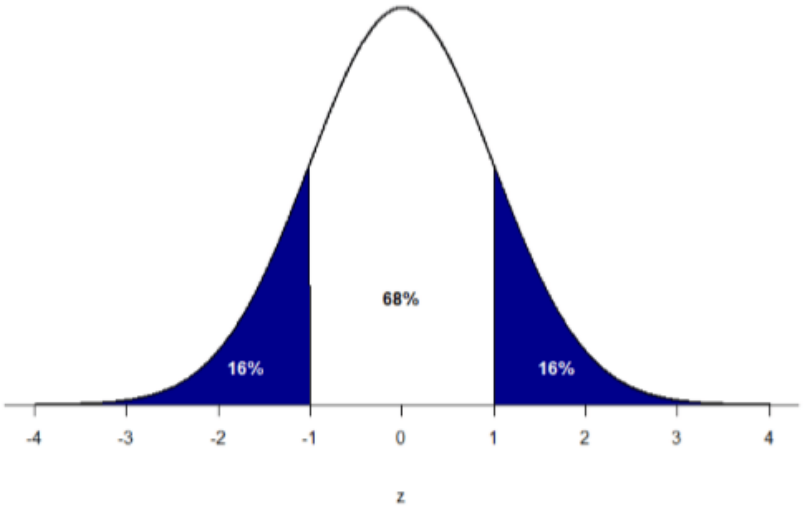


Figure 6.1.1: Shaded areas represent the area under the curve in the tails

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 6.1: 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.

6.2: 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 -1.00 tells us that this z-score is 1 standard deviation (because of the magnitude 1.00) below (because of the negative sign) the mean. Similarly, a z-score value of 1.00 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!

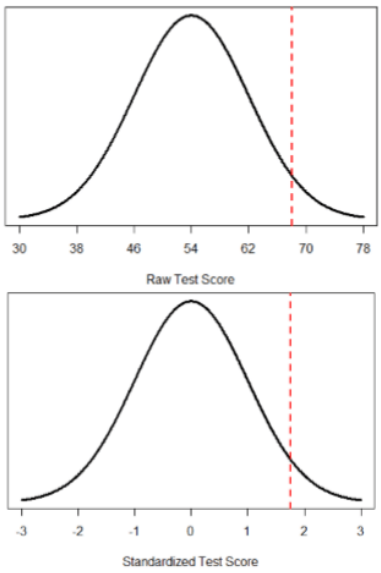


Figure 6.2.1: Raw and standardized versions of a single score (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Figure 6.2.1 shows both the raw score and the z-score on their respective distributions. Notice that the red line indicating where each score lies is in the same relative spot for both. This is because transforming a raw score into a z-score does not change its relative location, it only makes it easier to know precisely where it is.

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

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

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

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

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

Finally, z-scores are incredibly useful if we need to combine information from different measures that are on different scales. Let's say we give a set of employees a series of tests on things like job knowledge, personality, and leadership. We may want to combine these into a single score we can use to rate employees for development or promotion, but look what happens when we take the average of raw scores from different scales, as shown in Table 6.2.1:

Table 6.2.1: Raw test scores on different scales (ranges in parentheses).

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

Because the job knowledge scores were so big and the scores were so similar, they overpowered the other scores and removed almost all variability in the average. However, if we standardize these scores into z-scores, our averages retain more variability and it is easier to assess differences between employees, as shown in Table 6.2.2.

Table 6.2.2: Standardized scores.

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

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

6.3: Practice z-score Calculation on IQ Scores

Example 6.3.1

Assume the following scores represent a sample of statistics classes taken by five psychology professors: 2, 3, 5, 5, 6. If the standard deviation is 1.64, what is the z-score for each of these professor's?

Solution

The mean is 4.2 stats classes taken ($\overline{X} = \frac{\sum X}{N} = \frac{21}{5} = 4.2$).

$$z_2 = \frac{x - \overline{X}}{s} = \frac{2 - 4.2}{1.64} = \frac{-2.2}{1.64} = -1.34$$

$$z_3 = \frac{x - \overline{X}}{s} = \frac{3 - 4.2}{1.64} = \frac{-1.2}{1.64} = -0.73$$

$$z_Both5 = \frac{x - \overline{X}}{s} = \frac{5 - 4.2}{1.64} = \frac{0.80}{1.64} = 0.49$$

$$z_6 = \frac{x - \overline{X}}{s} = \frac{6 - 4.2}{1.64} = \frac{1.8}{1.64} = 1.10$$

PS You might want to practice calculating the standard deviation yourself to make sure that you haven't forgotten how!

Your turn!

Exercise 6.3.1

Calculate z-scores for the three IQ scores provided, which were taken from a population with a mean of 100 and standard deviation of 16: 112, 109, 88.

Answer available at the end of this section.

This time, you'll get the z-score and will need to find the IQ scores. Remember, you can do this with the z-score formula that you used above and to algebra to find \(x|), or you an use the other z-score formula.

Exercise 6.3.2

Use the z-scores provided to find two IQ scores taken from a population with a mean of 100 and standard deviation of 16:

z = 2.19

z = -0.06

Answer available at the end of this section.

Contributors and Attributions

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

- Dr. MO (Taft College)

Answers

Exercise 6.3.1

z_112 = 0.75

z_109 = 0.56

z_88 = -0.75

Exercise 6.3.2

The IQ scores are 135 (for z = 2.19) and 99 (z = -0.06) .

This page titled 6.3: Practice z-score Calculation on IQ Scores is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.4: 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 $6.4.1$. 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.

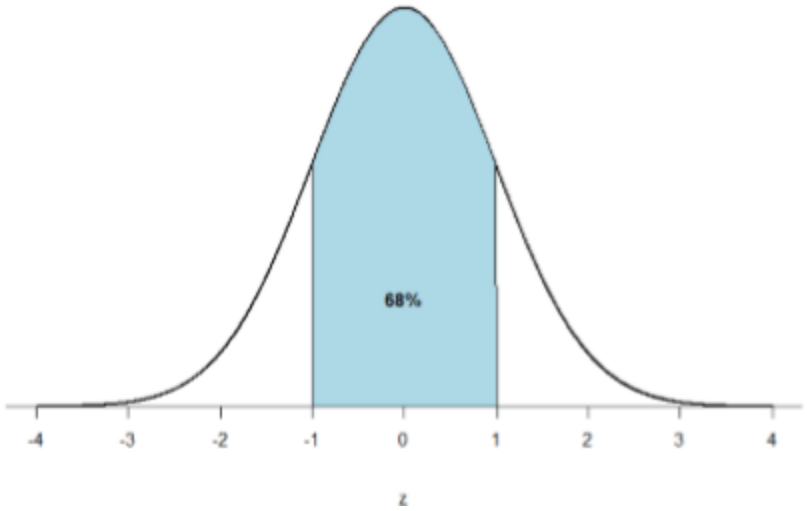


Figure $6.4.1$: There is a 68% chance of selection a z-score from the blue-shaded region (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Take a look at the normal distribution in Figure $6.4.2$ 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.

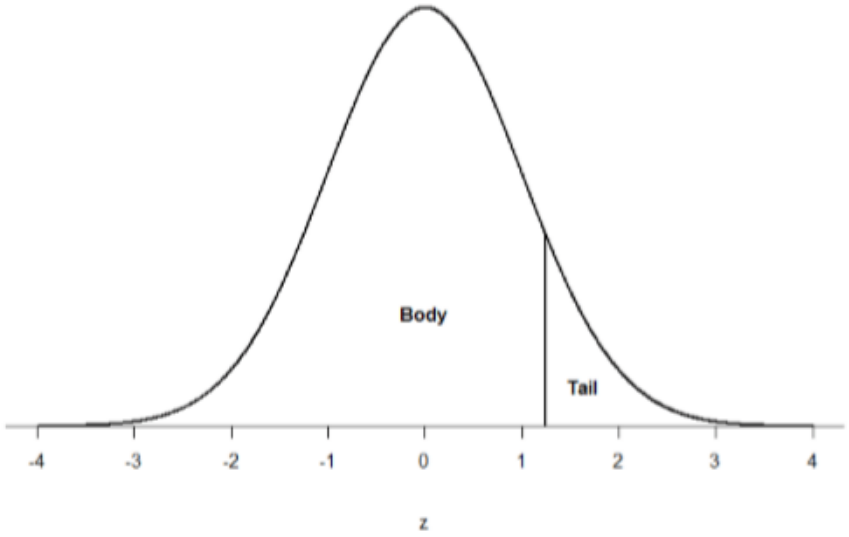


Figure 6.4.2: Body and tail of the normal distribution (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 6.4.3, which corresponds to the area more extreme than z = -1.96 and z = 1.96.

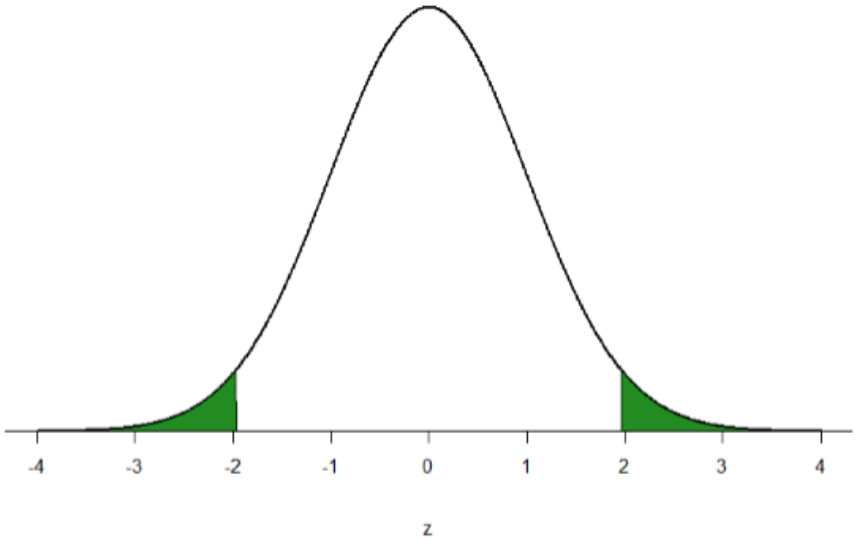


Figure 6.4.3: Area in the tails beyond z = -1.96 and z = 1.96 (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 6.4.1?

Example 6.4.1

What p-value corresponds to 1.96 from the z-table in Table 6.4.1?

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 6.4.3 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 6.4.3, 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 6.4: 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.

6.5: Practice Using the z Table

It's time to practice with the z-table!

Example 6.5.1

Find the z-score that bounds the *top* 9% of the distribution.

Solution

Because we are looking for *top* 9%, we need to look for the p-value closest to p = .91000 ($100\% - 9\% = 91\%$) because the p-values (probabilities) in the z Table show the probability of score being lower, but this question is asking for top 9%, not the portion lower than 9%. There should be 91% of scores lower than the top 9%.

The closest p-value to p = .91000 (91%) is 0.90988. The z-score for p = 0.90988 iz z=1.34.

The z-score for the *top* \(9\%\) of the distribution is z=1.34 (for p=0.90988, the closest probability to 91%, which marks everyone lower than the top 9%).

Your turn!

Exercise 6.5.1

Find the z-score that bounds 25% (p=0.25000) of the *lower* tail of the distribution.

Hint: You don't have to subtract anything for this one because the question is asking about the scores that are *lower*.

Answer available at the end of this section.

Now, let's try some scenarios...

Example 6.5.2

The heights of women in the United States are normally distributed with a mean of 63.7 inches and a standard deviation of 2.7 inches. If you randomly select a woman in the United States, what is the probability that she will be between taller than 64 inches?

Solution

X (raw score) = 64 inches

$\overline{X}$ = 63.7 inches

s=2.7 inches

$$z = \frac{x - \overline{X}}{s} = \frac{64 - 63.7}{2.7} = \frac{0.30}{2.7} = 0.11$$

Finding z=0.11 on the z Table, we see that p = 0.54380. This is the probability that a score will be lower than our raw score, but the question asked the proportion who would be taller.

$1 - 0.54380 = 0.4562$

$p \times 100 = 0.4562 \times 100 = 45.62\%$

Final Answer (in words): The probability that a woman in the U.S. would be 64 inches or taller is 0.4562, or 45.62%.

Your turn!

Exercise 6.5.2

The heights of men in the United States are normally distributed with a mean of 69.1 inches and a standard deviation of 2.9 inches. What proportion of men are taller than 6 feet (72 inches)?

Answer available at the end of this section.

Last one, on something that you might find relevant!

Exercise 6.5.3

Imagine that you scored 82 points on a final exam. After the final, you find out that the average score on the exam was 78 with a standard deviation of 7. What proportion (in a percentage) did worse than you (earned a *lower* score)?

Answer available at the end of this section.

Contributors and Attributions

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

Answers

Exercise 6.5.1

The z-score for 25% of the lower tail of the distribution is$z = -0.67$ (for p=0.25143, the closes probability to 0.25000 (25%)).

Exercise 6.5.2

Final Answer: The probability that a man in the U.S. would be 72 inches or taller is 0.15866, or 15.87%.

Exercise 6.5.3

The proportion of students in the class who did worse than you (earned a *lower* score) should be 71.57%.

This page titled 6.5: Practice Using the z Table is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.6: 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 6.6.1

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 6.6.2

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 6.6.1

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 available at the end of this section.

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.

Answers

Exercise 6.6.1

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.

This page titled 6.6: Predicting Amounts is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

6.7: 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:

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

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

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

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.

Write-Ups

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.

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! For now, you know how to describe a distribution of quantitative numbers with measures of central tendency and standard deviations (a measure of variability), and make predictions about the shape of the distribution based on these measures. You also know how to create a

frequency line graph to compare to see if your predictions were correct. You also know how to answer some research questions related to predicting the percentage and amount of scores that should fall standard deviation units around the mean using the standard normal curve, and the percentage and amount of scores that are above or below a target score using the z-table. This means that you can now start writing your conclusions up in a paper! In the social sciences, the predominant writing format is APA Style. The next chapter will help you learn what you need to know to write about statistical analyses.

Contributors and Attributions

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

This page titled 6.7: Summary of z Scores is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Danielle Navarro.

7: APA STYLE

We have already discussed how to format the title of a table and the title of a graph in APA Style, and now it is time to talk about formatting a whole paper in APA Style.

7.1: APA and APA Style

7.2: APA Style Resources

7.3: General Paper Format

7.4: Formatting by Section

7.5: Tables and Figures

This page titled 7: APA Style is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

7.1: APA and APA Style

The American Psychological Association (APA) citation style is the commonly used format for manuscripts in the social and health sciences.

APA style regulates style and formatting. These regulations try to make it so that all papers that are submitted for publication are evaluated based on the research involved, not on fancy graphics or writing ability. There are writing and grammar suggestions, as well, so that all manuscripts use similar writing styles and are free from bias.

Edition?

The APA manual was recently updated to the 7th edition. The 7th edition has a Professional style and a Student style. **We will be discussing the Professional style**. Many resources on the internet are still using the 6th edition, or even earlier editions, so don't trust them!

This page titled 7.1: APA and APA Style is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

7.2: APA Style Resources

Speaking of resources, here are some of my favorites!

This chapter

This chapter will be a great resource to explain APA Style formatting, how to do it, and sometimes why we do things the way that we do.

OWL Purdue

The Purdue University's Online Writing Lab (website address: http://owl.english.purdue.edu) has an immense collection of webpages dedicated to APA Style, including specific pages that you might need like a whole section on Tables and Figures. There's a PDF of a sample Professional Style paper with explanations of formatting. There's also a sample paper for students, but we are using the professional version.

APA Style Checklist

This PDF is a checklist developed by Dr. MO to go through before you turn in any statistics paper that is supposed to be formatted in APA Style. You can use this for non-statistics APA Style papers, too, but there are other important features in APA Style that aren't included because they are less important when describing a distribution compared to a full manuscript on an experiment that you conducted.

Video

Many people learn best by watching (and re-watching) videos, so here's a good video in Figure 7.2.1 showing how to format a Word document into an APA Style paper. Although sometimes the features are in different places in a Google doc or on a Mac, all word processing software will allow you to make these modifications.

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Figure 7.2.1: APA Essay Format (7th Edition). (The Nature of Writing via YouTube)

DO NOT USE THE APA STYLE TEMPLATE PROVIDED BY WORD. I know that this seems like an easy way to format you paper, but the template is really difficult to modify. For example, if you don't need, say, an abstract or a reference page for your paper, it is nearly impossible to delete those from your paper when created from the template without messing up the formatting for the rest of your paper.

Sample Paper

Here is a sample paper as a Word document (.docx). Save this to your own drive or computer so that you can save and modify it. This is an example paper to show APA Style formatting, and the basic information in a descriptive statistics paper. Careful reading shows that the topic and numbers change throughout the paper because it is a combination of several different students'

papers. This is just a guide; do NOT copy the words, just the format.

American Psychological Association

Publication Manual

Dr. MO encourages you to purchase the APA Style manual if you are a psychology or sociology major. You can find resources at https://apastyle.apa.org/ . Additionally, your school's library has the *Publication Manual of the American Psychological Association, 7th ed.* Note that it is typically considered a reference (like a dictionary or encyclopedia), so you probably can't take it out of the library. Your library probably has a webpage devoted to APA Style, as well.

The APA also provides instructional aids (website address: https://apastyle.apa.org/instructional-aids), some for free and some are not. When using these, make sure to use the resources for the Professional type of paper, not the Student version (unless your instructor prefers the Student version).

Writing Guides

While your primary writing guide should be the APA Style publication manual. The following writing guides developed by the APA provide guidelines for specific types of research and topics:

- JARS-Quant (2018) standards (includes additional guides for many different types of designs and methodologies): https://apastyle.apa.org/jars/quantitative
- JARS-Qual (2018) standards: https://apastyle.apa.org/jars/qualitative
- JARS-Mixed (2018) standards: https://apastyle.apa.org/jars/mixed-methods
- Three key things you should know about APA's new inclusive language guidelines (Akbar, 2022) *APA Style*. https://apastyle.apa.org/blog/inclus...age-guidelines (which is supposed to link to this document: Equity, Diversity, and Inclusion: Inclusive Language Guidelines (APA, 2023): https://pages.apa.org/edi-guidelines/)
- JARS- Race, Ethnicity, and Culture (2023) (includes video towards the bottom of the page): https://apastyle.apa.org/jars/race-ethnicity-culture
- Sexual Orientation and Gender Diversity Research Manuscript Writing Guide (Veldhuis et al., 2024): https://psycnet.apa.org/fulltext/2024-79040-001.pd

Contributors and Attributions

- Dr. MO (Taft College)

- The Nature of Writing's APA Essay Format: https://www.youtube.com/watch?v=Mrh5OC3T6dc

This page titled 7.2: APA Style Resources is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

7.3: General Paper Format

APA papers have different sections that have different slightly different formatting styles, however, the **entire** paper should:

The ENTIRE paper should:

- Be double-spaced (except Tables and Figures)
- Be 12 pt. Times New Roman font (or similar)
- Have 1" margins on all sides
- Not have any extra spacing before or after each paragraph.
  - Check **Paragraph** settings for this; the default in Word is to add 10 points of space after each paragraph (regardless of line spacing).
- Have a page number in the upper right-hand side of every page that is the same font as the rest of the paper.
- Have the paper's title, all capitalize, as a page header in the upper left-hand of every page. When all papers were submitted by mailing hard-copies to journal publishers, the running head was used to keep the papers together and organized, but anonymous. Dr. MO has also used this when she has literally dropped a stack of printed-but-not-yet-stapled student papers.

There should be a page break (starting on a new page) *between* each of these sections. There should not be a page break *within* any of these sections. Even if you have part of a sentence dangling on a new page. *Insert a Page Break* to start a new page, don't just hit Enter a bunch of times. Doing that messes up other formatting.

In-Text Citations

If you cite any sources, there is a particular way to format them in your sentences. Surprised?

In-text citations help readers locate the source of your information in the References section of the paper. Whenever you use a source, provide the author's last name(s), *in the order that they appear on the article*, and the year of publication. You can use the authors' names in the sentence, or only include their names in the reference at the end of the sentence.

Example with One Author

- Using their name in the sentence: Wilson (2014) developed mechanical wings.
- Not using their name in the sentence: Mechanical wings can be used as shielding or for flight (Wilson, 2014).

Example with Two Authors

- Using their names in the sentence: According to Prince and Trevor (2021), you must sometimes sacrifice to live your values.
- Not using their name in the sentence: There is still controversy over the cause of World War I, also known as the Great War (Prince & Trevor, 2017).

Notice that everything cited is in the *past tense*. It was written before today, so it is in the past.

> **Note**
>
> What is different with two authors between using the authors' names in the sentence, compared to at the end of the sentence?

Example with Three or More Authors

When citing a work with three or more authors, you can use the first author's last name followed by "et al." in the sentence or in parentheses:

- Using their names in the sentence: Kirk et al. (1968) found many similarities between their different cultures.
- Not using their name in the sentence: Exploration shows that there is infinite diversity in infinite combinations (Kirk et al., 1968).

Students often want to cite once at the end of a whole paragraph. Resist this urge. In APA Style, each sentence with information from an outside source must be cited. Citing at the end of the sentence makes it difficult to recognize which parts of the paragraph are the author's thoughts, and which portions are the ideas or findings for another author.

This page titled 7.3: General Paper Format is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

7.4: Formatting by Section

APA papers generally include the following sections:

1. Cover page (sometimes called a title page)
2. Abstract
3. Main body
4. References
5. Appendices (including Tables and Figures)

Each of these sections have different slightly different formatting styles. The formatting discussed in the last section that should be throughout the paper will *not* be repeated here, so make sure to format your whole paper that way first, then move on to making sure each section is correctly formatted.

First Section: Title Page

In the upper half of the page, centered:

- The title of your paper, in **bold**.
- Your name
- Affiliation (your college/university)

Insert a Page Break after the affiliation so that the first page of the main body starts on the next page. [Don't just hit Enter a bunch of times. Doing that messes up other formatting.]

Second Section: Abstract

An abstract is a 150- to 250- word summary of your paper. The abstract page should include:

- Page header and page number (2, because this will always be on the second page)
- The abstract title (which should be the word Abstract), centered and **bold**, at the top of the page.
- The abstract paragraph is **not indented**.

Third Section: Main Body

This is the section of the paper where most of your writing will be; this is where you describe what was measured, on whom, and what was found.

The first page of the main body includes the **title in the header *and* at the top of the page (as a heading).** This title is in the style of a Level 1 heading (see below). Every page of your main body should include:

- Page header and Page number
- Indented paragraphs (**with no extra space between the paragraphs**).

There should *not* be a page break *within* the main body; Just let the paragraphs break where they naturally do.

Headings

A heading is a title for different sections of your paper. It is not a header, which is the title at the top of each page.

If you choose to use headings, they should be formatted in correct APA Style. APA uses a system of five heading levels, as shown in the Note below. Start with the Level 1 style for your titles and headings, then move down for lower level sections.

Note

Level 1: Centered, Bold, Title Case (first letter of important words is capitalized)

Level 2: Flush Left, Bold, Title Case

Level 3: Flush Left, Bold & Italicized, Title Case

It is exceedingly unlikely that you will need a Level 4 or Level 5 heading, but here they are:

Level 4: Indented, Bold, Title Case, Ending with a Period.

Level 5: Indented, Bold and Italicized, Title Case, Ending with a Period.

Fourth Section: Reference Page

You need a reference page when you cite someone else's work. You might think that you won't be citing anything in a statistics paper, but you actually might! For example, if you were comparing the number of siblings of your classmates, you might cite a website showing the average number of siblings of college students in the U.S. for your population mean.

Like every other page, the reference page should include:

- Running head (only the paper's title, all capitalized)
- The page number
- Double-spaced, with no extra space between paragraphs

It should also include a title (**References**) at the top of the page in bold (Level 1 heading). APA uses References**, *not* works cited *or* bibliography**.

The indention of references are opposite of paragraphs in the main body. Each reference should start all of the way to the left, and each subsequent line should be indented. This is called a *hanging indent*, and Word will do this automatically for you through the paragraph settings.

Fifth Section: Appendix (Tables and Figures)

For our purposes, an appendix will be used for tables and figures.

Note

What other types of information might be put in an appendix?

The next section will discuss tables and figures in more detail.

Summary

That's the basics of formatting in APA Style!

Exercise 7.4.1

1. Should the words "Running head" be in your paper?
2. Should your paper include a cover page?
3. Should your paper have your name on it?

4. Should your paper have your professor's name on it?
5. Do you have to cite this textbook?

Answer available at the end of this section.

The last section of this chapter will review formatting tables and figures in APA Style, but you learned that in the chapter on graphs already.

Answers

Exercise 7.4.1

1. No
2. Yes
3. Yes
4. No
5. No, unless you are using it as a direct source of information.

This page titled 7.4: Formatting by Section is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

7.5: Tables and Figures

Tables are where you put rows or columns of numbers, like a frequency table or your calculation results. A figure is a chart/graph or other picture. The 7th Edition allows you to put the tables and figures in the main body of the paper, but you can also put them on their own pages at the end of your paper.

- Tables and Figures can be single-spaced.
- Tables and Figures must be numbered, in **bold** (Example: **Figure 1** or **Table 1**). The table or figure number should be on a line above the title.
- Tables and Figures must have a title (*in italics*) on the line below the number.
- In a Table, the column titles should be underlined.
- There should be a line across the whole bottom edge of each table.

If you are putting a table or figure in an appendix, then each set of tables or figures should be started on a new page [Insert a Page Break, don't hit enter a bunch of times].

Example Table Formatting

Table 1

Measures of Central Tendency and Standard Deviation

Table 7.5.1- Example of APA Style Table Formatting

Groups	N	Mean	Median	Mode	SD
Group 1	34	9	2	6.50	2.51
Group 2	550	246.99	45.5	22.22	5.54
Note: "SD" is "standard deviation.					

Table 7.5.1: Example of number, title, and table formatted in APA Style.. (CC-BY-NC-SA; Michelle Oja)

Example Chart Formatting

Figure 1

Frequency of Final Exam Scores

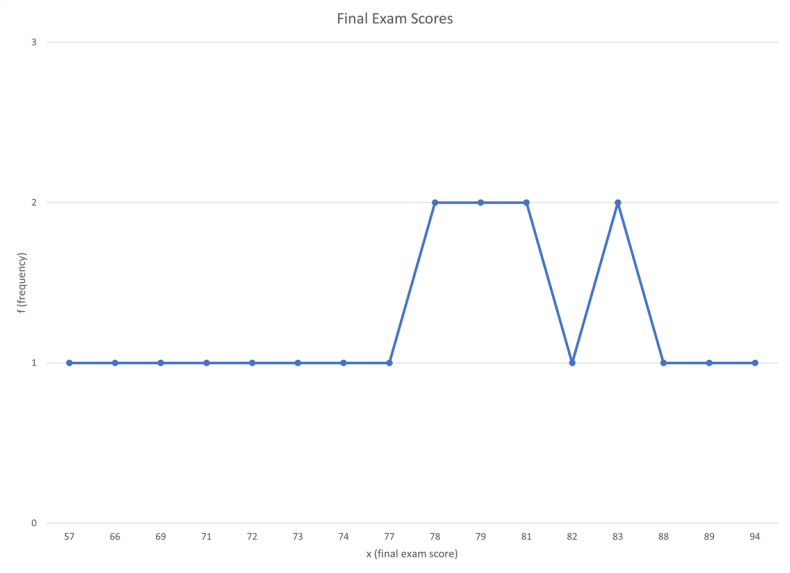


Figure 7.5.1: Example of number, title, and chart formatted in APA Style. (CC-BY-NC-SA; created by Michelle Oja with data from OpenIntro.org)

That's it! You have all you need to identify a distribution's variables and describe what that distribution's measures of central tendency, measure of variability and appropriate graphs tell us, plus report all of that information in APA Style. Wow, you've learned a lot!

This page titled 7.5: Tables and Figures is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

8: INFERENTIAL STATISTICS AND HYPOTHESIS TESTING

8.1: Growth Mindset

8.2: Samples and Populations Refresher

8.3: Can Samples Predict Populations?

8.4: Descriptive versus Inferential Statistics

8.5: The Research Hypothesis and the Null Hypothesis

8.6: Null Hypothesis Significance Testing

8.7: Critical Values

8.7.1: Rejection Region Visualization

8.8: Practice with p-values and NHST

8.9: Steps of the Hypothesis Testing Process

8.10: The Two Errors in Null Hypothesis Significance Testing

8.11: Power and Sample Size

8.12: Views of NHST

This page titled 8: Inferential Statistics and Hypothesis Testing is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

8.1: 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 8.1: Growth Mindset is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

8.2: Samples and Populations Refresher

Population Parameters and Sample Statistics

Up to this point we have been talking about populations the way a scientist might. To a psychologist, a population might be a group of people. To an ecologist, a population might be a group of bears. In most cases the populations that scientists care about are concrete things that actually exist in the real world. Statisticians, however, are a funny lot. On the one hand, they *are* interested in real world data and real science in the same way that scientists are. On the other hand, they also operate in the realm of pure abstraction in the way that mathematicians do. As a consequence, statistical theory tends to be a bit abstract in how a population is defined. Statisticians operationalize the concept of a "population" in terms of mathematical objects that they know how to work with, namely: probability distributions.

The idea is quite simple. Let's say we're talking about IQ scores. To a psychologist, the population of interest is a group of actual humans who have IQ scores. A statistician "simplifies" this by operationally defining the population as the probability distribution depicted in Figure 8.2.1. IQ tests are designed so that the average IQ is 100, the standard deviation of IQ scores is 15, and the distribution of IQ scores is normal. These values are referred to as the ***population parameters*** because they are characteristics of the entire population. That is, we say that the population mean μ (mu) is 100, and the population standard deviation is 15. [Although not directly related to this conversation, we also covered non-parametric analyses in Ch. 4.5, which were analyses that we could do when we don't think that the population is normally distributed.]

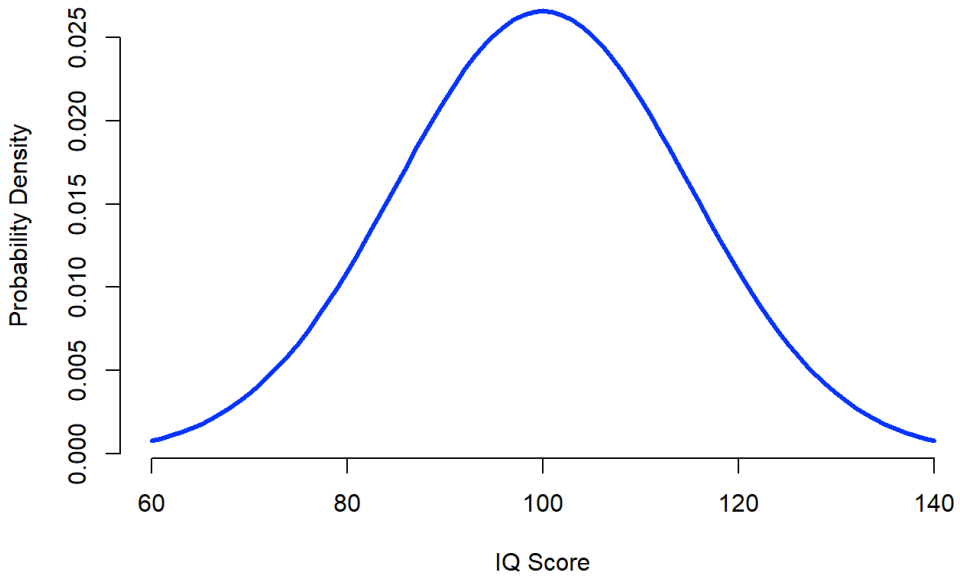


Figure 8.2.1- Probability Distribution of IQ (CC-BY-SA- Danielle Navarro from Learning Statistics with R)

Now suppose I run an experiment. I select 100 people at random and administer an IQ test, giving me a random sample from the population. Each of these IQ scores is sampled from a normal distribution with mean 100 and standard deviation 15. So if I plot a histogram of the sample, I get something like the one shown in Figure 8.2.2.

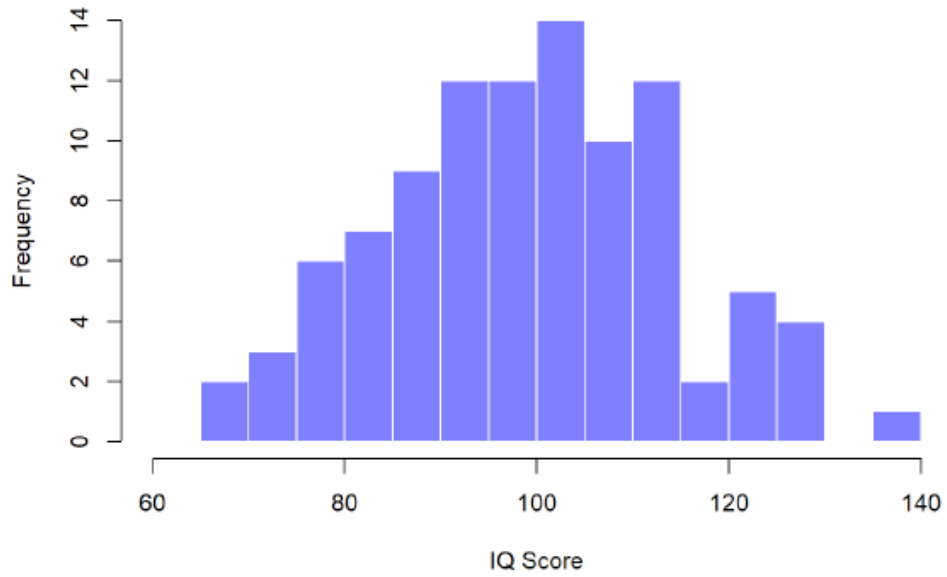


Figure 8.2.2- Probability Distribution of 100 IQ Scores (CC-BY-SA- Danielle Navarro from Learning Statistics with R)

As you can see, the histogram is *roughly* the right shape, but it's a very crude approximation to the true population distribution shown in Figure 8.2.1. Even though there seems to be an outlier who scored near 140 points on the IQ test, the mean of my sample is fairly close to the population mean 100, but not identical. In this case, it turns out that the people in my sample have a mean IQ of 98.5 ($\overline{X} = 98.5$), and the standard deviation of their IQ scores is 15.9. These ***sample statistics*** are *descriptive* of the data set, and although they are fairly similar to the true population values, they are not the same the population parameters. In general, sample statistics are the things you can calculate from your data set, and the population parameters are the things you want to learn about.

As we learned in the chapter on distributions, the bigger the sample, the more like a normal curve is is. You can see this in Figure 8.2.3, a random selection of 10,000 IQ scores from a regular population looks very much like a normal distribution.

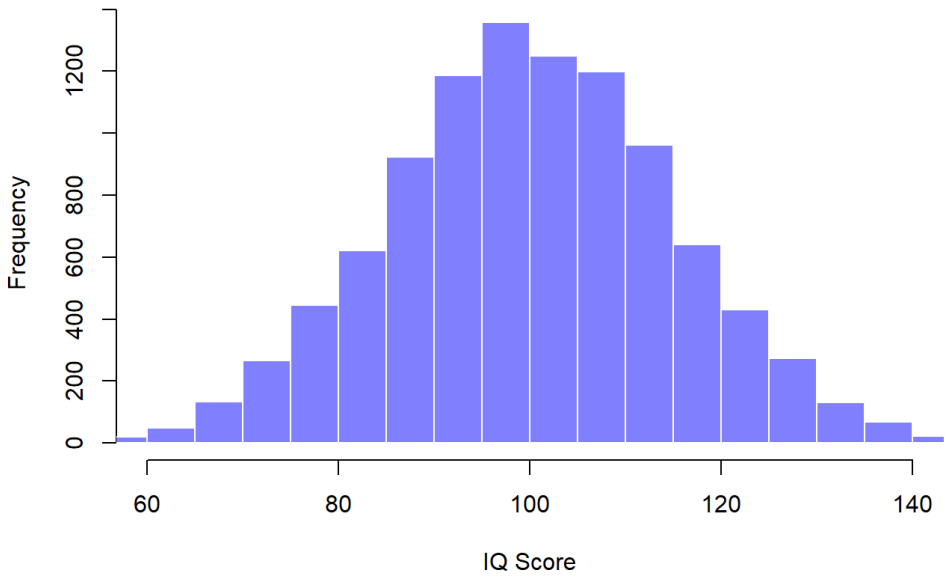


Figure 8.2.3- Probability Distribution of 10,000 IQ Scores (CC-BY-SA- Danielle Navarro from Learning Statistics with R)

Onward and upward!

The next section talks a little more about how samples are sometimes similar to their population, and sometimes not...

This page titled 8.2: Samples and Populations Refresher is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 10.1: Samples, Populations and Sampling by Danielle Navarro is licensed CC BY-SA 4.0. Original source: https://bookdown.org/ekothe/navarro26/.

8.3: Can Samples Predict Populations?

To understand whether samples can be used to describe population's, let start with a silly experiment Imagine that you had a gumball machine. We twist the dial with our right hand, and see how many green gumballs come out. We put the gumballs back in, then twist the dial with our left hand and count how many green gumballs come out. In sum, in each experiment we counted the green balls for the left and right hand. What we really want to know is if there is a *difference* between them. So, we can calculate the **difference score** (which is just subtracting one score from the other).

$$\text{Difference} = X_L - X_R$$

To show that the difference score is the of green gumballs from the left hand minus the number of green gumballs from the right hand. The difference should be zero, but sampling error produces some differences. The bars in Figure $8.3.1$ shows the differences (note that it is not frequency chart) in the number of green gumballs ($Difference = X_L - X_R$). Missing bars mean that there were an equal number of green gumballs chosen by the left and right hands (difference score is 0). A positive value means that more green gumballs were chosen by the left than right hand. A negative value means that more green gumballs were chosen by the right than left hand. Note that if we decided (and we get to decide) to calculate the difference in reverse (right hand - left hand), the signs of the differences scores would flip around.

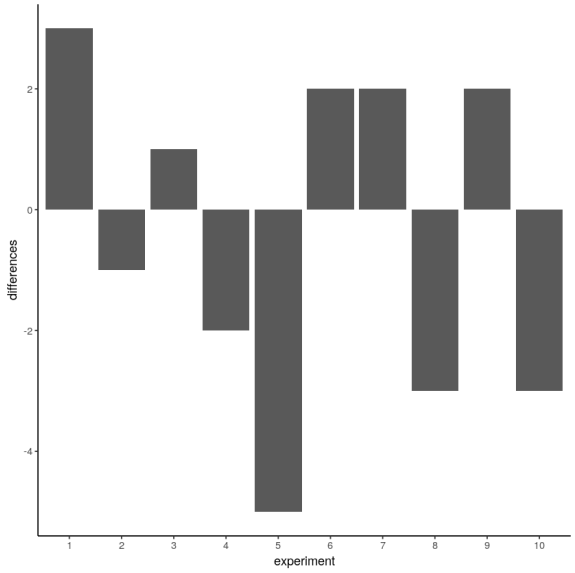


Figure $8.3.1$- Difference in Number of Green Gumballs from 10 Draws from Left and 10 Draws from Right Hand (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

We are starting to see the differences that chance can produce. The difference scores are mostly between -2 to +2, meaning that there are usually not more than 2 more green gumballs when the left hand was chosen, and usually not few than 2 green gumballs, either. We could get an even better impression by running this pretend experiment 100 times instead of only 10 times. How about we do that (Figure $8.3.2$).

Ooph, we just ran so many simulated experiments that the x-axis is unreadable, but it goes from 1 to 100. Each bar represents the *difference* of number of green balls chosen randomly by the left or right hand. There are man kinds of differences that chance alone can produce!

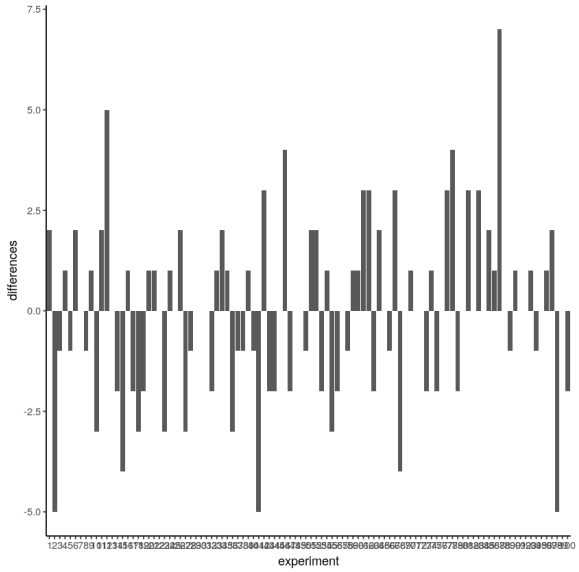


Figure 8.3.2- Difference in Number of Green Gumballs from 100 Draws from Left and 100 Draws from Right Hand (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

Beginning to notice anything? Look at the y-axis, this shows the size of the difference. Yes, there are lots of bars of different sizes, this shows us that many kinds of differences do occur by chance. However, the y-axis is also restricted. It does not go from -10 to +10. Big differences (greater than 5 or -5) don't happen very often.

Now that we have a method for simulating differences due to chance, let's run 10,000 simulated experiments. But, instead of plotting the differences in a bar graph for each experiment, how about we look at the histogram of frequency of each *difference* score. This will give us a clearer picture about which differences happen most often, and which ones do not. This will be another window into chance. The chance window of differences.

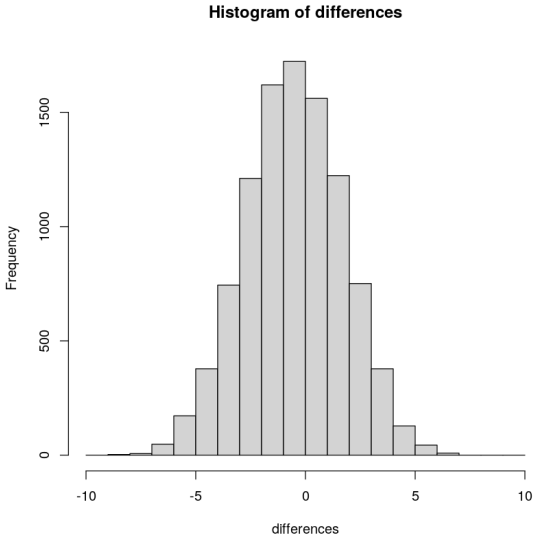


Figure 8.3.3- Difference in Number of Green Gumballs from 10,000 Draws from Left and 10,000 Draws from Right Hand (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

The most frequency difference in Figure 8.3.3 is 0, which is what we expect by chance, that neither hand has a special power to attract green gumballs. Although there can be differences as large as 10 green gumballs between each hand, larger differences occur less often by chance.

Exercise 8.3.1

What does this histogram remind you of?

Answer available at the end of this section.

What else can this histogram tell us about chance? It can show that chance produces some differences more often than others. First, chance usually produces 0 differences, that's the biggest bar in the middle. Second, chance can produce larger differences, but as the differences get larger (positive or negative), they occur less frequently. The shape of this histogram is your chance window, it tells you what chance can do, it tells you what chance usually does, and what it usually does not do.

OK, we have seen that chance can produce differences here. But, we still don't have a good idea about what chance usually does and doesn't do. For example, if we could find the window of opportunity here, we would be able find out that chance usually does not produce differences of a certain large size. If we knew what the size was, then if we ran experiment and our difference was bigger than what chance can do, we could be confident that chance did not produce our difference.

You can use this chance window to help you make inferences. If you ran yourself in the gumball experiment and found that your left hand chose 2 more green gumballs than red gumballs, would you conclude that you left hand was special, and caused you to choose more green gumballs? Hopefully not. You could look at the chance window and see that differences of size +2 do happen fairly often by chance alone. You should not be surprised if you got a +2 difference. However, what if your left hand chose 5 more green gumballs than red gumballs. Well, chance doesn't do this very often, you might think something is up with your left hand. If you got a whopping 9 more green gumballs than red gumballs, you might really start to wonder. This is the kind of thing that could happen (it's possible), but is very rare if the sample is randomly chosen. When you get things that almost never happen by chance, you can be more confident that the difference reflects a causal force that is not chance.

Bringing it Together

Okay, so what have we covered so far in this chapter.

First, populations are different from samples.

Second, that samples might be able to be used to predict population characteristics (parameters).

Finally, that even when the sample is from the population, it is not exactly like the population; the sample's mean and standard deviation are not going to be exactly the same at the population mean and standard deviation.

So, how can we tell when your sample is similar enough to the population that you can use the sample to predict the population? That's inferential statistics!

Answers

Exercise 8.3.1

Yep, a normal curve!

This page titled 8.3: Can Samples Predict Populations? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 5.4: Chance makes some differences more likely than others by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.
- 7.2.1: Can Samples Predict Populations? by Michelle Oja is licensed CC BY-SA 4.0.

8.4: 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 8.4.1. The table shows the average salaries for various occupations in the United States in 1999.

Table 8.4.1: 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 8.4.2. 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 8.4.2: 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

Cities with Mostly Men	Men per 100 Women	Cities with Mostly Women	Men per 100 Women
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 8.4.3: 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 8.4.4 shows the same statistics, but for men.

Table 8.4.4: 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

Year	Winner	Country	Time
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 8.4.3 and Table 8.4.4 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 8.4.1

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 8.4.1

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 available at the end of this section.

Answers

Exercise 8.4.1

#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 8.4: Descriptive versus Inferential Statistics is shared under a CC BY-NC-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.

8.5: The Research Hypothesis and the Null Hypothesis

Hypotheses

Hypotheses are predictions of expected findings.

The Research Hypothesis

A research hypothesis is a mathematical way of stating a research question. A research hypothesis names the groups (we'll start with a sample and a population), what was measured, and which we think will have a higher mean. The last one gives the research hypothesis a direction. In other words, a research hypothesis should include:

1. The name of the groups being compared. This is sometimes considered the IV.
2. What was measured. This is the DV.
3. Which group are we predicting will have the higher mean.

There are two types of research hypotheses related to sample means and population means: Directional Research Hypotheses and Non-Directional Research Hypotheses

Directional Research Hypothesis

If we expect our obtained sample mean to be above or below the other group's mean (the population mean, for example), we have a directional hypothesis. There are two options:

- "The sample mean is expected to be *bigger* than the population mean."
  - Symbol: $\overline{X} > \mu$
  - (The mean of the sample is greater than than the mean of the population.)
- "The sample mean is expected to be *smaller* than the population mean."
  - Symbol: $\overline{X} < \mu$
  - (The mean of the sample is less than than mean of the population.)

Example 8.5.1

A study by Blackwell, Trzesniewski, and Dweck (2007) measured growth mindset and how long the junior high student participants spent on their math homework. What's a *directional* hypothesis for how scoring higher on growth mindset (compared to the population of junior high students) would be related to how long students spent on their homework? Write this out in words and symbols.

Solution

Answer in Words: Students who scored high on growth mindset would spend more time on their homework than the population of junior high students.

Answer in Symbols: $\overline{X} > \mu$

Non-Directional Research Hypothesis

A non-directional hypothesis states that the means will be different, but does not specify which will be higher. In reality, there is rarely a situation in which we actually don't want one group to be higher than the other, so we will focus on directional research hypotheses. There is only one option for a non-directional research hypothesis: "The sample mean *differs* from the population mean." These types of research hypotheses don't give a direction, the hypothesis doesn't say which will be higher or lower.

A non-directional research hypothesis in symbols should look like this: $\overline{X} \neq \mu$ (The mean of the sample is not equal to the mean of the population).

Exercise 8.5.1

What's a *non-directional* hypothesis for how scoring higher on growth mindset higher on growth mindset (compared to the population of junior high students) would be related to how long students spent on their homework (Blackwell, Trzesniewski, & Dweck, 2007)? Write this out in words and symbols.

Answer available at the end of this section.

See how a non-directional research hypothesis doesn't really make sense? The big issue is not if the two groups differ, but if one group seems to improve what was measured (if having a growth mindset leads to more time spent on math homework). This textbook will only use *directional* research hypotheses because researchers almost always have a predicted direction (meaning that we almost always know which group we think will score higher).

The Null Hypothesis

The hypothesis that an apparent effect is due to chance is called the null hypothesis, written H_0 ("H-naught"). We usually test this through comparing an experimental group to a comparison (control) group. This null hypothesis can be written as:

$$\mathrm{H}_0 : \overline{X} = \mu$$

For most of this textbook, the null hypothesis is that the means of the two groups are similar. Much later, the null hypothesis will be that there is no relationship between the two groups. Either way, remember that a null hypothesis is always saying that nothing is different.

This is where descriptive statistics diverge from inferential statistics. We know what the value of $\overline{\mathrm{X}}$ is – it's not a mystery or a question, it is what we observed from the sample. What we are using inferential statistics to do is infer whether this sample's descriptive statistics probably represents the population's descriptive statistics. This is the null hypothesis, that the two groups are similar.

Keep in mind that the null hypothesis is typically the opposite of the research hypothesis. A research hypothesis for the ESP example is that those in my sample who say that they have ESP would get more correct answers than the population would get correct, while the null hypothesis is that the average number correct for the two groups will be similar.

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.

In sum, the null hypothesis is *always*:

There is **no difference between the groups' means**

OR

There is **no relationship between the variables**.

In the next chapter, the null hypothesis is that there's no difference between the *sample mean*_and **population mean.** In other words:

- There is no mean difference between the sample and population.
- The mean of the sample is the same as the mean of a specific population.
- $\mathrm{H}_0 : \overline{X} = \mu$

- We expect our sample's mean to be same as the population mean.

Exercise $8.5.2$

A study by Blackwell, Trzesniewski, and Dweck (2007) measured growth mindset and how long the junior high student participants spent on their math homework. What's the *null hypothesis* for scoring higher on growth mindset (compared to the population of junior high students) and how long students spent on their homework? Write this out in words and symbols.

Answer available at the end of this section.

Answers

Exercise $8.5.1$

Answer in Words: Students who scored high on growth mindset would spend a different amount of time on their homework than the population of junior high students.

Answer in Symbols: $\overline{X} \neq \mu$

Exercise $8.5.2$

Answer in Words: Students who scored high on growth mindset would spend a similar amount of time on their homework as the population of junior high students.

Answer in Symbols: $\overline{X} = \mu$

This page titled 8.5: The Research Hypothesis and the Null Hypothesis is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 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.
- 7.3: The Research Hypothesis and the Null Hypothesis by Michelle Oja is licensed CC BY-NC-SA 4.0.

8.6: 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.

Historical Context

In 1935, Ronald A. Fisher's book *Design of Experiments* gives ways of deciding which results of scientific experiments are significant and which are not (Champkin, 2014). Unfortunately, Fisher used his theoretical breakthroughs and statistical knowledge to support fascism (Salsburg, 2002).


The
Design of Experiments

Figure 8.6.1: Title page for *The Design of Experiments* (1935) by Ronald Fisher from the Internet Archive. Fair use

In Dr. Navarro's experience, the best way to think about null hypothesis significance testing 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, s*omeone* 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 $\overline{X} \neq \mu$.
  - 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 $\overline{X} = \mu$.
  - 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 8.6.1

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

Answer available at the end of this section.

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.

Reference

Champkin, J. (2014, January 24). *The timeline of statistics*. Significance. https://www.statsref.com/timeline.pdf

Salsburg, D. (2002). *The lady tasting tea*. Henry Holt and Company.

Contributors

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

Answers

Exercise 8.6.1

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

This page titled 8.6: 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.

8.7: 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 8.7.1 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 8.7.1 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.

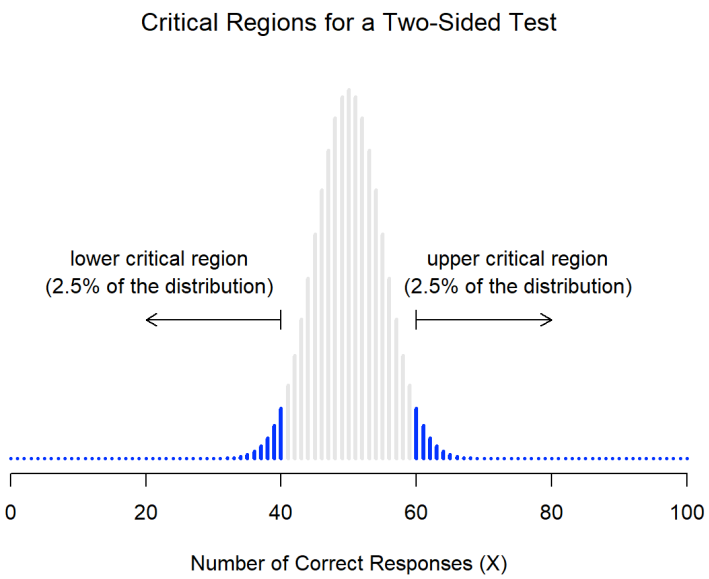


Figure 8.7.1- Example of Critical Regions for a Non-Directional Research Hypothesis (CC-BY-SA- Danielle Navarro from Learning Statistics with R)

Our critical region consists of the most *extreme values*, known as the *tails* of the distribution. The interactive visualization in the next sub-section may help understand this!

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 8.7.1.

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 8.7.2. 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 8.7.1.

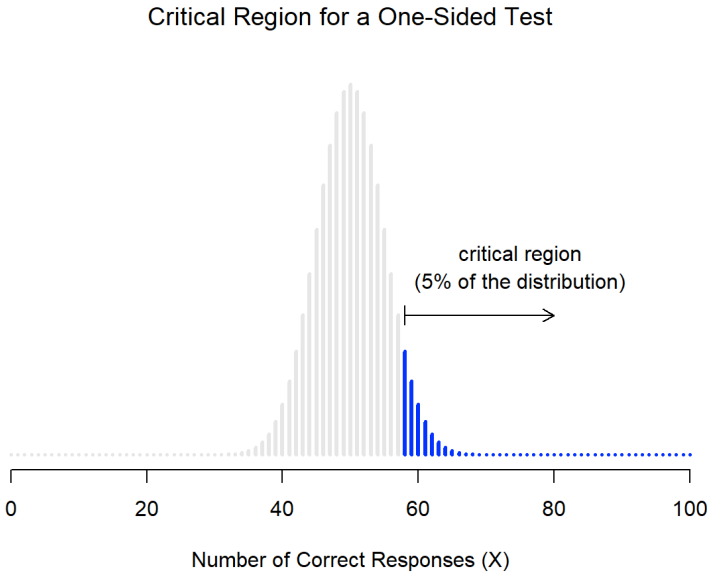


Figure 8.7.2- Critical Region for Directional Research Hypotheses (CC-BY-SA- Danielle Navarro from Learning Statistics with R)

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

8.7.1: Rejection Region Visualization

Visualize the Rejection Region for a Left, Right and Two Tailed Test

You can change to a left, right or two tailed test and adjust the level of significance and the rejection region will be shown.

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Graphic online

5

$\alpha = 0.05$

Alternative Hypothesis: ◉ $<$ ○ $>$ ○ $\neq$

This page titled 8.7.1: Rejection Region Visualization is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

8.8: Practice with 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 8.8.1

What does the null hypothesis say?

Answer available at the end of this section.

Example 8.8.1

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:

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

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

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 8.8.2

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:

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

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

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 8.8.2

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:

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

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

1. p > .05

2. p = .032

3. p = .049

Answer available at the end of this section.

Answers

Exercise 8.8.1

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.

Exercise 8.8.2

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 8.8: Practice with p-values and NHST is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

8.9: Steps of the Hypothesis Testing Process

Remember way back in the first chapter when we discussed the scientific method? Figure 8.9.1 shows the model, as a reminder.

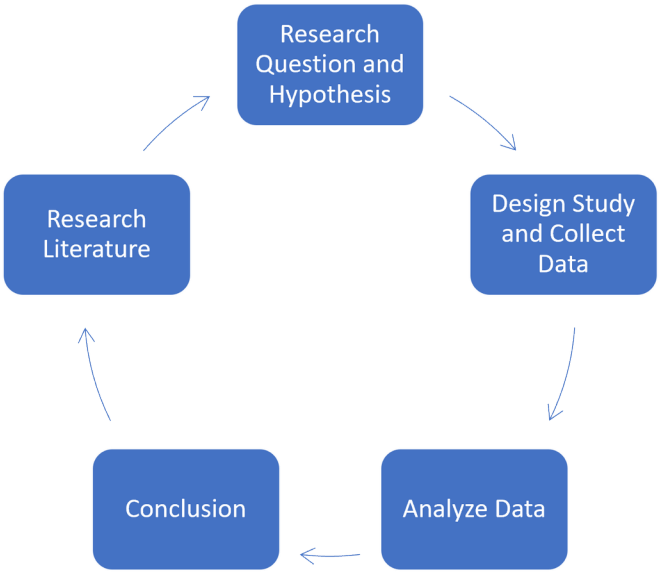


Figure 8.9.1: Model of Stages of the Scientific Method. (CC-BY-SA; Michelle Oja via source)

We're going to use that model to structure our process of null hypothesis significance testing (NHST) into steps. This process can be quite lengthy and involves a lot of pieces, so having a simple five-step model to refer to will help you put some of the puzzle pieces together. These steps will be what we use for the remainder of the textbook. The statistics we use will change, but this process will not.

Step 1: Hypotheses

In real life and real research, you would start with some questions about the world. In most examples in this course, you'll be provided data or a scenario to work with. Sometimes this scenario will include a prompt or research question. For those working through this five-step Null Hypothesis Significance Testing process, your first step will be to develop research hypotheses and null hypotheses. These hypotheses will usually be stated mathematically in symbols AND in words, with the words explaining what each symbols represents for the specific research scenario. These hypotheses will look different for each different type of statistical analysis used, but we will always state the research hypothesis (what we expect or are predicting) and the null hypothesis to remind us what we are testing (that nothing happened).

Step 2: Collect Data

Most versions of this course focus on analyzing data, designing studies, and collecting data. This step reminds us to identify what variables were collected and who the data was collected on.

Step 3: Analyze Data

Once we have our hypotheses and data, we can do calculations! 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: Conclusion

To conclude whether or not the research hypothesis is supported, we first need to find a critical value to compare our calculated statistic to in order to decide whether or not the null hypothesis is retained or rejected. Okay, that's not entirely true. If you are using statistical software, the software provides a p-value (probability that the null hypothesis is true). Typically, we use p<.05 (less than a 5% probability that nothing is happening) to reject the null hypothesis. Remember:

Reject null: $(\overline{X} \neq \mu) = p < .05$
Retain null: $(\overline{X} = \mu) = p > .05$

Whether we compare our calculated value to a critical value to determine if p<.05 or p>.05 (if the probability is higher than 5% or lower than 5% that the null hypothesis is true in the population), or statistical software provides a p-value, we use that information to decide whether the null hypothesis should be retained or rejected.

If the null hypothesis is retained, then we can conclude that the research hypothesis was not supported.

Example 8.9.1

Why can we conclude that the research hypothesis is automatically not supported if the null hypothesis is retained?

Solution

The null hypothesis always says that nothing is happening, and the research hypothesis (RH) is always saying that something is happening. In research hypotheses with two means, the research hypothesis says that one group's mean is bigger than another group's mean, so we can't support that something is happening (that one group's mean is bigger than another group's mean) if we just concluded that nothing is happening (retain the null).

If the null hypothesis is rejected, then we have one more step in making a conclusion. By rejecting the null hypothesis, we're saying that something happened. But in directional research hypothesis, we expected the results to be in a certain direction. For example, a directional research hypothesis with means would state that the mean number of times that a student re-tries an Example after they don't get quite the same number is higher for those with growth mindset than those with a fixed mindset. This research hypothesis could look like:

$$\overline{X}_{GM} > \overline{X}_{FM}$$

To support this research hypothesis, the means should actually be in that direction! Rejecting the null hypothesis says that the means are different, but we can't support a directional research hypothesis unless these means are different in the direction that we hypothesized.

Once you've made the decision, share what you've learned!

Step 5: Write-Up

Your answer is never just a number in behavioral statistics. In Null Hypothesis Significance Testing, your answer should include information about the research hypothesis and your calculated results.

First, the research hypothesis should be re-stated (in words), and then state your conclusion (from the previous step) about whether the research hypothesis was supported or not. Note that you don't have to mention the null hypothesis here at all! However, when a directional research hypothesis is not supported, it helps to clarify whether the research hypothesis was not supported because the means were in the opposite direction of what was predicted or because the null hypothesis was retained. When the research hypothesis is supported, most researchers don't mention the null hypothesis.

Second, your write-up should include your calculated results. When comparing means, this includes stating the mean of each group with the units of the variable that was measured. For example, if you measured final grade percentages, then the unit would

be percentages. If you measured the number of classes missed, then the unit would be absences or percentage of classes missed. If you measured scores on a quiz, the units might be points. The other calculated results would be the calculated statistic, e.g., your math results. Each chapter from now on will discuss a different type of statistical analysis, and the results for each formula would be your calculated statistic. You'll learn more in next chapter about how to format these results, but for now just remember that you'll need to include that information, too.

Conclusion

These processes are summarized in Table 8.9.1.

Table 8.9.1- Steps of the NHST Process and What to Do

Step of the Null Hypothesis Significance Testing Process	What To Do
1. Hypotheses	State a directional research hypothesis in words and symbols (include the groups being compared, the DV, and the direction word) State the null hypothesis in words and symbols.
2. Collect Data	You may not need to literally collect data, but you'll always want to determine: Who is the sample? Who is the population? What is being measured? What are the groups (if any)?
3. Analyze Data	Do the math!
4. Conclusion	Determine whether the research hypothesis is supported or not. To do that, you'll probably need to find a critical value to see if your null hypothesis is retained or rejected. If the null hypothesis is rejected, then determine if the direction of the actual means is the same as the expected direction in the research hypothesis (Is the mean of the group that you said would be bigger actually the bigger mean?).
5. Write-Up	Share your results with others by: Restating the research hypothesis and stating whether or not it was supported, and Providing the calculated results, which will probably include the actual means (with units) but will definitely include the calculated statistic.

Dr. MO is excited to start putting this five-step Null Hypothesis Significance Testing process into practice! But first, there's at least one more piece about testing null hypotheses that we should think about: probability errors.

This page titled 8.9: 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.

8.10: 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 8.10.1 illustrates, after we run the test and make our choice, one of four things might have happened:

Table 8.10.1- *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 8.10.2- *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 8.10: 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.

8.11: 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 8.11: 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/.

8.12: Views of NHST

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 8.12: Views of NHST is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

9: ONE SAMPLE T-TEST

Let's start putting what you've learned to use by using null hypothesis significance testing to compare the mean of a sample to a population.

9.1: Introduction to One-Sample t-tests

9.2: One-Sample t-test Calculations

9.3: Reporting Results

9.4: Practice One-Sample t-tests on Mindset Scores

9.5: Confidence Intervals

This page titled 9: One Sample t-test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

9.1: 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.

Historical Context

In 1908, William Sealy Gosset, chief brewer for Guinness in Dublin, described the t-test (Champkin, 2014). It uses a small number of samples to ensure that every brew tastes equally good.



Figure 9.1.1: Plaque in the Guinness storehouse commemorating Gosset by Tim bates on Wikipedia, CC BY-SA 4.0

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

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

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

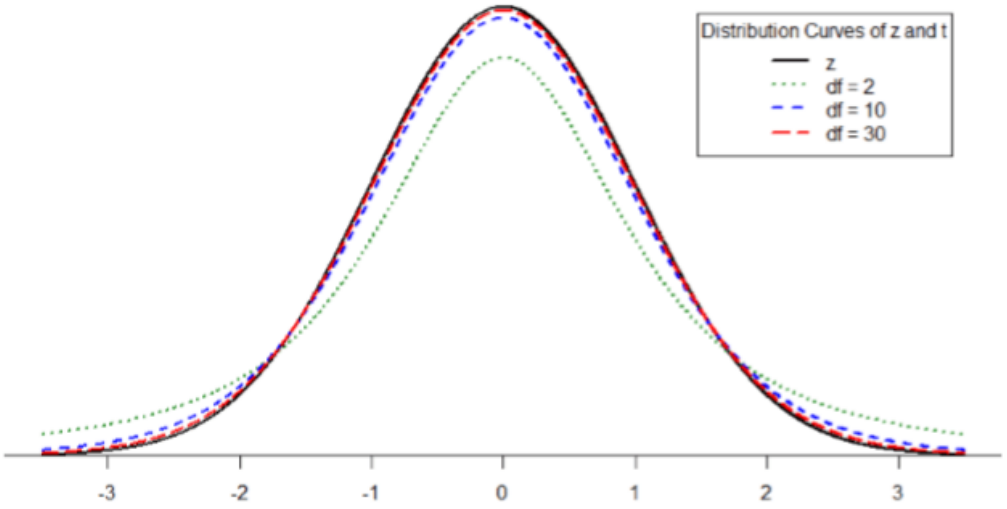


Figure 9.1.1: Distributions comparing effects of degrees of freedom. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

---

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

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

The $s_{\overline{X}}$ 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.

Reference

Champkin, J. (2014, January 24). *The timeline of statistics*. Significance. https://www.statsref.com/timeline.pdf

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 9.1: Introduction to One-Sample t-tests is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

9.2: One-Sample t-test Calculations

Statistical analyses can be confusing. To make it easier, Dr. Foster and his colleagues use these four steps. For a thorough analysis, these steps should be completed after you review the descriptive statistics (measures of central tendencies of mean, median, and mode; standard deviation as a measure of variability) and the data graphed in a frequency chart.

Note

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

Practice Example

We will work though an example.

Scenario:

Imagine that you hear that most students spend about 30 minutes studying for their weekly quizzes. You ask four students who sit next to you in your behavioral statistics course how long they study for their weekly quizzes.

Step 1: State the Hypotheses

We still state the research hypothesis and the null hypotheses mathematically in terms of the population parameter and written out in readable English. For our example:

- Research Hypothesis: The students who sit next to me in my behavioral statistics course study for longer than the the population of students studies for their quizzes.
  - Note that this has all of the required components of a complete research hypothesis discussed in 7.3: The Research Hypothesis and Null Hypotheses:
    - The name of the groups being compared: The four people sitting next to you in your behavioral statistics class, compared to all students.
    - What was measured: Time spend studying
    - Which group are we predicting will have the higher mean: The sample of four people will be higher ("study for longer").
  - Symbols: $\overline{X} > \mu$
- Null Hypothesis: The students who sit next to me in my behavioral statistics course study for about the same amount of time as the population of students study for their quizzes.
  - Symbols: $\overline{X} = \mu$

Step 2: Find the Critical Values

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. In almost all cases, degrees of freedom (or df) are N-1. So, for this example:

$$df = N–1 = 4–1 = 3$$

You can find the table of critical t-scores page, or go to the Common Critical Values page at the end of the book (in Back Matter) to find a link to all of the critical value tables that we'll be using.

Going to our table of critical t-scores page, we find the the number 4 in the Degrees of Freedom column (all the way on the left), then follow that to the p = .05 column (middle column). Doing that, our critical value is t = 2.353. You can use Figure 9.2.1 to visualize our rejection region.

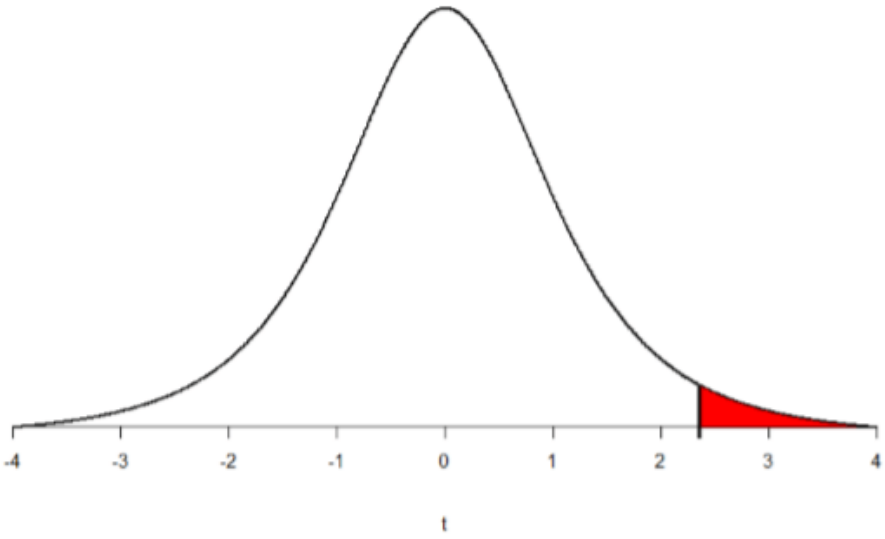


Figure 9.2.1: Rejection Region (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Step 3: Compute the Test Statistic

The answers to how long they study for the weekly quiz last week from your four classmates are below in Table 9.2.1. You can use these to calculate $\overline{X}$ and s by filling in the rest of the Table 9.2.1. These are worked out in the Exercise 9.2.1.

Table 9.2.1: Sum of Squares Table

$\overline{X}$	$X - \overline{X}$	$(X - \overline{X})^2$
46		
58		
40		
71		
Σ=215	Σ=0	Σ= ???

Exercise 9.2.1

Use the four data points in the table to determine the mean and standard deviation of the sample of four classmates.

Answer available at the end of this section.

Okay, now that we have the mean of the sample ($\overline{X} = 53.75$) and a standard deviation for the sample ($s = 13.72$), and we still know N (and the Degrees of Freedom), we plug it all into the t-test formula!

$$t = \frac{(\overline{X} - \mu)}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{(53.75 - 30)}{\left(\frac{13.72}{\sqrt{4}}\right)} = \frac{(23.75)}{\left(\frac{13.72}{2}\right)} = \frac{23.75}{6.86} = 3.46$$

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. 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

This is a behavioral statistics course, so you're not done!

Now that we have our critical value and test statistic, we can make our decision. Our calculated t-statistic was t = 3.46 and our critical value was $t = 2.353$. So, do we retain the null hypothesis or reject the null hypothesis? Since our calculated t-score was more extreme than our critical value, we reject the null hypothesis.

Let's add these critical and calculated values to our cheat sheet on understanding how to make the decision:

Note

$$(Critical < |Calculated|) = \textbf{Reject null} = \textbf{means are different} \ = p < .05$$

$$(Critical > |Calculated|) = \textbf{Retain null} = \textbf{means are similar} \ = p > .05$$

This shows us that when the critical value from the table is smaller than the absolute value of our calcluated, then we reject the null hypothesis, say that the means are different (the higher mean is larger), and that p<0.05 (the probability that the sample mean is similar to the population mean is small).

We're still not done!

Write-Up

"Based on our sample of four classmates, the sample studied longer on average ($\overline{X}$ = 53.75 minutes) than the population of students (μ = 30 minutes), $t(3)$ = 3.46, p < 0.05."

Notice that we also include the degrees of freedom in parentheses after the t.

Students always get confused about which way the sign goes between the "p" and the 0.05. The first thing to remember is that it's like Pac-Man, and will eat whatever is bigger. The sample mean is so much larger than the population mean that p=.05 is bigger than the probability of getting this sample mean by chance if the sample really was similar to the population.

The next section talks about this write-up part in more detail.

Answers

Exercise 9.2.1

Table 9.2.2: Sum of Squares Table

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=564.75

Mean:

As the table shows, the sum of the time spent studying for our four classmates was 215 minutes. We can use this to find that the mean:

$$\overline{X} = \frac{215}{4} = 53.75$$

Our four classmates studied for the weekly quiz an average of 53.75 minutes.

Standard Deviation:

We use that mean to subtract each score from (middle column), then square each score (right column). Summing that last column, you should get a sum of squares of SS = 564.75/. If you are close (anywhere from 564.70 to 564.80), then your calculator or software are rounding slightly differently. Don't worry! Your answer is close enough.

We then plug in to the formula for standard deviation from chapter 3:

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}} = \sqrt{\frac{SS}{df}} = \sqrt{\frac{564.75}{4 - 1}} = \sqrt{\frac{564.75}{3}} = \sqrt{188.25} = 13.72$$

This page titled 9.2: One-Sample t-test Calculations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

9.3: Reporting Results

In the previous chapter, we discussed how each stage of the scientific method (originally discussed in the first chapter) is applied in the Null Hypothesis Significance Testing process. This section provides a little more detail about the Write-Up step. Remember that when conducting statistics for the social sciences, the answer is never just the number. In the social sciences, 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.

You'll see a sample write-up below, but let's introduce some pieces before you read through the sample.

- *Provide the means first*. That is, you can tell the reader something about what the data looks like before going on to describing the results of 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 we're going to practice this 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 (just) in symbols. 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, the write-up won't just say that the result was not statistically significant, there's a "statistical sentence" (the dense mathematical-looking part in the parentheses in the sample below) which reports all the statistical results. For the t-test, the information that gets reported is the test statistic result (the calculated t-score), 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, include an interpretation of the result about 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.

Now that you have a sense of what a one-sample t-test analysis (from the previous section) and its write-up (from above) might look like, let's review a writing up of the results of a scenario an example with playing a card game.

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 ($\overline{X} = .55$ (my friends and I won 55% of the time), my research hypothesis is that the sample had a higher win rate than the population ($\overline{X} > \mu$). 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% ($\overline{X} = .55$), while the population average win rate was 43% ($\mu = .43$). We hypothesize that the sample (Dr. Navarro and her five friends) would have a higher average win rate than the population ($\overline{X} > \mu$). A one-sample t-test found that the sample's win rate is similar to the population's win rate. ($t(3) = 1.49$, $p > 0.05$). The research hypothesis that the sample would have a higher win rate than the population was not supported.*

Anything Missing?

The section on write-ups in the previous chapter indicated that the following should be included; were they?

- Restating the research hypothesis and stating whether or not it was supported, and
- Providing the calculated results, which will probably include the actual means (with units) but will definitely include the calculated statistic.

Example 9.3.1

1. Was the research hypothesis (re-)stated in the write-up?
2. Did the write-up include whether or not the research hypothesis was supported?
3. Were the means of each group shared? Were the units of what was measured for the means stated?
4. Was the calculated statistic include? In other words, was there a statistical sentence showing the results?

Solution

1. Yes, in both words and symbols as the write-up says "*We hypothesize that the sample (Dr. Navarro and her five friends) would have a higher average win rate than the population* ($\overline{X} > \mu$)."
2. Yes, as the write-up says "*the research hypothesis that the sample would have a higher win rate than the population was not supported*".
3. Yes, the means for both groups were named in the sentence and in symbols.
4. Yes, the results of the statistical analysis were provided in a statistical sentence.

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 she has 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.

In sum, write-ups should include:

- Research Hypothesis, including:
  - Re-state the RH (including the groups being compared, the DV, and the direction word), and
  - If the RH was supported or not.
- Calculation results:
  - Calculated statistical (statistical sentence):
    - *t*(df) = calculated t-score, *p*<.05
    - or
    - *t*(df) = calculated t-score, *p*>.05
  - Means, with units, of each group (name the groups)

To put it all together, 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 9.3: Reporting Results is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

9.4: Practice One-Sample t-tests on Mindset Scores

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 interpret what these descriptive statistics can show us!

Descriptive Statistics

Measures of Central Tendency

The measures of central tendency are shown in Table 9.4.1.

Table 9.4.1- *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 9.4.1

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 available at the end of this section.

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 9.4.2

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 available at the end of this section.

I literally was just guessing in Exercise 9.4.3. Let's look at the actual frequencies to see if I guessed correctly!

Frequency

Table 9.4.2 is a frequency table for this sample's mindset scores, and Figure 9.4.1 shows the same information in a line graph.

Table 9.4.2- 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

X (Mindset Score)	Frequency
56	1
57	1
60	2

And Figure 9.4.1:

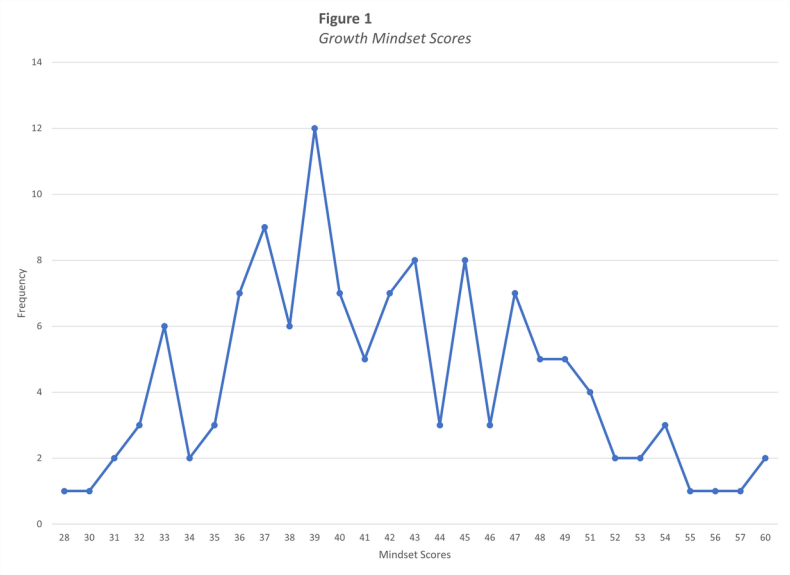


Figure 9.4.1: Frequency Line Graph of Mindset Scores. (CC-BY-NC-SA; Michelle Oja)

Let's start unpacking Figure 9.4.1.

Exercise 9.4.3

1. What kind of graph is Figure 9.4.1?
2. What does the x-axis measure in Figure 9.4.1?
3. What does the y-axis measure in Figure 9.4.1?
4. Is Figure 9.4.1 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 9.4.1? What pops out to you?
6. What does Figure 9.4.1 make you wonder about?
7. What is a catchy headline for Figure 9.4.1?
8. How could you summarize the info in Figure 9.4.1 into one sentence?
9. Who might want to know the information in Figure 9.4.1?

Answer available at the end of this section.

Okay, now that we've looked at Figure 9.4.1 closely, let's think about it in relation to the measures of central tendencies.

Exercise 9.4.4

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 available at the end of this section.

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

We'll go through each of the five steps of the Null Hypothesis Significance Testing process for this scenario. Here we go!

Step 1: Hypotheses

When the student researchers first started on this research project, we were very surprised by how high the mindset scores started 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: $\overline{X} > \mu$

Exercise 9.4.5

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

Answer available at the end of this section.

Step 2: Collect Data

As the data was already collected, it's now your job to describe what we know already.

Exercise 9.4.6

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?
4. What is the IV? In other words, what are the two groups that are being compared?

Answer available at the end of this section.

That concludes the second step of the NHST process. When you don't have to collect your own data, this step only requires you to identify the sample, population, measured outcome (DV), and groups (IV).

Step 3: Analyze Data

Okay, you have everything you need to calculate the one-sample t-test! To make sure it's all in one place, Table 9.4.3 puts all of the numbers in one place. This might be a good practice everytime that you conduct a statistical analysis...

Table 9.4.3- 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

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

Step 4: Conclusion

Using the most common level of significance is α = 0.05 and the table of critical values for the t-distribution, find the critical value that we should use to compare the calculated t-score (completed in the previous step).

Example 9.4.1

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:

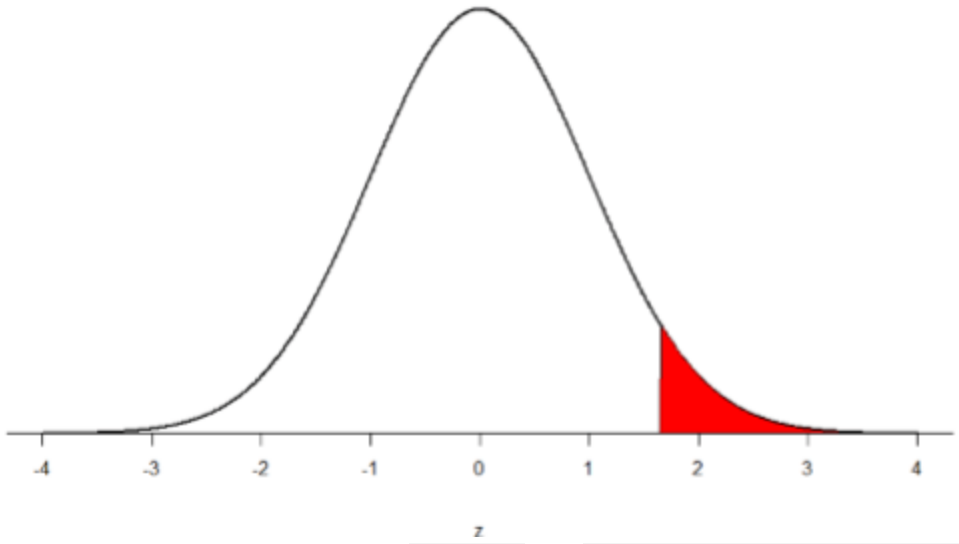


Figure 9.4.2: Rejection region (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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.

To determine if our research hypothesis is supported or not, we first decide if the null hypothesis is rejected or retained.

Remember:

$$(Critical < |Calculated|) = \textbf{Reject null} \ = (\overline{X} \neq \mu) = p < .05$$
$$(Critical > |Calculated|) = \textbf{Retain null} \ = (\overline{X} = \mu) = p > .05$$

So, we reject the null hypothesis.

Now that we know that the means are statistically different, we review our research hypothesis and the actual means to see if our research hypothesis is support or not. Our research hypothesis predicted that the average mindset score of our sample would be higher than the average mindset score of the population. As can be seen in Table 9.4.3, the mean of the sample was mathematically higher. As we rejected the null hypothesis, this means that the mean of the sample was also statistically significantly higher.

Step 5: Write-Up

The last step of our Null Hypothesis Significance Testing process is to share our findings in a write-up.

Exercise 9.4.7

Share your results with others by describing the following in a paragraph:

- Restating the research hypothesis and stating whether or not it was supported, and
- Providing the calculated results, includinc the actual means (with units) and the calculated statistic in a statistical sentence.

Answer available at the end of this section.

Cool, huh?!

Answers

Exercise 9.4.1

The mean ($\overline{X} = 42.06 points$) 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.

Exercise 9.4.2

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.

Exercise 9.4.3

1. What kind of graph is Figure 9.4.1?
    1. Figure 9.4.1 is a line graph showing frequencies.
2. What does the x-axis measure in Figure 9.4.1?
    1. The x-axis measures mindset scores.
3. What does the y-axis measure in Figure 9.4.1?
    1. The y-axis measures frequencies of each mindset score.
4. Is Figure 9.4.1 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 9.4.1? 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 9.4.1 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 9.4.1?
    1. Community College Students Start with Growth!
8. How could you summarize the info in Figure 9.4.1 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 9.4.1?
    1. Definitely community college faculty and adminstrators, as well as researchers who want to learn about growth mindset or improve student performance.

Exercise 9.4.4

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.

Exercise 9.4.5

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

Exercise 9.4.6

1. The sample is 126 community college students.
2. Any reasonable population based on the sample, but probably community college students in the U.S.
3. The DV (outcome that was measured) was scores on the mindset assessment.
4. The IV (groups) were the sample of community college students and the population.

Exercise 9.4.7

I hypothesized that the sample's mindset scores ($\overline{X} = 42.06 points$ 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.

This page titled 9.4: Practice One-Sample t-tests on Mindset Scores is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

9.5: Confidence Intervals

All statistics have sampling error, which means 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. To do this, we calculate what is known as a confidence interval.

A confidence interval starts with our sample mean, called a point estimate here, then creates a range of scores (this is the "interval" part) 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 t-score by our standard error:

$$\text{Margin of Error } = t \times \left(\frac{s}{\sqrt{N}}\right)$$

The critical value we use will be based on a chosen level of confidence, which is equal to $1–\alpha$. Thus, a 95% level of confidence corresponds to $\alpha = 0.05$. 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:

$$\text{Upper Bound} = \overline{X} + \text{Margin of Error}$$
$$\text{Lower Bound } = \overline{X} - \text{Margin of Error}$$

Or simply:

$$\text{Confidence Interval } = \overline{X} \pm (t \times \left(\frac{s}{\sqrt{N}}\right))$$

Let's see what this looks like with some actual numbers by using studying for weekly quizzes data to create a 95% confidence interval estimating the average length of study time for our sample. We already found that our average was $\overline{X}$= 53.75 minutes, our standard error (the denominator) was 6.86, and our critical t-score was 2.353. With that, we have all the pieces we need to construct our confidence interval:

$$95\%CI = 53.75 \pm 2.353(6.86) = 53.75 \pm 16.14$$

$$\text{Lower Bound (LB)} = 53.75 - 16.14 = 37.61$$

$$\text{Upper Bound (UB)} = 53.75 + 16.14 = 69.88$$

$$95\%CI = (37.61, 69.88)$$

So we find that our 95% confidence interval runs from 37.61 minutes to 69.88 minutes, but what does that actually mean? The range (37.61 to 69.88) 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 (37.61 to 69.88), 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 (37.61 to 69.88) brackets the true population mean. This is a very subtle difference, but it is an important one.

Hypothesis Testing with Confidence Intervals

Once a confidence interval has been constructed, using it to test a hypothesis is simple. When the range of the confidence interval brackets (or contains, or is around) the null hypothesis value, we retain the null hypothesis. If the interval 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 (retain). 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.

Scenario

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: Hypotheses

Start by laying out the research hypothesis and null hypothesis. Testing null hypotheses with Confidence Intervals requires that the research hypothesis is non-directional. 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.

- Research Hypothesis: There is a difference in how friendly the local community is compared to the national average
  - Symbols: $\overline{X} \neq \mu$
- Null Hypothesis: There is no difference in how friendly the local community is compared to the national average
  - Symbols: $\mu = 38$

Step 2: Collect Data

Imagine that we collected data and found that the average person in our community scored 39.85, or $\overline{X} = 39.85$, and our standard deviation was $s = 5.61$. Additionally, let's identify some basic information about the variables and who it was collected on by answer the questions in Exercise $9.5.1$; these questions may become very familiar to you!

Exercise $9.5.1$

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?
4. What is the IV? In other words, what are the two groups that are being compared?

Answer available at the end of this section.

Step 3: Analyze Data

In our one-sample t-test examples (and most other statistical analyses that you'll learn), you can find the critical value *after* you analyze the data. But in confidence intervals, you actually need 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). A *two-tailed (non-directional)* critical value at $\alpha = 0.05$ is actually p=0.025 on the table of critical values for t. With 29 degrees of freedom ($N-1 = 30-1 = 29$) and p-value of 0.025, the critical t-score is 2.045.

Now we can construct our confidence interval. We can put our sample mean, and our critical value into the formula for a

confidence interval:

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

$$95 = \overline{X} \pm (t \times \left(\frac{s}{\sqrt{N}}\right)) = 39.85 \pm (2.045 \times \left(\frac{5.61}{\sqrt{30}}\right)) = 39.85 \pm (2.045 \times \left(\frac{5.61}{5.48}\right)) = 39.85 \pm (2.045 \times 1.02) = 39.85 \pm (2.09)$$

$$\text{Lower Bound} = 39.85 - 2.09 = 37.76$$

$$\text{Upper Bound} = 39.85 + 2.09 = 41.94$$

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

Step 4: Conclusion

In this step, we 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 retain (fail to reject) the null hypothesis.

Step 5: Write-Up

The write-up is a little simpler for confidence intervals, too; 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.

> 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). This supports the research hypothesis.

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.

But what if you're in the very common predicament in which you don't have the population mean? You'll learn other t-tests to compare two different samples, or groups, next!

Answers

Exercise $9.5.1$

1. The sample is 30 people in the town.
2. Any reasonable population based on the sample, but probably community college students in the U.S.
3. The DV (outcome that was measured) was an assessment of friendliness.
4. The IV (groups) was the people in the town compared to the population of the nation.

This page titled 9.5: Confidence Intervals 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.

10: INDEPENDENT SAMPLES T-TEST

10.1: Introduction to Independent Sample t-tests
10.2: Independent Samples t-test Equation
10.3: Hypotheses with Two Samples
10.4: Practice Independent t-test on Movies and Mood
10.5: Practice Independent t-test on Growth Mindset
10.6: When to NOT use the Independent Samples t-test
10.7: Non-Parametric Independent Sample t-Test

This page titled 10: Independent Samples t-test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

10.1: Introduction to Independent Sample t-tests

Although the one sample t-test has its uses, it's not the most typical example of a t-test. A much more common situation arises when you've got two different groups of observations. In psychology, this tends to correspond to two different groups of participants, where each group corresponds to a different condition in your study. For each person in the study, you measure some outcome variable of interest (DV), and the research question that you're asking is whether or not the two groups have the same population mean (each group is considered a level of the IV). This is the situation that the independent samples t-test is designed for.

Example

Suppose we two tutors (Anastasia and Bernadette) for a class of 33 students taking a behavioral statistics course, with 15 students being randomly assigned to Anastasia as a tutor and 18 students who were assigned to Bernadette. The research question that the professor might be interested in is whether Anastasia or Bernadette is a better tutor, or if it doesn't make much of a difference. Table 10.1.1 shows some descriptive statistics related to course grades.

Table 10.1.1- *Descriptive Statistics for Grades by Tutor*

Tutor	Mean	Standard Deviation	N
Anastasia's students	74.53	9.00	15
Bernadette's students	69.06	5.77	18

To give you a more detailed sense of what's going on here, Dr. Navarro plotted histograms showing the distribution of grades for both tutors (Figure 10.1.1 and Figure 10.1.2). Inspection of these histograms suggests that the students in Anastasia's class may be getting slightly better grades on average, though they also seem a little more variable.

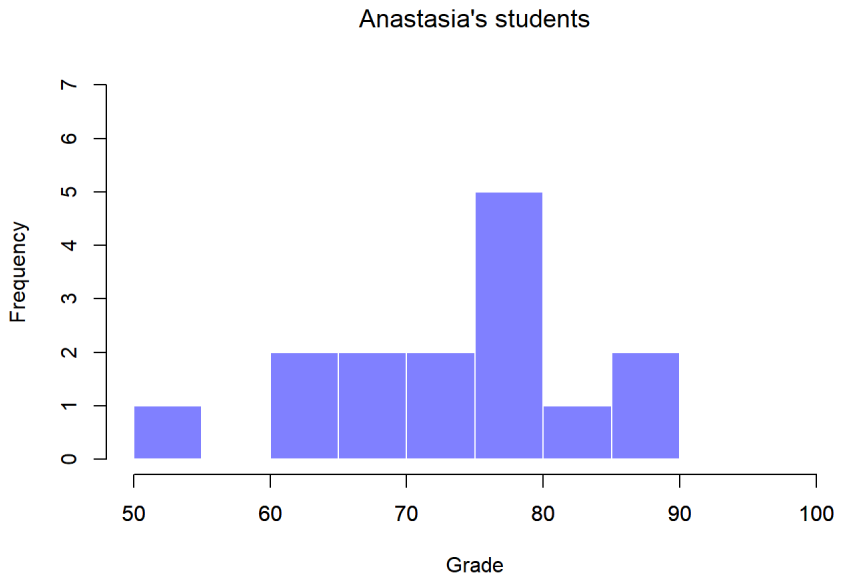


Figure 10.1.1- Histogram showing the overall distribution of grades for students in Anastasia's class (CC-BY-SA Danielle Navarro from Learning Statistics with R)

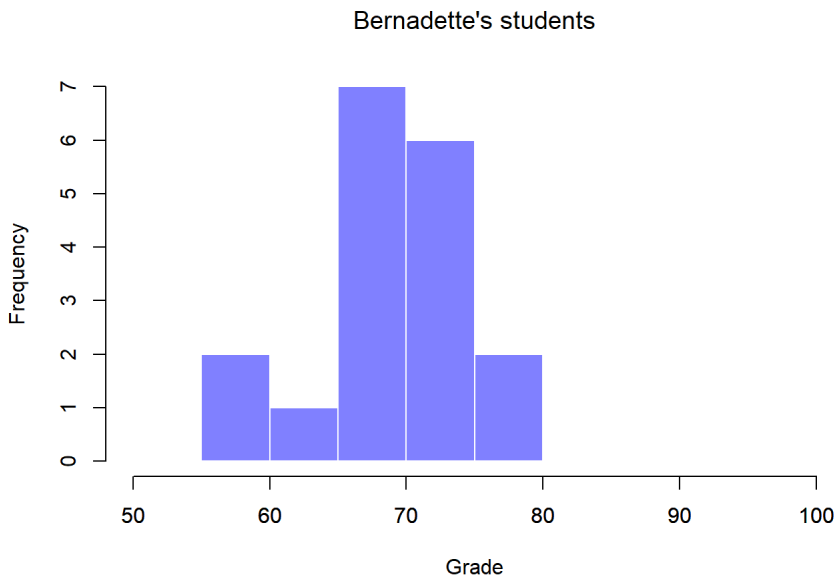


Figure 10.1.2- Histogram showing the overall distribution of grades for students in Bernadette's class (CC-BY-SA Danielle Navarro from Learning Statistics with R)

Here is a simpler plot showing the means and corresponding confidence intervals for both groups of students (Figure 10.1.3). Note that Figure 10.1.3 is not the typical frequency chart that you've become familiar with, but instead shows the mean (with "whiskers" showing one standard deviation around the mean). On the basis of visual inspection of Figure 10.1.3, it does look like there's a real difference between the groups, though it's hard to say for sure since there is some overlap of the "whiskers" that show the variability.

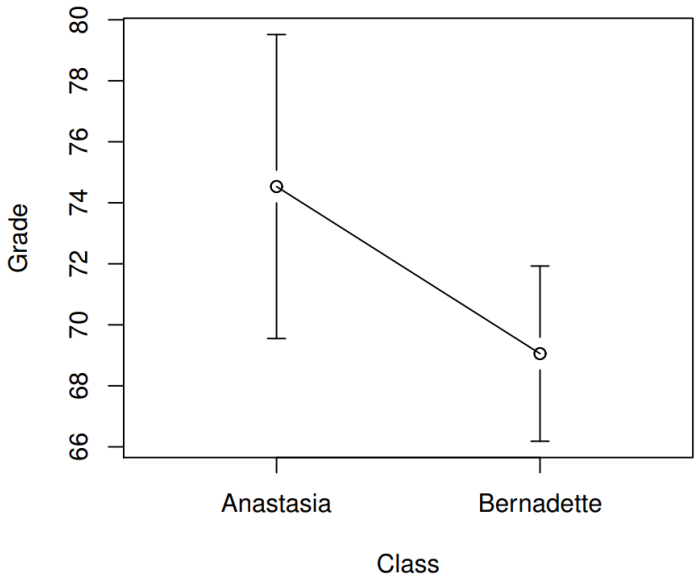


Figure 10.1.3- Plots showing the mean grade for the students in Anastasia's and Bernadette's tutorials. Error bars depict 95% confidence intervals around the mean. (CC-BY-SA Danielle Navarro from Learning Statistics with R)

This is where an independent sample t-test would come in! We aren't comparing a sample's mean score to a population's mean score like we did last chapter with the one-sample t-test. Instead, we are comparing the means of two groups to see if they are from the same population (in other words, to see if the two groups are similar).

Independent Sample t-test

Let's talk about the Student t-test. The goal is to determine whether two "independent samples" of data are drawn from populations with the same mean (the null hypothesis) or different means. When we say "independent" samples, what we really

mean here is that there's no special relationship between observations in the two samples. This probably doesn't make a lot of sense right now, but it will be clearer when we come to talk about the paired samples t-test later on. For now, let's just point out that if we have an experimental design where participants are randomly allocated to one of two groups, and we want to compare the two groups' mean performance on some outcome measure, then an independent samples t-test (rather than a paired samples t-test) is what we're after.

Okay, so let's let μ_1 denote the true population mean for group 1 (e.g., Anastasia's students), and μ_2 will be the true population mean for group 2 (e.g., Bernadette's students), and as usual we'll let $\overline{X}_1$ and $\overline{X}_2$ denote the observed sample means for both of these groups. Our null hypothesis states that the two population means are identical ($\mu_1=\mu_2$) and the alternative to this is that they are not ($\mu_1\neq\mu_2$). Figure $10.1.4$ shows this concept graphically. Notice that it is assumed that the population distributions are normal, and that, although the alternative hypothesis allows the group to have different means, it assumes they have the same standard deviation.

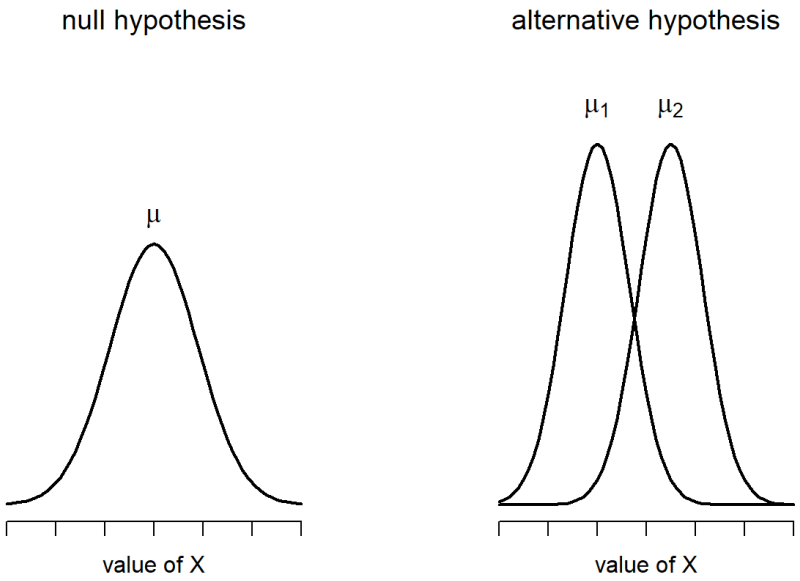


Figure $10.1.4$- Graphical illustration of the null and alternative hypotheses assumed by the Student t-test. (CC-BY-SA Danielle Navarro from Learning Statistics with R)

To construct a hypothesis test that handles this scenario, we start by noting that if the null hypothesis is true, then the difference between the population means is *exactly* zero, $\mu_1 - \mu_2 = 0$ As a consequence, a diagnostic test statistic will be based on the difference between the two sample means. Because if the null hypothesis is true, then we'd expect

$$\overline{X}_1 \text{ - } \overline{X}_2$$

to be *pretty close* to zero. However, just like we saw with our one-sample t-tests, we have to be precise about exactly *how close* to zero this difference

$$t = \frac{\overline{X}_1 - \overline{X}_2}{SE}$$

We just need to figure out what this standard error estimate actually is. This is a bit trickier than was the case for either of the two tests we've looked at so far, so we need to go through it a lot more carefully to understand how it works. We will do that later when discussing the actual formula. For now, just recognize that you're using something like a standard deviation to divide everything by in order to create something like an average difference between the two groups.

Results

If we used statistical software to calculate this t-test, we'd get a statistical sentence of: t(31)=2.12, p<.05 This is telling you that the results for the calculated t-test was 2.12, that the degrees of freedom are 31, and that the probability is less than 5% that the groups are similar (that they are from the same popuolation). The degrees of freedom might look new. You can think of the degrees of freedom to be equal to the number of data points minus the number of constraints. In this case, we have N observations (N_1 in sample 1, and N_2 in sample 2), and 2 constraints (the sample means). So the total degrees of freedom for this

test are $N_1 + N_2 - 2$. These results tell you that the difference between the two groups is statistically significant (just barely), so we might write up the result using text like this:

The mean grade in Anastasia's class was 74.5% (s=9.0), whereas the mean in Bernadette's class was 69.1% (s=5.8). A Student's independent samples t-test showed that this difference was significant (t(31)=2.12, p<.05), suggesting that a genuine difference in learning outcomes has occurred.

Positive and Negative t-scores

Before moving on to talk about the assumptions of the t-test, there's one additional point I want to make about the use of t-tests in practice. The first one relates to the sign of the t-statistic (that is, whether it is a positive number or a negative one). One very common worry that students have when they start running their first t-test is that they often end up with negative values for the t-statistic, and don't know how to interpret it. In fact, it's not at all uncommon for two people working independently to end up with results that are almost identical, except that one person has a negative t values and the other one has a positive t value. This is perfectly okay: whenever this happens, what you'll find is that the two versions are based on which sample mean you put as $\overline{X}_1$ and which is $\overline{X}_2$. If you always put the bigger sample mean first, then you will always have a positive t-score.

Okay, that's pretty straightforward when you think about it, but now consider our t-test comparing Anastasia's class to Bernadette's class. Which one should we call "mean 1" and which one should we call "mean 2". It's arbitrary. Whenever I get a significant t-test result, and I want to figure out which mean is the larger one, I don't try to figure it out by looking at the t-statistic. Why would I bother doing that? It's foolish. It's easier just look at the actual group means!

Here's the important thing. Try to *report* the t-statistic in such a way that the numbers match up with the text. Here's what I mean... suppose that what I want to write in my report is "Anastasia's class had higher grades than Bernadette's class". The phrasing here implies that Anastasia's group comes first, so it makes sense to report the t-statistic as if Anastasia's class corresponded to group 1. If so, I would write

Anastasia's class had higher grades than Bernadette's class (t(31)=2.12,p<.05).

(I wouldn't actually emphasize the word "higher" in real life, I'm just doing it to emphasize the point that "higher" corresponds to positive t values). On the other hand, suppose the phrasing I wanted to use has Bernadette's class listed first. If so, it makes more sense to treat her class as group 1, and if so, the write up looks like this:

Bernadette's class had lower grades than Anastasia's class (t(31)=−2.12,p<.05).

Because I'm talking about one group having "lower" scores this time around, it is more sensible to use the negative form of the t-statistic. It just makes it read more cleanly.

One last thing: please note that you *can't* do this for other types of test statistics. It works for t-tests, but don't overgeneralize this advice! I'm really just talking about t-tests here and nothing else!

Assumptions of the t-test

As always, our statistical test relies on some assumptions. So what are they? For the Student t-test there are three assumptions:

- *Normality*. Like the one-sample t-test, it is assumed that the data are normally distributed. Specifically, we assume that both groups are normally distributed. In Section 13.9 we'll discuss how to test for normality, and in Section 13.10 we'll discuss possible solutions.
- *Independence*. Once again, it is assumed that the observations are independently sampled. In the context of the Student test this has two aspects to it. Firstly, we assume that the observations within each sample are independent of one another (exactly the same as for the one-sample test). However, we also assume that there are no cross-sample dependencies. If, for instance, it turns out that you included some participants in both experimental conditions of your study (e.g., by accidentally allowing the same person to sign up to different conditions), then there are some cross sample dependencies that you'd need to take into account.
- *Homogeneity of variance* (also called "homoscedasticity"). The third assumption is that the population standard deviation is the same in both groups. Statistical software can test this assumption using the Levene test, but nothing that you really need to worry about right now.

Contributors and Attributions

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

This page titled 10.1: Introduction to Independent Sample t-tests is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

10.2: Independent Samples t-test Equation

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 (one mean subtracted from the other mean), all divided by the standard error:

$$t = \frac{(\overline{X_1} - \overline{X_2})}{SE}$$

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 as the information needed in a complete conclusion.

The following explains the conceptual mathematics behind this new denominator. If this makes sense to you, it will help you to understand what the t-test is doing so that it's easier to interpret. However, it's not necessary to understand the mathematical reasoning that lies behind the standard error to calculate it. To calculate, you just need to plug in the correct numbers and follow the order of operations!

Estimated Standard Error

If you are here for the reasoning underlying the denominator, here you go!

Recall that the standard error is the average distance between any given sample mean and the center of its corresponding sampling distribution (a distribution of means from many samples from the same population), and it is a function of the standard deviation 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 (variance is the standard devation squared) – 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:

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

This formula can look daunting at first, but it is in fact just a weighted average. Note that the subscript (the little number sorta under and to the right of some of the symbols) denotes which sample the symbol is representing (the first sample or the second sample).

Unfortunately, that is just part of the denominator. Once we have our pooled variance calculated, we can drop it into the equation for our standard error:

$$s_p^2 = \sqrt{\left[\frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}\right] * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

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:

$$t = \frac{(\overline{X_1} - \overline{X_2})}{SE}$$

Independent t-test Formula

As you can see, we are once again, not done yet! This is denominator. And once again, although this formula is different, it is accomplishing the same task of standardizing and averaging the differences between the mean.

The final formula to compare two independent means with a t-test is:

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\left[\frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}\right] * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

I promise that this is not as hard as it looks! Let's see an example in action. If you lose track of this page, remember that all formulas are listed at the back of the book in the Common Formulas page.

This page titled 10.2: Independent Samples t-test Equation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

10.3: Hypotheses with Two Samples

The process of testing hypotheses using an independent samples t-test is the same as it was when comparing a sample to a population with the one-sample t-test. It all starts with stating our hypotheses and laying out the criteria we will use to test them.

Research Hypotheses

Our research hypotheses are still predicting that one group will have a bigger mean than the other group (direction) or that the two means will be different (non-directional). The only difference in the research hypothesis between an independent samples t-test and a one-sample t-test is that instead of comparing a sample mean ($\overline{X}$) to a population mean (μ), we are now comparing two sample means (and no population mean). Because we have two sample means (two $\overline{X}$), we have to label them somehow. We do this with a subscript. Often this is a little number 1 or number 2, but when you know who the sample is, it's often helpful to use letters related to the group name as the subscript.

$$RH : \overline{X}_1 > \overline{X}_2$$
$$RH : \overline{X}_1 < \overline{X}_2$$
$$RH : \overline{X}_1 \neq \overline{X}_2$$

Null Hypotheses

Our null hypothesis for an independent samples t-test is the same as always: there is no difference between the means. The means of the two groups are the same under the null hypothesis, no matter how those groups were formed. In symbols, this looks like:

$$NH : \overline{X}_1 = \overline{X}_2$$

$$NH : \text{ There is no difference between the means of the two groups}$$

Your Research Hypothesis and the Null Hypothesis Work Together

Again, 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 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 an experimental medication example:

Words and Symbols:

- Research Hypothesis: The mean of the treatment group will be higher than the mean of the control group.
  - Symbols: $\overline{X}_{\text{treatment}} > \overline{X}_{\text{control}}$
- Null Hypothesis: There is no difference between the means of the treatment and control groups
  - Symbols: $\overline{X}_{\text{treatment}} = \overline{X}_{\text{control}}$

Next Step: Decision Criteria

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). For an independent samples t-test, the Degrees of Freedom 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 ($df = (N_1 - 1) + (N_2 - 1)$). 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.

- Dr. MO (Taft College)

This page titled 10.3: Hypotheses with Two Samples is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

10.4: Practice Independent t-test on Movies and Mood

Dr. Foster was interested in whether the type of movie someone sees at the theater affects their mood when they leave. He decide to ask people about their mood as they leave one of two movies: a comedy (group 1, N_C = 35) or a horror film (group 2, N_H = 29). The data are coded so that higher scores indicate a more positive mood.

Step 1: Hypotheses

As always, we start with hypotheses:

- Research Hypothesis: The average mood for the sample who watched a comedy will be higher than the average mood for the sample who watched a horror film.
  - Symbols: $\overline{X}_C > \overline{X}_H$
- Null Hypothesis: The average mood for the sample who watched a comedy will be similar to the average mood for the sample who watched a horror film. In order words, there will be no difference between the means.
  - Symbols: $\overline{X}_C = \overline{X}_H$

One thing to notice about the research hypothesis is that we put the mean of the comedy first. And since we think that mean will be bigger, our calculated t-score should be positive because we will have a larger number minus a smaller number. Be sure to pay attention to which group is which and how your data are coded to make sure your hypothesis makes sense; higher score usually indicate better outcome, but not always!

Step 2: Collect Data

Table 10.4.1 provides descriptive statistics about each group, which lets you know that the data has already been collected.

Table 10.4.1- *Descriptive Statistics for Mood by Type of Movie*

Group	N	Mean	Standard Deviation
Group 1: Comedy	35	24.00	12.20
Group 2: Horror Movie	29	16.50	11.80

You might also be able to identify the variables and the sample size from Table 10.4.1, although reading a little about the scenario (like the first paragraph in this section) can help, too. Let's use that information to answer the questions in Example 10.4.1.

Example 10.4.1

1. Based on the Table 10.4.1,, who is the sample? *Your answer should include the* ***total*** *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?
4. What is the IV? In other words, what are the two groups that are being compared?

Solution

1. The total sample is 64 movie-goers.
2. Any reasonable population based on the sample, but probably people who watch movies in a movie theater.
3. The DV (outcome that was measured) was an assessment of mood, with higher scores indicating a more positive mood.
4. The IV (groups) were the two types of movies: comedy or horror

Step 3: Analyze Data

The descriptive statistics for both groups are provided in Table 10.4.1 so that you can focus on the independent samples t-test equation in Example 10.4.1, and not get caught up in other calculations right now

Take a deep breath!

Example 10.4.1

Calculate an independent t-test using the descriptive statistics from Table 10.4.1.

After your deep breath, copy down the independent samples t-test formula:

$$t = \frac{(\overline{X}_C - \overline{X}_H)}{\sqrt{\left[\frac{(N_C - 1) * s_C^2 + (N_H - 1) * s_H^2}{N_C + N_H - 2}\right] * \left(\frac{1}{N_C} + \frac{1}{N_H}\right)}}$$

And then fill in all of the numbers from Table 10.4.1 into the formula:

$$t = \frac{(24.00 - 16.50)}{\sqrt{\left[\frac{(35 - 1) * (12.20)^2 + (29 - 1) * (11.80)^2}{35 + 29 - 2}\right] * \left(\frac{1}{35} + \frac{1}{29}\right)}}$$

At this point, you do the calculations by following the order of operations (PEMDAS from the Math Refresher in chapter 3!).

Solution

I like to start at the top right and work my way left and down because that's how reading goes in English. Let's try that by doing a few of the easy calculations first. As long as you follow the rules of the order of operations, you can do this in whatever order you'd like, though.

$$
\begin{aligned}
t &= \frac{(24.00 - 16.50)}{\sqrt{\left[\frac{(35-1)*(12.20)^2 + (29-1)*(11.80)^2}{35+29-2}\right]*\left(\frac{1}{35}+\frac{1}{29}\right)}} \\
t_{Add/Subtract} &= \frac{7.50}{\sqrt{\left[\frac{(34)*(12.20)^2 + (28)*(11.80)^2}{62}\right]*\left(\frac{1}{35}+\frac{1}{29}\right)}} \\
t_{Square} &= \frac{7.50}{\sqrt{\left[\frac{(34*148.84) + (28*139.24)}{62}\right]*\left(\frac{1}{35}+\frac{1}{29}\right)}} \\
t_{Multiply} &= \frac{7.50}{\sqrt{\left[\frac{(5060.56) + (3898.72)}{62}\right]*\left(\frac{1}{35}+\frac{1}{29}\right)}} \\
t_{AddAgain} &= \frac{7.50}{\sqrt{\left(\frac{8959.28}{62}\right)*\left(\frac{1}{35}+\frac{1}{29}\right)}} \\
t_{Divide} &= \frac{7.50}{\sqrt{(144.50)*(0.03+0.03)}} \\
t_{AddAgainAgain} &= \frac{7.50}{\sqrt{(144.50)*(0.06)}} \\
t_{MultiplyAgain} &= \frac{7.50}{\sqrt{8.67}} \\
t_{SquareRoot} &= \frac{7.50}{2.94} \\
t_{TheEnd!} &= 2.55
\end{aligned}
\tag{10.4.1}
$$

If you use Excel your final calculations will be slightly different because Excel keeps all of the numbers after the decimal point. For Excel, the calculation will look like:

$$
\begin{aligned}
t &= \frac{(24.00 - 16.50)}{\sqrt{\left[\frac{(35-1)*(12.20)^2 + (29-1)*(11.80)^2}{35+29-2}\right]*\left(\frac{1}{35}+\frac{1}{29}\right)}} \\
t_{Add/Subtract} &= \frac{7.50}{\sqrt{\left[\frac{(34)*(12.20)^2 + (28)*(11.80)^2}{62}\right]*\left(\frac{1}{35}+\frac{1}{29}\right)}} \\
t_{Square} &= \frac{7.50}{\sqrt{\left[\frac{(34*148.84) + (28*139.24)}{62}\right]*\left(\frac{1}{35}+\frac{1}{29}\right)}} \\
t_{Multiply} &= \frac{7.50}{\sqrt{\left[\frac{(5060.56) + (3898.72)}{62}\right]*\left(\frac{1}{35}+\frac{1}{29}\right)}} \\
t_{AddAgain} &= \frac{7.50}{\sqrt{\left(\frac{8959.28}{62}\right)*\left(\frac{1}{35}+\frac{1}{29}\right)}} \\
t_{Divide} &= \frac{7.50}{\sqrt{(144.50)*(0.03+0.03)}} \\
t_{AddAgainAgain} &= \frac{7.50}{\sqrt{(144.50)*(0.06)}} \\
t_{MultiplyAgain} &= \frac{7.50}{\sqrt{9.11}} \\
t_{SquareRoot} &= \frac{7.50}{3.02} \\
t_{TheEnd!} &= 2.48
\end{aligned}
\tag{10.4.2}
$$

Phew!

A few things to notice. I added subscripts inappropriately to indicate what I was doing on each step. Also, on my own, I might have combined a few of these calculations in one step, but I don't encourage too much of that because it's easy to miss a step. Especially when you are re-writing the equation over and over again.

Before we can move on to the final step of the hypothesis testing procedure, take a break and congratulate yourself! That was a scary formula, and you did it! I know that you're not supposed to use food as a reward, so I petted my dog as a reward. But I also had a dark chocolate truffle.

Step 4: Conclusion

Just like before, we will need critical values, which come from out of the same t-table that we used for one-sample t-tests (and found in the Common Critical Value Tables page at the back of this book). In this example, we have a one-tailed test at α = 0.05. Our degrees of freedom for our independent samples t-test is just the degrees of freedom from each group added together:

$$df = N_1 + N_2 - 2 = 35 + 29 - 2 = 62$$

OR

$$df = (N_1 - 1) + (N_2 - 1) = (35 - 1) + (29 - 1) = 34 + 28 = 62$$

From our t-table, we find that our the closest Degrees of Freedom to 62 that is listed is $df = 60$, with a critical value of$t*$ = 1.671. Note that just because we use the $df = 60$ row in the table doesn't mean that the Degrees of Freedom is 60. The table doesn't have all critical t-scores, so we're using the closes Degrees of Freedom to find a critical value that is close to what we need for our two samples.

Our calculated test statistic has a value of t = 2.48, and the critical value is $t*$ = 1.671. As 1.671 is smaller than 2.48 (the critical value is smaller than the calculated value), we reject the null hypothesis because:

$$(Critical < |Calculated|) = \textbf{Reject null} = \textbf{means are different} \ \ = p < .05$$
$$(Critical > |Calculated|) = \textbf{Retain null} = \textbf{means are similar} \ \ = p > .05$$

We found that the calculated t-score was more extreme than the critical t-score. The critical t-score was marking the 5% probability, so a calculated score more extreme than that shows that the probability of the calculated t-score being this extreme (the means being this different) is less than 5%.

To make conclusions, let's look back at the research hypothesis:

- Research Hypothesis: The average mood for the sample who watched a comedy will be higher than the average mood for the sample who watched a horror film.
- Symbols: $\overline{X}_C > \overline{X}_H$

By rejecting the null hypothesis, we're saying that the means are different. Our research hypothesis suggests that those who watch comedies will be in a better mood (have a higher score on our DV) than those who watched a horror movie. Looking back at Table $10.4.1$, we find that this is indeed the case. The participants who watched the comedy scored both mathematically and statistically significantly higher than the horror movie group. Use Example $10.4.2$ to create a statistical sentence to indicate this conclusion.

Example 10.4.2

What is the statistical sentence for these results?

Solution

$t(62) = 2.48, p < .05$

Step 5: Write-Up

You have all the pieces to practice a write-up now!

Example 10.4.3

Complete the write-up for this scenario.

Solution

Based on our sample data from people who watched different kinds of movies, the average mood for the sample who watched a comedy ($\overline{X_C} = 24.00$) was higher than the average mood for the sample who watched a horror film ($\overline{X_H} = 16.50$), $t(62) = 2.48, p < .05$. This supports our research hypothesis.

You're doing great, keep it up!

- Dr. MO (Taft College)

This page titled 10.4: Practice Independent t-test on Movies and Mood is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

10.5: Practice Independent t-test on Growth Mindset

Scenario

Dr. MO's student researchers at a small community college were interested in changing the growth mindset of students, so they asked faculty to have students complete activities that might improve growth mindset. One faculty member, an English professor, had 39 students journal about growth mindset ideas throughout the semester. This activity required students to reflect on barriers that they had overcome, their attitudes about learning, and much more. Another faculty member held discussions about growth mindset ideas throughout the semester in her statistics course; she had 22 students complete the Mindset Quiz. Higher scores on the Mindset Quiz means that the person has a more growth mindset (rather than fixed). The research question was: Which activity worked better to increase scores on the Mindset Quiz: journaling or in-class activities?

Table 10.5.1- *Descriptive Statistics of Mindset Quiz Scores by Intervention Type*

Sample	Mean	SD	N
Journal	44.77	8.01	39
Mindset Discussions	46.91	6.18	22

Okay, let's go!

Step 1: Hypotheses

As usual, we will start with the research hypothesis. You can use the means in Table 10.5.1 to make your predictions.

Exercise 10.5.2

What is a direction research hypothesis related to the research question? Provide a research hypothesis in words and in symbols.

Answer available at the end of this section.

And now, the null hypothesis!

Exercise 10.5.3

What is the null hypothesis for our variables? Provide the null hypothesis in words and in symbols.

Answer available at the end of this section.

Step 2: Collect Data

Table 10.5.1 provides data, so you are not collecting your own. Instead, we'll use this step to answer questions to make sure that we understand the scenario. These questions are asking about terms that we learned in the first chapter, if you'd like to review!

Exercise 10.5.1

1. Who is the sample? *Your answer should include the **total** number of people in 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)? Add exercises text here.

Answer available at the end of this section.

We are now prepared for the next step of the null hypothesis significance testing process. Although calculating the t-test might seem like it's the most difficult part of the process, students actually have more trouble with the last steps (Conclusion and Write-Up).

Okay, let's go!

Step 3: Compute the Test Statistic

The descriptive statistics for both groups were provided in Table $10.5.1$. Don't forget to take a deep breath before you start!

Exercise $10.5.1$

Calculate an independent t-test using the descriptive statistics from Table $10.5.1$.

After your deep breath, copy down the independent samples t-test formula:

$$t = \frac{(\overline{X}_J - \overline{X}_D)}{\sqrt{\left[\frac{(N_J - 1) * s_J^2 + (N_D - 1) * s_D^2}{N_J + N_D - 2}\right] * \left(\frac{1}{N_J} + \frac{1}{N_D}\right)}}$$

And then fill in all of the numbers from Table $10.5.1$ into the formula:

$$t = \frac{(44.77 - 46.91)}{\sqrt{\left[\frac{(39 - 1) * (8.01)^2 + (22 - 1) * (6.18)^2}{39 + 22 - 2}\right] * \left(\frac{1}{39} + \frac{1}{22}\right)}}$$

Answer available at the end of this section.

Ack, this is a negative calculated t-score! But that's okay, it's not a computational error. It just means that the mean that I started with was smaller than the second mean. What to note is that you use the absolute value when comparing to the critical t-score (so, act like it's a positive calculated t-score when determining if it's bigger or smaller than the critical t-score).

Step 4: Conclusion

To conclude whether or not to support the research hypothesis, we need to first decide if we should retain or reject the null hypothesis; we can't say that the journaling activity worked better than the discussions if we don't even know if the means are statistically significantly different.

The critical values will come from the t-table that we used for one-sample t-tests (and found in the Common Critical Value Tables at the back of this book). In this example, we have a one-tailed test at α = 0.05. To find the critical t-score, we must first find the Degrees of Freedom for our independent samples.

Exercise $10.5.1$

What are the Degrees of Freedom for this independent sample scenario?

Answer available at the end of this section.

Using that Degrees of Freedom, we now can look at the t-table that we used for one-sample t-tests (and found in the Common Critical Value Tables at the back of this book).

Exercise 10.5.1

What is the critical t-score for this scenario?

Answer available at the end of this section.

Now that we have a critical score and a calculated scorew, we can compare there to determine if the null hypothesis is retained or rejected. Our calculated test statistic has a value of t = -1.08, and we found that the critical value is $t*$ = 1.671 (if you used the closest Degrees of Freedom). 1.671 > |-1.08|, so we retain (fail to reject) the null hypothesis because:

$$(Critical < |Calculated|) = \textbf{Reject null} = \textbf{means are different } = p < .05$$
$$(Critical > |Calculated|) = \textbf{Retain null} = \textbf{means are similar } = p > .05$$

Since we retained our null hypothesis (which said that the means would be similar), we cannot support our research hypothesis that journaling would be more effective than discussions. What would our statistical sentence look like?

Exercise 10.5.3

Write out the statistical sentence for this conclusion.

Answer available at the end of this section.

There are some things to note about the statistical sentence. First, the Degrees of Freedom in the parentheses are what was calculated (df = 59), not the Degrees of Freedom that we had to use from the table. Second, the negative calculated t-score is included in the statistical sentence, even though we used the absolute value when comparing to the critical value from the table.

Step 5: Write-Up

You did all of the math and made the conclusion about the research hypothesis, so the last step is to write-it up!

Exercise 10.5.4

What could the write-up look like for this scenario?

Answer available at the end of this section.

Your writing probably won't look exactly like the Answer in Exercise 10.5.4 because your brain, voice, and writing style is different. As long as you have all of the following pieces, then you're doing it correctly!

Note

Information to include in a write-up:

- Research Hypothesis, including:
  - Re-state the RH (including the groups being compared, the DV, and the direction word), and
  - If the RH was supported or not.
- Calculation results:
  - Calculated statistical (statistical sentence): t(df) = calculated t-score, p <.05 [or p>.05]
  - Means, with units, of each group (name the groups)

You did it! One final note about this scenario: if you read the scenario closely, you might have noticed that the IV (type of mindset intervention) was confounded with the type of class that the students were in. This is the type of problem that you run into when conducting field experiments (experiments in real-life settings)! Because of this, an alternate conclusion is that there is no different in Mindset Quiz scores between the two class types (English or Statistics) $t(59) = -1.08, p > .05$. If there had been a difference, we wouldn't know if improved mindset was because of the intervention of journaling or because students in English classes have better mindsets. #ScienceIsHard

- Dr. MO (Taft College)

Answers

Exercise 10.5.2

- Research Hypothesis: The average Mindset Quiz scores for the sample who journalled will be higher than the average Mindset Quiz scores for the sample who had class discussions about growth mindset.
- Symbols: $\overline{X}_J > \overline{X}_D$

Exercise 10.5.3

- Null Hypothesis: The average Mindset Quiz scores for the sample who journalled will be similar to the average Mindset Quiz scores for the sample who had class discussions about growth mindset. In order words, there will be no difference between the means.
- Symbols: $\overline{X}_J = \overline{X}_D$

Exercise 10.5.1

1. The sample is 61 students from a small community college.
2. There isn't one right answer for this one; I might say "all community college students," but any larger group that the full sample can represent works.
3. The levels of the IV are the two groups being compared, the group of students who journalled about mindset and the group of students who had class discussions. This IV might be called Intervention Type or something similar.
4. The DV is the Mindset Quiz, measured in points.

Exercise 10.5.1

$$
\begin{aligned}
t &= \frac{(44.77 - 46.91)}{\sqrt{\left[\frac{(39-1)*(8.01)^2 + (22-1)*(6.18)^2}{39+22-2}\right] * \left(\frac{1}{39} + \frac{1}{22}\right)}} \\
t_{Add/Subtract} &= \frac{-2.14}{\sqrt{\left[\frac{(38)*(8.01)^2 + (21)*(6.18)^2}{59}\right] * \left(\frac{1}{39} + \frac{1}{22}\right)}} \\
t_{Square} &= \frac{-2.14}{\sqrt{\left[\frac{(38*64.16) + (21*38.19)}{59}\right] * \left(\frac{1}{39} + \frac{1}{22}\right)}} \\
t_{Multiply} &= \frac{-2.14}{\sqrt{\left[\frac{(2438.08) + (802.04)}{59}\right] * \left(\frac{1}{39} + \frac{1}{22}\right)}} \\
t_{AddAgain} &= \frac{-2.14}{\sqrt{\left(\frac{3240.12}{59}\right) * \left(\frac{1}{39} + \frac{1}{22}\right)}} \\
t_{Divide} &= \frac{-2.14}{\sqrt{(54.92) * (0.03 + 0.05)}} \\
t_{AddAgainAgain} &= \frac{-2.17}{\sqrt{(54.92) * (0.07)}} \\
t_{MultiplyAgain} &= \frac{-2.14}{\sqrt{3.90}} \\
t_{SquareRoot} &= \frac{-2.14}{1.98} \\
t_{TheEnd!} &= -1.08
\end{aligned}
\tag{10.5.1}
$$

Exercise 10.5.1

$$df = N_J + N_D - 2 = 39 + 22 - 2 = 59$$

OR

$$df = (N_J - 1) + (N_D - 1) = (39 - 1) + (22 - 1) = 38 + 21 = 59$$

Exercise 10.5.1

The closest Degrees of Freedom in the table is df = 60. For that, the critical t-score for α = 0.05 is 1.671.

The Degrees of Freedom in the table if you are rounding down is df = 40. For that, the critical t-score for α = 0.05 is 1.684.

Use the rule that your professor tells you to use!

Exercise 10.5.3

$t(59) = -1.08, p > .05$

Exercise 10.5.4

We hypothesized that The average Mindset Quiz scores for the sample who journalled will be higher than the average Mindset Quiz scores for the sample who had

class discussions about growth mindset; This was ~~does~~ *not* support our research hypothesis ($t(59) = -1.08, p > .05$). Based on our sample data from students who learned about growth mindset topics, the average Mindset Quiz score for the group who journalled on growth mindset topics ($\overline{X_J} = 44.77$) was s*imilar* ~~than~~ to the average Mindset Quiz score for the group who had discussions about growth mindset ($\overline{X_D} = 46.91$), $t(59)$-= $-1.08, p$-> $.05$. In the end, the two types of intervention work equally well.

This page titled 10.5: Practice Independent t-test on Growth Mindset is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

10.6: When to NOT use the Independent Samples t-test

Let's say that you have two groups, that are independent, and you measured a quantitative variable. Sounds like the perfect time to use an independent samples t-test, right?

This is not a trick question. Yes, when you have a quantitative variable for two independent groups, the independent samples t-test is the analysis to complete.

But what if it's not?

When NOT to Use the Independent Samples t-test

When There Are More Than Two Groups

If you wanted to compare more than two groups, you *could* compare pairs of each group by conducting multiple t-tests. Other than being tedious, what do you think that we don't do that? Maybe review Type I errors to help figure out why we don't do multiple t-tests?

After reminding yourself about Type I errors, you now can see that we increase our chances of rejecting a true null hypothesis (a true null hypothesis says that the means are similar in the population) when we conduct multiple tests on the same data. But don't fret, we have another kind of statistical analysis for when we have more than two groups: An ANOVA! We'll learn about those a little alter.

When the Groups are Dependent

Independent is in its name, so it might be a bit obvious, but we wouldn't conduct an independent t-test when the two samples are somehow related. This actually happens a lot, like:

- When we compare the same person Before some experience and then After that experience. This is call a pretest/posttest design.
- When we compare two people who are linked in some way. Often, this is based on a relationship (twins, a parent and child, a romantic couple), but it could also be when two participants are matched on some characteristics (say, GPA), and then the participants are randomly assigned into two conditions (one person with a high GPA is assigned to the tutoring condition and one person with a high GPA is assigned to a no-tutoring condition).

Regardless of the reason why the groups are related, if we have pairs of participants, then an Independent Samples t-test is not the appropriate test for us. But don't frest, we have another kind of statistical test for dependent samples: the Dependent Samples t-test. We'll learn about that in the next chapter!

When the Distribution Doesn't Fit the t-test Assumptions

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.

When the Two Standard Deviations are Very Different

Using the pooled variance to calculate the test statistic relies on an assumption known as homogeneity of variance. This is fancy statistical talk for the idea that the true standard deviation for each group is the similar to the other group, and that any differences in the samples' standard deviations 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.

Although it hasn't been highlighted, we've been talking about something called the Student's t-test (there's a fun story about beer and privacy about why it's called that), and that is the most common t-test and is all of the statistics textbooks. However, Welch's t-test is probably a better statistical analysis of two independent groups because it doesn't require that the two standard deviations are similar. The thing is, since no one learns about Welch's t-test, no one uses it. And since no one uses it, if you try to use it, then your supervisor or publisher or professor will be like, "Why are you using this crazy kind of t-test?" which translates to "I don't understand what you're doing, so you can't do it." If you stay in the social sciences, you are encouraged to read a little about Welch's t-test so that you can convince the supervisor/publisher/professor in your life that it's a better statistical analysis than the Student's t-tests. Standard statistical software will run it, and the interpretation is similar to the Student's t-test, so it's really about explaining how cool it is to others to be able to use it.

When the Distribution is Not Normally Distributed

Similar to the assumption of homogeneity of variance, an assumption of t-tests is that the distributions are normally distributed. With the examples that we've been using with small sample sizes, this has probably not been the case, so we are lucky that the t-test is robust.

As we will learn in the next section, when we are worried about the distribution not being normally distributed, we can use non-parametric alternatives.

This page titled 10.6: When to NOT use the Independent Samples t-test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

10.7: Non-Parametric Independent Sample t-Test

The alternatives to all statistical analyses comparing means are non-parametric analyses. A parameter is a statistic that describes the population. Non-parametric statistics 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! 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".)

All of the non-parametric statistics for use with quantitative variables (means) work with the *ranks* of the variables, rather than the values themselves.

Exercise 10.7.1

From Ch. 1, what is the scale of measurement called for ranked variables?

Answer available at the end of this section.

If your data started out as a quantative variable but you need to convert it to rank for these analyses, you give the smallest score of either of your groups the smallest rank, making each higher score a bigger rank. The tied values get the mean of the involved ranks. For example, if April's score and Primavera's score are tied for 3rd and 4th ranks, both get a rank of 3.5.

Mann-Whitney U Test

The Mann-Whitney U-test is a non-parametric alternative to an independent samples t-test that some people recommend for non-normal data. An independent samples t-test can usually handle if the standard deviations are similar or are not normally distributed, so there's little reason to use the Mann-Whitney U-test unless you have a true ranked variable instead of a quantitative variable..

Despite that fact, this is a behavioral statistics textbook, so we're going to talk about statistical alternative.

You can use the Mann-Whitney when:

- Your data is already in ranks (ordinal), *or*
- When you'd like to use an independent sample t-test, but the data is probably not normally distributed. (When the data is not normally distributed, the mean is sorta... meaningless.)

Formula

The formulas are below, but they are so uncommonly used that they won't be in the Common Formulas page at the back of the textbook. Similarly, there is a critical value table for U-scores, but that will also not be included in the Common Critical Values page.

To calculate this formula, you would need to list out all of the scores in order, and identify which is from which group. Then, you would calculate R1, which is the sum of the *ranks* of all of the scores (not the scores themselves) from the first group.

$$U_1 = (N_1 * N_2) + \left(\frac{(N_1 * (N_1 + 1))}{2}\right) - R_1$$

$$U_2 = (N_1 * N_2) - U_1$$

Mann-Whitney steps:

1. Calculate the *two* formulas,
2. Then compare the *smallest* of the two calculated U values to a critical U from a critical U table.

Interpreting Results

Imagine that I wanted to compare the mean of Exam #1 of two sections of my behavioral statistics classes, one in the morning and one in the evening.

- Research Hypothesis: The morning class's average Exam #1 score will be higher than the average Exam #1 score of the evening section.
- Symbols: $\overline{X_M} > \overline{X_E}$

Exercise 10.7.2

What's the null hypothesis in words and symbols?

Answer available at the end of this section.

I don't think that the data is normally distributed, so I would run a Mann-Whitney U. If I got results like these: (U(62) = 1.11, p>.05). After looking at the actual ranks, I would conclude that the morning class's Exam #1 ranks were higher than the ranks of the evening class on Exam #1.

And that's it!

Contributors and Attributions

- John H. McDonald (University of Delaware)
- Dr. MO (Taft College)

Answers

Exercise 10.7.1

Ordinal scale of measurement

Exercise 10.7.2

- Null Hypothesis: The morning class's average Exam #1 score will be similar ton the average Exam #1 score of the evening section.
- Symbols: $\overline{X_M} = \overline{X_E}$

This page titled 10.7: Non-Parametric Independent Sample t-Test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

11: DEPENDENT SAMPLES T-TEST

11.1: Introduction to Dependent Samples
11.2: Dependent Sample t-test Calculations
11.3: Practice Dependent t-test on Job Satisfaction
11.4: Practice Dependent t-test on Changes in Mindset
11.5: Non-Parametric Analysis of Dependent Samples
11.6: Choosing Which Statistic- t-test Edition

This page titled 11: Dependent Samples t-test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

11.1: Introduction to Dependent Samples

Another Kind of t-test

Remember way back when we learned about one-sample t-tests? We compared the mean of a sample to a population mean.

And then last chapter, we learned about mean differences between two different groups. Independent t-tests compare TWO unrelated groups from ONE time point.

Now, we will find the average value of difference scores, that is: difference scores came from ONE group at TWO time points (or two perspectives).

It is important to 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 dependent samples t-test because the repeated measures become pairs of data that are related. 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? Twins? If so, we call those data matched and also use a 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 last chapter. And if we only have one group and population mean, then we use the one-sample t-test.

Dependent t-test

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 11.1.1*: Raw and difference scores before and after training.*

Employee	Before	After	Difference
Alberto	8	9	1
Bekah	7	7	0
Carlos	4	1	-3
Diamond	6	8	2
Elphaba	2	8	6

Table 11.1.1 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 (Difference) is what we look at when assessing whether or not our training was effective. We want to see positive scores, which indicate improvement (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 (Employee E with a score of 2) improved so much that they ended up (After) as skilled as the highest scoring employee from Before (Employee A (with a score of 8), and ended up only 1 point lower at the end. There's also one Difference score of 0 (Employee B), meaning that the training did not help this employee, and one negative Difference score, meaning that Employee C actually performed worse after the training! 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 11.1.1, we calculate our difference scores by taking the time 2 score and subtracting the time 1 score. That is:

$$X_D = \mathrm{X_{T2}} - \mathrm{X_{T1}}$$

Where X_D is the difference score, $\mathrm{X_{T1}}$ is the score on the variable at Time 1 (Before), and $\mathrm{X_{T2}}$ is the score on the variable at Time 2 (After). Notice that we start with Time 2 so that positive scores show improvement, and negative scores show that skills decreased. Often, Time 1 is called a Pretest and Time 2 is called a Post-Test, making this a Pretest/Post-Test design. The difference score, X_D, will be the data we use to test for improvement.

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 who weren't related at the individual level, such as samples of people from different states that we gathered independently. Such datasets and analyses should be analyzed with an independent t-test, not the dependent t-test that we are discussing in this 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.

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 or two independent samples, but now with a different formula that focuses on the difference between each pair.

Interpreting Dependent t-tests

Null hypothesis significance testing and p-values are the same with dependent t-tests as the other types of t-tests:

Table 11.1.2- Small p-values Versus Large p-values

Small p-values (p<.05)	Large p-values (p>.05)
A small p-value means a small probability that the two means are similar (suggesting that the means are different...).	A large p-value means a large probability that the two means are similar.
We conclude that: • The means are different. • The samples are *not* from the same population	We conclude that • The means are similar. • The samples *are* from the same population.

Small p-values (p<.05)	Large p-values (p>.05)
The calculated t-score is further from zero (more extreme) than the **critical t-score**. (Draw the standard normal curve and mark the calculated t-score and the **critical t-score** to help visualize this.)	The calculated t-score is closer to zero (less extreme) than the **critical t-score**. (Draw the standard normal curve and mark the calculated t-score and the **critical t-score** to help visualize this.)
Reject the null hypothesis (which says that the means are similar).	Retain (or fail to reject) the null hypothesis (which says that the means are similar).
Support the Research Hypothesis? MAYBE. Look at the actual means: · *Support the Research Hypothesis if the mean that was hypothesized to be bigger really is bigger.* · *Do not support the Research Hypothesis if the mean that was hypothesized to be bigger is actually smaller.*	Do not support the Research Hypothesis (which said that one mean would be bigger, but the means are similar).
Write "*The mean from Sample 1 (= ##) differs from the mean of Sample 2 ((, suggesting that the samples are from different populations (t(df) = _____, p<.05). This supports (OR DOES NOT SUPPORT) the Research Hypothesis.*"	Write "*The mean from Sample 1 (= ##) is similar to the mean from Sample 2 (suggesting that the samples are from the same population (t(df) = _____, p>.05). This does not support the Research Hypothesis.*"

Let's figure out what we do with dependent pairs of scores.

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 11.1: Introduction to Dependent Samples is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

11.2: Dependent Sample t-test Calculations

Hypotheses

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. As with our other hypotheses, we express the 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 mean at Time 1 will be lower than the mean at Time 2 after training.").

Research Hypothesis

Our research hypotheses will follow the same format that they did before:

- Research Hypothesis: The average score increases, such that Time 1 will have a lower mean than Time 2.
  - Symbols: $\overline{X}_{T1} < \overline{X}_{T2}$

or

- Research Hypothesis: The average score decreases, such that Time 1 will have a higher mean than Time 2.
  - Symbols: $\overline{X}_{T1} > \overline{X}_{T2}$

When might you want scores to decrease? There are plenty of examples! Off the top of my head, I can imagine that a weight loss program would want lower scores after the program than before. Or a therapist might want their clients to score lower on a measure of depression (being less depressed) after the treatment. Or a police chief might want fewer citizen complaints after initiating a community advisory board than before the board. For mindset, we would want scores to be higher after the treament (more growth, less fixed).

> **Note**
>
> What Before/After test (pretest/post-test) can you think of for your future career? Would you expect scores to be higher or lower after the intervention?

As before, you choice of which research 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 change. However, since we are just beginning to learn all of this stuff, Dr. MO might let you peak at the group means before you're asked for a research hypothesis.

Null Hypothesis

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:

- Null Hypothesis: The means of Time 1 and Time 2 will be similar; there is no change or difference.
- Symbols: $\overline{X}_{T1} = \overline{X}_{T2}$

The mathematical version of the null hypothesis is always exactly the same when comparing two means: the average score of one group is equal to the average score of another group.

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 (still usually one-tailed), and the degrees of freedom. With degrees of freedom, we go back to $df = N–1$, but the "N" is the *number of pairs*. If you are doing a Before/After (pretest/post-test) design, the number of people will be the number of pairs. However, if you have matched pairs (say, 30 pairs of romantic partners), then N is the number of pairs (N = 30), even though the study has 60 people. 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. After we calculate our test statistic, our decision criteria are the same as well:

$$(Critical < |Calculated|) = \textbf{Reject null} = \textbf{means are different} \ = p < .05$$

$$(Critical > |Calculated|) = \textbf{Retain null} = \textbf{means are similar} \ = p > .05$$

Test Statistic

Our test statistic for our change scores follows similar format as our prior t-tests; we subtract one mean from the other, and divide by a standard deviation. Sure, the formulas changes, but the idea stays the same.

$$\frac{\left(\overline{X}_D \times \sqrt{N}\right)}{s_D}$$

This formula is mostly symbols of other formulas, so it's only useful when you are provided mean of the difference ($\overline{X}_D$) and the standard deviation of the difference (s_D). Below, we'll go through how to get the numerator and the denominator, then combine them into the full formula

Numerator (Mean Differences and Stuff)

Let's start with the numerator (top) which deals with the mean differences (subtracting one mean from another).

It turns out, you already found the mean differences! That's the Differences column in the table. But what we need is an average of the differences between the mean, so that looks like:

$$\overline{X}_D = \frac{\Sigma X_D}{N}$$

The mean of the difference is calculated in the same way as any other mean: sum each of the individual difference scores and divide by the sample size. But remember, *the sample size is the number of pairs*! The X_D is the difference score for each pair.

Then, for mathematical reasons related to standardizing, you'll multiply that mean of the differences by the square root of N:

$$\overline{X}_D \times \sqrt{N}$$

Denominator (Standard Deviation of the Difference)

The denominator is just the standard deviation of the differences.

The standard deviation of the difference is the same formula as the standard deviation for a sample, but using difference scores for each participant, instead of their raw scores.

$$s_D = \sqrt{\frac{\sum\left((X_D - \overline{X}_D)^2\right)}{N-1}} = \sqrt{\frac{SS}{df}}$$

And just like in the standard deviation of a sample, the Sum of Squares (the numerator in the equation directly above) is most easily completed in the table of scores (and differences), using the same table format that we learned in chapter 3.

Full Formula for Dependent t-test

Finally, putting that all together, we can the full formula!

$$\frac{\left(\frac{\Sigma X_D}{N}\right) \times \left(\sqrt{N}\right)}{\sqrt{\left(\frac{\sum\left((X_D - \overline{X}_D)^2\right)}{(N-1)}\right)}}$$

I know that looks like a lot. And there are lots of parentheses to try to make clear the order of operations. But really, this is only finding a finding a mean of the difference, multiplying it one the square root of N, then dividing that by the standard deviation of the difference. Basically,

1. Calculate the mean of the difference ($\overline{X}_D$)), and
2. Multiply the mean of the difference by the square root of the number of pairs, then
3. Calculate the standard deviation of the difference (s_D), and finally
4. Divide!

Easy-peasy!

Don't worry, we'll walk through a couple of examples so that you can see what this looks like next!

This page titled 11.2: Dependent Sample t-test Calculations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

11.3: Practice Dependent t-test on Job Satisfaction

You've learned a complicated formula, and what it means. Now, let's try using it to answer a research question! This first scenario will provide the mean of the difference ($\overline{X}_D$) and the standard deviation of the difference (s_D), so you can focus on how the parts work together, not on all of the calculations. The next section will let you use the full formula.

Scenario

In order to be more competitive and hire the best talent, a CEO would like her employees to love working for the company. The CEO hires a consultant to first assesses a sample of the employee's level of job satisfaction (N = 40) with a job satisfaction survey that he developed to identify specific changes that might help. The company institutes some of the changes that the consultant suggests, and six months later the consultant returns to measure job satisfaction with the same survey on the same 40 employees again. You are hired by the consultant to crunch the numbers using an $\alpha = 0.05$ level of significance.

Step 1: Hypotheses

First, we state our hypotheses. Let's hope that the changes that the consultant has the knowledge and expertise to suggest useful changes that we can predict an improvement in job satisfaction.

- Research Hypothesis: The average score on the job satisfaction survey in Time 2 will be higher than average score on the job satisfaction survey in Time 1.

Note that the name of the two groups, what we were measuring, and that we are comparing means was all included. The format seems a little backwards to me because it's stated in chronological order, but it highlights that there will be improvement. If you started with Time 1, then it reads like the DV is decreasing. But either is correct.

Exercise 11.3.1

What would the research hypothesis look like in symbols?

Answer available at the end of this section.

Let's move on to the null hypothesis.

Exercise 11.3.2

What is the null hypothesis in words and then in symbols?

Answer available at the end of this section.

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 hypothesis.

Step 2: Collect Data

This scenario was first discussed in a previous section, and that section provided the data. For us, let's identify the variables and sample so that we know what and who we're talking about.

Example 11.3.1

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 the group of 40 employees who completed the job satisfaction survey twice.
2. The population might be all employees in that company? Other answers may also be correct.
3. The IV is before and after the survey, the time periods. This can also be said as the pretest and post-test.
4. The DV is the job satisfaction survey.

Step 3: Analyze Data

It's time to calculate the statistic! This first example will be simplified a little by providing the mean of the difference and the standard deviation of the difference.

Example 11.3.2

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). What is the calculated t-score?

Solution

You can start with any parentheses (Please Excuse My Dear Aunt Sally); I like to start at the top and work my way down and to the left. For this formula, starting at the top means to start with the mean of the difference in the numerator!

$$\frac{\left(\frac{\Sigma X_D}{N}\right) \times \left(\sqrt{N}\right)}{\sqrt{\left(\frac{\sum\left((X_D - \overline{X}_D)^2\right)}{(N-1)}\right)}}$$

The mean of the difference was already calculated: $\overline{X}_D = \frac{\Sigma X_D}{N} = 2.96$

We then multiply this by the square root of N to get the full numerator:

$$\left(\overline{X}_D \times \sqrt{N}\right) = \frac{2.96}{\sqrt{40}} = \frac{2.96}{6.32} = 18.72$$

Next is the denominator, which is the standard deviation of the difference:

$$s_D = \sqrt{\frac{\sum\left((X_D - \overline{X}_D)^2\right)}{N-1}} = 2.85 \tag{11.3.1}$$

Now, we can put all of those values into the formula:

$$\frac{18.72}{2.85} = 6.57$$

Step 4: Conclusion

We first need to find the critical value for this combined sample to determine if the null hypothesis is retained or rejected.

For our dependent-samples t-test, the degrees of freedom are still given as $df = N–1$ in which N is the number of pairs. For this problem, we have 40 people, so our degrees of freedom are 39. Using the same t-table from when we learned about one-sample t-tests (or going to the Common Critical Values Table page), we find that the critical value is 1.684 if we use the closest df from the table (df=40), or 1.697 if you round the df down (df=30); ask your professor which rule you should follow (closest or rounding down).

The critical t-score from the table (based on either decision rule) from Step 2 is smaller than the calculated test statistic of $t = 6.57$.

$$(Critical < |Calculated|) = \textbf{Reject null} = \textbf{means are different} \ = p < .05$$

$$(Critical > |Calculated|) = \textbf{Retain null} = \textbf{means are similar} \ = p > .05$$

Exercise 11.3.3

Based on the critical value from the table and the calculated t-score, should we reject the null hypothesis? Does this mean that the means are similar or different?

Answer available at the end of this section.

We're getting somewhere! Now that we know that the means are statistically different (because we rejected the null hypothesis that said that they were similar), we should compare the actual means to the pattern in our research hypothesis. However, these means were never provided! This means that we can't conclude whether or not the research hypothesis was supported; all we can conclude is that the means are different.

Step 5: Write-Up

For a complete write-up, you should include the means of each group. However, these were not provided in our scenario. That'll make this easier since there'll be less information that we need to include.

Example 11.3.3

Write up a conclusion for this job satisfaction scenario.

Solution

Based on the sample data from 40 workers, we can say that the employees' average job satisfaction was statistically significantly different after the interventions compared to before the interventions ($t(39) = 6.57$, $p < 0.05$.

Hopefully the above example made it clear that running a dependent samples t-test to look for differences before and after some treatment works similarly to the other t-tests that we've covered. 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.

Let's try another example with real data on mindset scores!

Answers

Exercise 11.3.1

$\overline{X}_{T2} > \overline{X}_{T1}$

If you but Time 1 first in your sentence, then it would be: $\overline{X}_{T1} < \overline{X}_{T2}$

Exercise 11.3.2

- Null Hypothesis: The average score on the job satisfaction survey in Time 2 will be similar to the average score on the job satisfaction survey in Time 1. In other words, average job satisfaction scores will not differ between the two time periods.
- Symbols: $\overline{X}_{T2} = \overline{X}_{T1}$

Exercise 11.3.3

$Critical(1.6xxx) < |6.57| =$ **Reject null** $=$ **means are different** $= p < .05$

We reject the null hypothesis, and state that the means are statistically different from each other.

This page titled 11.3: Practice Dependent t-test on Job Satisfaction is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

11.4: Practice Dependent t-test on Changes in Mindset

Let's try this full process, starting with a table of raw scores, and moving through the steps and the formula to analyze paired Mindset Quiz scores. Check out the section on growth mindset to refresh yourself on what that idea means.

Scenario

Dr. MO's student research group conducted a study in which one English faculty member tried to use journaling to improve growth mindset in a class that was one level below college-level English (so, remedial or basic skills English) at a community college. The Mindset Quiz was completed by students at the beginning of the semester, and then again at the end of the semester. We had complete data (both pretest and posttest scores) for 10 students (N=10).

Step 1: Hypotheses

Without knowing the means yet, we should just hope that the journalling intervention improved mindset. With this assumption, develop the hypotheses.

Exercise 11.4.2

Develop the research hypothesis, in words and symbols. Then, determine the null hypothesis in words and symbols.

Answer available at the end of this section.

In this case, we are hoping that the changes we made will improve mindset scores, so we will use a one-directional research hypothesis.

Step 2: Collect Data

You may have noticed that we have yet to actually collect data for this step of the Null Hypothesis Significance Testing process, and that hasn't changed with this scenario! However, you may collect data in your class (or your future career) in this step. For now, we'll again identify the sample, population, and variables.

Exercise 11.4.1

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 available at the end of this section.

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, and our degrees of freedom.

Exercise 11.4.3

For our dependent-samples t-test, what is the critical t-score from the the same t-table from when we learned about one-sample t-tests (or going to the Common Critical Values Table page)?

Answer available at the end of this section.

Step 3: Analyze Data

You can use the information in Table 11.4.1 to calculate the mean and standard deviation for the pretest and posttest. To complete the dependent t-test equation, additional columns will be needed to find and sum the difference, and to square and sum the square of each difference.

Table 11.4.1- Pretest and Posttest Mindset Quiz Scores

Participant	Mindset Quiz Pretest	Mindset Quiz Posttest
A	38	33
B	43	39
C	35	32
D	36	37
E	47	51
F	39	45
G	49	55
H	47	54
I	31	39
J	42	52
Σ	Σ **= 407**	Σ **= 437**

Although not required right now, you might be interested in the actual means for the Mindset Quiz at the pretest and posttest, and now you can calculate them!

For the pretest:

$$\frac{\Sigma X}{N} = \frac{407}{10} = 40.70$$

This score means that the students had many growth mindset ideas, but still held some fixed beliefs.

For the posted:

$$\frac{\Sigma X}{N} = \frac{437}{10} = 43.70$$

This score also shows that the students had many growth mindset ideas, but still held some fixed beliefs. However, scores of 45 or higher are considered having only growth mindset beliefs, so the posttest group was close!

Exercise 11.4.4

Use the dependent t-test formula to calculate a dependent t-test.

Answer available at the end of this section.

If your answer is close to 1.76 but not exactly (say, 1.73ish to 1.79ish), you probably did the equation correctly! The difference may be based on rounding; Dr. MO uses spreadsheets to do her calculations, rather than a calculator and hand-writing each difference score; spreadsheets keep hundreds of numbers after the decimal point, rather than just two, which affects the two numbers after the decimal point at the end of the calculation.

Step 4: Conclusion

The Conclusion step is about deciding whether or not to support the research hypothesis. To get to that, we first need to decide if the null hypothesis should be retained or rejected. And to get to that, we need to find the critical value so that we can compare the critical value to the t-score that you just calculated. 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, and our degrees of freedom.

Exercise 11.4.3

For our dependent-samples t-test, what is the critical t-score from the same t-table from when we learned about one-sample t-tests (or going to the Common Critical Values Table page)?

Answer available at the end of this section.

The critical t-score from the table (based on either decision rule) from Step 2 is larger than the calculated test statistic of $t = 1.76$.

$$(Critical < |Calculated|) = \textbf{Reject null} = \textbf{means are different} \ = p < .05$$

$$(Critical > |Calculated|) = \textbf{Retain null} = \textbf{means are similar} \ = p > .05$$

Exercise 11.4.5

1. Based on the critical value from the table and the calculated t-score, should we reject the null hypothesis?
2. What is the statistical sentence for this scenario's results?
3. Does this mean that the means are similar or different?

Answer available at the end of this section.

We retained the null hypothesis; now what?

Exercise 11.4.6

Based on the means, our research hypothesis, and the statistical sentence is the research hypothesis supported or not supported? Why or why not?

Answer available at the end of this section.

Step 5: Write-Up

What should the write-up look like? This write-up will sound sorta weird because the Posttest mean is higher than the Pretest mean, but the significance test shows that the means are similar.

Example 11.4.1

Write up a conclusion for the mindset data.

Solution

I hypothesized that the average Mindset Quiz Posttest ($\overline{X} = 43.70$) would be higher than the average Mindset Quiz Pretest ($\overline{X} = 40.70$) after the journaling intervention. This research hypothesis is not supported; the means are not different enough (too similar) to state that they are from different populations (t(9)=1.76,p>.05). It appears that the journaling intervention did not statistically improve mindset scores for this sample of 10 community college students in an English course below college-level.

How are you feeling?

If you didn't quite get it yet, keep trying! Maybe check out the Learning (Statistics) section the very first chapter?

And when you do get the right calculation, interpret it correctly, and have a complete write-up, **celebrate**!

Answers

Exercise 11.4.2

- Research Hypothesis: The average score will be higher on the Mindset Quiz at the end of the semester (Time 2, posttest) than the average score at the beginning of the semester (Time 1, pretest).
  - Symbols: $\overline{X}_{T2} > \overline{X}_{T1}$
- Null Hypothesis: The average score will be similar to the Mindset Quiz at the end of the semester (Time 2, posttest) than the average score at the beginning of the semester (Time 1, pretest).
  - Symbols: $\overline{X}_{T2} = \overline{X}_{T1}$

1. The sample is the group of 10 students who completed the Mindset Quiz at the beginning and the end of the semester (twice).
2. The population might be all community college students who haven't passed college-level English, although the specific population that you think this sample represents might be slightly different.
3. The IV is before and after the survey, the time periods. This can also be said as the pretest and post-test.
4. The DV is the Mindset Quiz.

Exercise 11.4.3

For our dependent-samples t-test, the degrees of freedom are still given as $df = N–1$ in which N is the number of pairs.

$$\text{Degrees of Freedom } = N - 1 = 10 - 1 = 9$$

With a Degrees of Freedom of 9, the critical t-score is 1.833 from the critical t-score table.

Exercise 11.4.4

To use the dependent t-test formula, you will need to calculate the Difference scores between each participant's pretest and posttest, find how different those Difference scores are from the mean of the difference, and square how different the Difference scores are from the mean of the difference. Table 11.4.2 is provided to help complete this step for each participant. The sum for each column is provided so that you can check your subtraction, squaring, and adding, but try to do it yourself so that you know how!

Table 11.4.2- *Mindset Quiz Pretest and Posttest Scores with Columns for Calculating*

Participant	Pretest	Posttest	Difference (Posttest minus Pretest)	Difference minus Mean of Difference	Square of the Difference minus Mean of Difference
A	38	33			
B	43	39			
C	35	32			
D	36	37			
E	47	51			
F	39	45			
G	49	55			

Participant	Pretest	Posttest	Difference (Posttest minus Pretest)	Difference minus Mean of Difference	Square of the Difference minus Mean of Difference
H	47	54			
I	31	39			
J	42	52			
Σ	Σ = **407**	Σ = **437**	Σ = **30**	Σ = **0.00**	Σ = **262.00**

Based on this table, we have what we need to complete the dependent t-test formula:

$$t = \frac{\left(\frac{\Sigma X_D}{N}\right) \times \left(\sqrt{N}\right)}{\sqrt{\left(\frac{\sum\left((X_D - \overline{X}_D)^2\right)}{(N-1)}\right)}}$$

So we fill in all of the numbers. Notice that the all of the sums were computed on the Table 11.4.2. Dr. MO even did the sums for you, although you should practice filling in the table and doing all of the adding so that you can do it on your own when needed. Here, we are only plugging in the answers from the table into the equation.

$$t_{numbers} = \frac{\left(\frac{30}{10}\right)\left(\sqrt{10}\right)}{\sqrt{\left(\frac{262}{(10-1)}\right)}}$$

Then we start calculating each set of parentheses:

$$t_{Parentheses} = \frac{(3 \times 3.16)}{\sqrt{\left(\frac{262}{9}\right)}}$$

And a few more parentheses:

$$t_{ParenthesesAgain} = \frac{9.49}{\left(\sqrt{29.11}\right)}$$

Then a square root:

$$t_{SquareRoot} = \frac{9.49}{5.40}$$

And then some division to finish it off!

$$t_{TheEnd!} = 1.76$$

Exercise 11.4.3

For our dependent-samples t-test, the degrees of freedom are still given as $df = N–1$ in which N is the number of pairs.

$$\text{Degrees of Freedom } = N - 1 = 10 - 1 = 9$$

With a Degrees of Freedom of 9, the critical t-score is 1.833 from the critical t-score table.

Exercise 11.4.5

$Critical(1.833) > |Calculated(1.76)|$ = **Retain null** = **means are similar** $= p > .05$

1. We retain (fail to reject) the null hypothesis.
2. t(9)=1.76,p>.05
3. The means are statistically similar to each other.

Exercise 11.4.6

The average Mindset Quiz score on the pretest was 40.70 points and the posttest was 43.70 points. Our research hypothesis was that the average score will be higher on the Mindset Quiz at the end of the semester (Time 2, posttest) than the average score at the beginning of the semester (Time 1, pretest). This is not supported because the statistical sentence of t(9)=1.76,p>.05 tells us that the means are similar (the probability that the null hypothesis is true in the population based on these samples is higher than our cut-off of 5% chance).

This page titled 11.4: Practice Dependent t-test on Changes in Mindset is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

11.5: Non-Parametric Analysis of Dependent Samples

Do you remember when to use non-parametric analyses?

Exercise 11.5.1

Do you remember when to use non-parametric analyses?

Answer available at the end of this section.

So what is the non-parametric alternative analysis for dependent t-tests? The Wilcoxon Matched-Pairs Signed-Rank test!

When to Use the Wilcoxon?

Use the Wilcoxon signed-rank test when there are two paired quantitative variables that are not normally distributed, or two paired variables that are ranks. This is the non-parametric analogue to the paired *t*–test, and you should use it if the distribution of differences between pairs is severely non-normally distributed.

Hypothesis

Because we are dealing with ranks, the hypotheses will discuss medians instead of means.

Research Hypothesis: The median of the pretest will be lower than the median of the posttest.

Null Hypothesis: The median of the pretest will be similar to the median of the posttest.

Note that this is different from the research hypothesis of the dependent t–test, which is that the *means* will be different.

How To

There's no formula for this statistical analysis! Instead, you find the difference of each pairs' scores, rank them, and compare to a critical value. Here are the steps:

1. Find the Differences between each pair of scores.
2. Rank the absolute value of the Differences for each participant from smallest to largest, with the smallest difference getting a rank of 1, then next larger difference getting a rank of 2, etc. Give average ranks to ties.
3. Add the ranks of all of the positive Difference scores, then add the ranks of all of the negative Difference scores.
4. The smaller of these two sums is the test statistic, W (sometimes symbolized T_s).
5. Compare the test statistic to a value in a critical value table. Unlike most test statistics, *smaller* values of W are less likely under the null hypothesis. However, it is unlikely that you will need to calculate this by hand, so you can rely on statistical software to provide the actual probability of a Type I error of rejecting a null hypothesis when it's true (p-value).

In Closing

Non-parametric analyses are fairly rare because most of the parametric analyses are robust to non-normal distributions so we just continue to use them. That's why we won't spend too much time on the actual calculations now; if you ever do need to conduct a non-parametric analysis, you will probably already be using statistical software to do it. For now, the important thing to remember is that there are alternative statistical analyses if your data does not appear to be normally distributed.

Contributor

- John H. McDonald (University of Delaware)
- Dr. MO (Taft College)

Answers

Exercise 11.5.1

Use non-parametric analyses when you expect that your distribution or the population is not normally distributed, or if your data starts out as ranked information.

This page titled 11.5: Non-Parametric Analysis of Dependent Samples is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

11.6: Choosing Which Statistic- t-test Edition

Time to stop and reflect. You've now learned about z-scores, three different kinds of t-tests, non-parametric alternatives for some of the t-tests, and confidence intervals. Good job!

Exercise 11.6.1

What are the three t-tests that you've learned?

Answer available at the end of this section.

You might be wondering how to decide when to use which analysis. It all depends on how many groups you have, what kind of variable your DV is (quantitative or qualitative), and whether the groups are related or not. Below is a list of all of the analyses that will be covered in this textbook; what we have covered already is in bold.

- **Quantitative DV:**
  - **One group: One-sample t-test (comparing your one sample to the population)**
  - **One IV with 2 groups: independent or dependent t-tests**
    - **Related groups (pairs): dependent samples t-test**
    - **Unrelated groups: independent samples t-test**
  - One IV with 3+ groups: ANOVA (next!)
    - Related groups: Repeated Measures ANOVA
    - Unrelated groups: Between Groups ANOVA
  - Two IVs with 2+ groups: Factorial ANOVA (later)
  - Two IVs with a range of values (no groups): Correlation (later)
- Qualitative DV:
  - One IV: Goodness of Fit Chi-Square (later)
  - Two IVs: Test of Independence Chi-Square (later)
- Combination of IVs & DVs: Regression
- **Ranked DV:**
  - **One IV with 2 groups:**
  - **Related groups: Wilcoxon Matched-Pairs Signed-Rank Test**
  - **Unrelated groups: Mann-Whitney U**
  - One IV with 3+ groups: Kruskal-Wallis One-Way ANOVA (later)

Keep this list or the attached decision tree so that you can decide which analysis to conduct whenever you get new data sets.

Answers

Exercise 11.6.1

- One-sample t-test
- Independent sample t-test
- Dependent sample t-test

This page titled 11.6: Choosing Which Statistic- t-test Edition is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

12: BG ANOVA

12.1: Why ANOVA?

12.1.1: Observing and Interpreting Variability

12.1.2: Ratio of Variability

12.2: Introduction to ANOVA's Sum of Squares

12.3: Summary of ANOVA Summary Table

12.4: Hypotheses in ANOVA

12.5: Practice BG ANOVA on Job Applicants

12.6: Introduction to Pairwise Comparisons

12.7: Pairwise Comparison Post Hoc Tests

12.8: Practice BG ANOVA and Pairwise Comparisons on Mindset Data

12.9: ANOVA versus t-test

12.10: Non-Parametric Analysis Between Multiple Groups

This page titled 12: BG ANOVA is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

12.1: 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 12.1: Why ANOVA? is shared under a CC BY-NC-SA 4.0 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.

12.1.1: 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 12.1.1.1, 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).

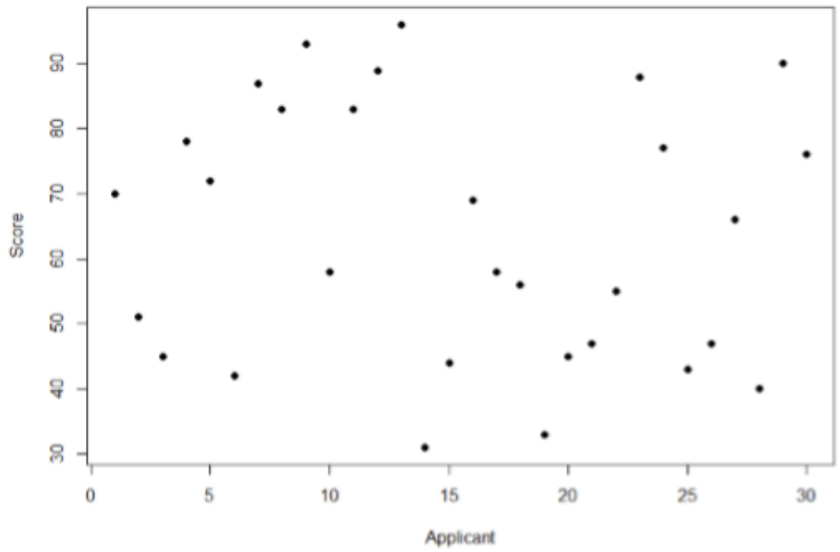


Figure 12.1.1.1: Scores on a job test (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 12.1.1.2 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 12.1.1.2.

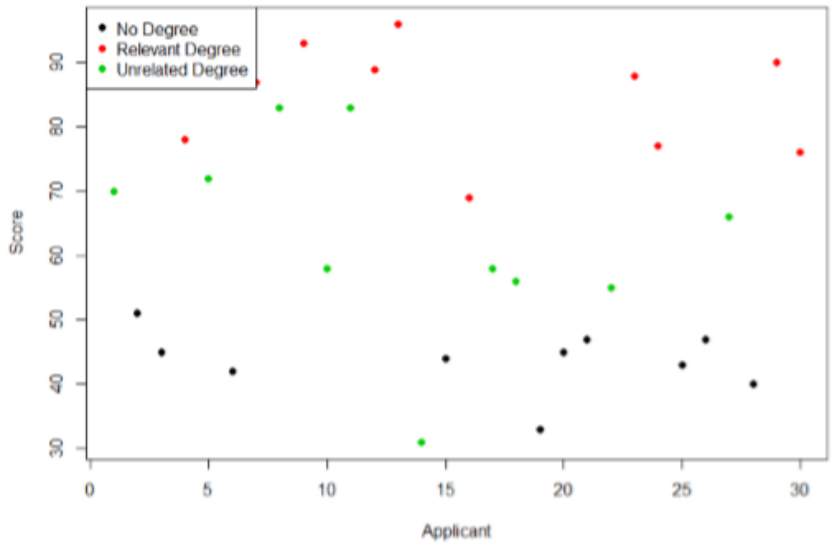


Figure 12.1.1.2: Applicant scores coded by degree earned (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

This pattern is even easier to see when the applicants are sorted and organized into their respective groups, as shown in Figure 12.1.1.3.

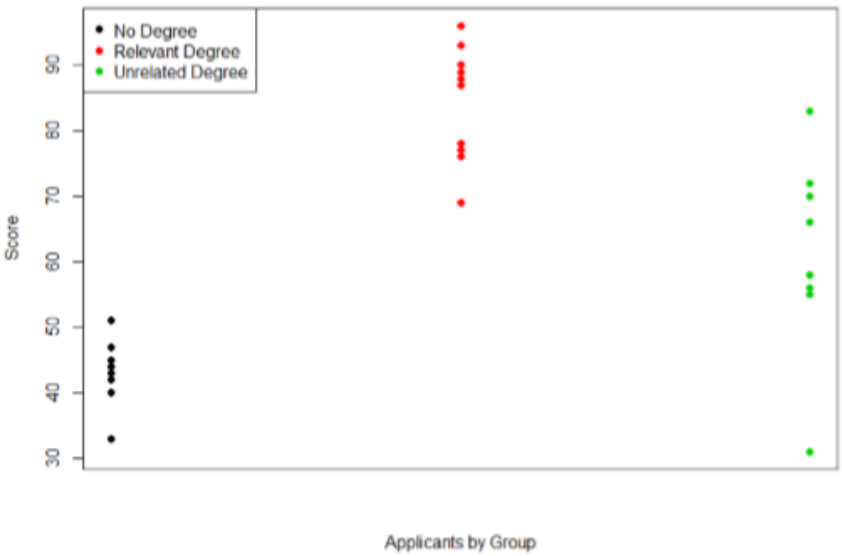


Figure 12.1.1.3: Applicant scores by group (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

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 12.1.1: 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.

12.1.2: 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 12.1.2.1 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 12.1.2.1- *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 12.1.2.1, 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 $df_{\text{BG}} + df_{\text{WG or Error}}$ 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:

$$\text{F} = \frac{\text{measure of effect}}{\text{measure of error}}$$

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 $SS_{\text{WG or Error}}$ both represent the variation due to the effect, and the leftover variation that is unexplained. Why don't we just do this?

$$\frac{SS_{\text{BG}}}{SS_{\text{WG or Error}}}$$

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}.

$$MS_{\text{BG}} = \frac{SS_{\text{BG}}}{df_{\text{BG}}}$$

$$MS_{\text{BG}} = \frac{72}{2} = 36$$

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 $SS_{\text{WG or Error}}$ represents the sum of variation for nine scores in our study. That's a lot more scores, so the $SS_{\text{WG or Error}}$ 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 $SS_{\text{WG or Error}}$ is so big. What we need to do is bring it down to the average size. So, we might want to divide our $SS_{\text{WG or Error}}$ 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 $MS_{\text{WG or Error}}$.

$$MS_{\text{WG or Error}} = \frac{SS_{\text{WG Error}}}{df_{\text{WG or Error}}}$$

$$MS_{\text{WG or Error}} = \frac{230}{6} = 38.33$$

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!

$$\text{F} = \frac{\text{measure of effect}}{\text{measure of error}}$$

$$F = \frac{MS_{BG}}{MS_{WG \text{ or } Error}}$$

$$F = \frac{36}{38.33} = .94$$

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 12.1.2: Ratio of Variability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

12.2: 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:

$$SS_B = \sum_{EachGroup} \left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 * (n_{group}) \right]$$

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 12.2: 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.

12.3: 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 12.3.1), which shows the formulas for everything other than the sum of squares, to calculate our test statistic.

Table 12.3.1: 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

The first column of the ANOVA table, labeled “Source”, indicates which of our sources of variability we are using: between groups, within groups, or total. The second column, labeled “SS”, contains our values for the sums of squares that we learned to calculate above. As noted previously, calculating these by hand takes too long, and so the formulas are not presented in Table 12.3.1. However, remember that the Total is the sum of the other two, in case you are only given two SS values and need to calculate the third.

The next column, labeled “df”, is our degrees of freedom. As with the sums of squares, there is a different df for each group, and the formulas are presented in the table. Notice that the total degrees of freedom, N–1, is the same as it was for our regular variance. This matches the SS_T formulation to again indicate that we are simply taking our familiar variance term and breaking it up into difference sources. Also remember that the capital N in the df calculations 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. Because there's so much going on in ANOVAs (in the Between Groups ANOVA that we're learning now but also in future types of ANOVAs that you'll learn), this final F statistic is called the omnibus F. Dr. MO tells her students that "omnibus" means something like "big or biggest," but it really identifies that

the F statistic includes many things. The omnibus F, 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 with t-tests. 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 $12.3.1$ 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 12.3: Summary of ANOVA Summary Table is shared under a CC BY-NC-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.

12.4: 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 12.4.1- *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 12.4: Hypotheses in ANOVA is shared under a CC BY-NC-SA 4.0 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.

12.5: Practice BG ANOVA on Job Applicants

Let's practice our five-step process using the ANOVA Summary Table to complete the calculations, but the Sum of Squares will be provided in this 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.

Step 1: 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 12.5.1

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:

- $\overline{X}_N$ ____ $\overline{X}_R$
- $\overline{X}_N$ ____ $\overline{X}_U$
- $\overline{X}_R$ ____ $\overline{X}_U$

Answer available at the end of this section.

What about the null hypothesis?

Exercise 12.5.2

State the null hypothesis in words and symbols.

Answer available at the end of this section.

Step 2: Collect Data

As we didn't design the study and collect the data, our job is to identify who the data was collected from, who they might represent, and what the IV groups and DV were.

Exercise 12.5.3

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 available at the end of this section.

Step 3: Analyze Data

Now that we have our hypotheses, 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 12.5.1.

Table 12.5.1: 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 12.5.2 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.

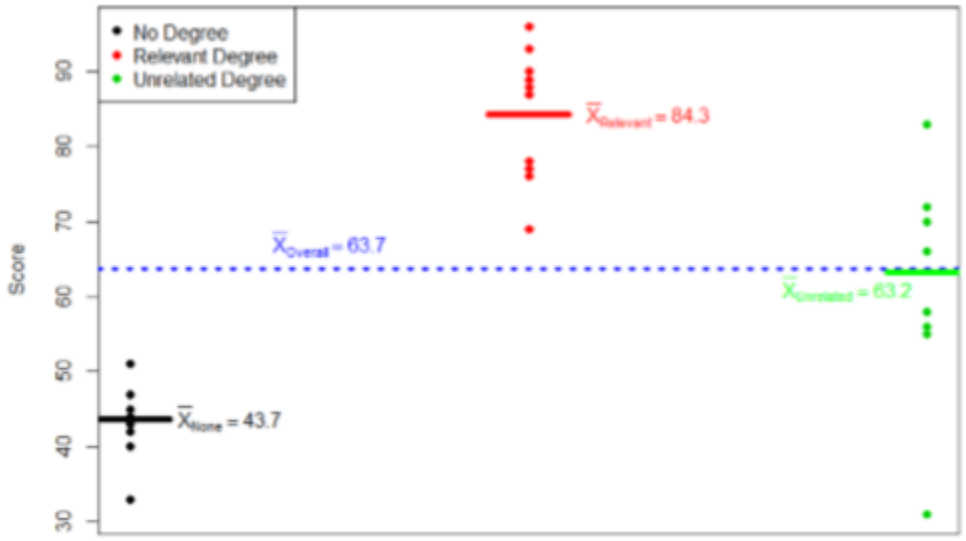


Figure 12.5.2: Means (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Example 12.5.1

Using the information provided in the scenario and Table 12.5.1, fill in the rest of the ANOVA Summary Table in Table 12.5.2

Table 12.5.2: 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 12.5.3:

Table 12.5.3- 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 cells that say N/A blank; no information is needed from them. So, that leaves us with the final table looking like Table 12.5.4 showing calculated F-score of 36.86.

Table 12.5.4- 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: Conclusion

Our first step to decide if any part of the research hypothesis might be true is to see if all of the means are similar to each other or not. To do this, we do what we've been doing: compare our calculated score (now an F-score, sometimes called an omnibus F) to a critical score. Because we are using a new test statistic, we will get a new critical value table: the F distribution table (shown in the Common Critical Value Tables in the Back Matter of this book).

For the critical values of F, there are three separate tables. Each table is for a different α-level. The first tables shows when the p-value is 1% (p=.01), the second table shows p-values of 5% (p=.05) and this is the one that we'll be using, and the third table is for when the p-value is 10% (p=.10).

Each of these tables has Degrees of Freedom on the top row (for the numerator) and on the left column (for the 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 Freedom 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 formulas for how to calculate all Degrees of Freedom are included in the relevant critical value page. Let's go through this first set for ANOVA since we're adding more to do. First, the formula for df_B is k–1; 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 third 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 third 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 on the second table (for p=.05).

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. 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.

(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. $= p > .05$

Based on our three groups of 10 people, we can conclude that average job test scores are statistically significantly different based on education level. The statistical sentence for this ANOVA will be: $F(2, 27) = 36.86, p < .05$. 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 to see if any means were different from any other means, 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, which will help us decide which parts of our research hypothesis were supported.

Step 5: Write-Up is Interrupted!

Because we can't yet know which parts, if any, of our research hypothesis was supported, we cannot complete our Null Hypothesis Significance Testing process quite yet! What we know now, based on rejecting the null hypothesis, is that at least one pair of means are statistically significantly different. We do not know which means are statistically different from which other means yet.

Answers

Exercise 12.5.1

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:
  - $\overline{X}_N < \overline{X}_R$
  - $\overline{X}_N = \overline{X}_U$
  - $\overline{X}_R > \overline{X}_U$

Exercise 12.5.2

- Null Hypothesis: The average test score will be similar for each group; the degree does not affect the hiring rate.
- Symbols: $\overline{X}_N = \overline{X}_U = \overline{X}_R$

Exercise 12.5.3

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.

This page titled 12.5: Practice BG ANOVA on Job Applicants is shared under a CC BY-NC-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.

12.6: 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 omnibus F 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 set of 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 12.6.1

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 12.6.1-*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 12.6: Introduction to Pairwise Comparisons is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

12.7: Pairwise Comparison Post Hoc Tests

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 12.7.1 shows the mean test scores for the three IV levels in our job applicant scenario.

Table 12.7.1- *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 12.7.1

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 12.7.2.

Table 12.7.2- *Interpreting Pairwise Comparison Results*

Means Compared
$\overline{X}_N - \overline{X}_R = 43.7–84.3$
$\overline{X}_N - \overline{X}_U = 43.7–63.2$
$\overline{X}_R - \overline{X}_U = 84.3–63.2$

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!

Step 5: Write-Up

We can now finish the Null Hypothesis Significance Testing process that we started pages and pages ago!

Exercise 12.7.1

Write up a conclusion to the job applicant scenario.

Answer available at the end of this section.

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!

Answers

Exercise 12.7.1

Look at your conclusion. Did you include the four components that should be in all write-ups?

- Research Hypothesis, including:
  - Re-state the RH (including the groups being compared, the DV, and the direction word), and
  - If the RH was supported or not.
- Calculation results:
  - Calculated statistical (statistical sentence): F(df) = calculated omnibus F, p <.05 [or p>.05]
- Means, with units, of each group (name the groups)

This page titled 12.7: Pairwise Comparison Post Hoc Tests is shared under a CC BY-NC-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.

12.8: Practice BG ANOVA and Pairwise Comparisons 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. 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).

Let's begin!

Step 1: Hypotheses

Using the following the research question and the descriptive statistics in Table 12.8.1, 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 12.8.1- Descriptive Statistics for each IV level and the Total

Intervention:	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 12.8.2 in Example 12.8.2 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 12.8.1

What is a research hypothesis in words and symbols?

Answer available at the end of this section.

What is the null hypothesis with this scenario?

Exercise 12.8.2

State the null hypothesis in words and symbols.

Answer available at the end of this section.

Step 2: Collect Data

Student researchers designed this study and collected the data. After reading the scenario, can you identify the sample, population, IV, and DV?

Exercise 12.8.3

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 available at the end of this section.

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 12.8.4

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 $\alpha = 0.05$ with three groups and a total of 30 people?

Answer available at the end of this section.

Step 3: Analyze Data

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 12.8.4) 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 12.8.1

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 12.8.1. 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 12.8.2

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 12.8.2 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 12.8.2- 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:

$$SS_W = \sum_{EachGroup} \left[\sum \left(\left(X - \overline{X}_{group} \right)^2 \right) \right]$$

$$SS_W = \sum_{EachGroup} = [602.10 + 310.40 + 218.90]$$

$$SS_W = 1131.40$$

And, on to the final Sum of Squares!

Example 12.8.3

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 ($\overline{X}_T$ = 46.07, found in Table 12.8.1). This is again easiest to compute in a table.

Table 12.8.3 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 12.8.3- 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

IV Levels	Mindset Quiz Scores	minus Total mean	squared
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 12.8.4) so that you can fill out the rest of the table to calculate the final ANOVA F-value.

Table 12.8.4- ANOVA Summary Table with SS only

Source	SS	df	MS	F
Between	394.60			
Within	1131.40			
Total	1525.87			

Example 12.8.4

Fill out the rest of the ANOVA Summary Table.

Solution

Table 12.8.5- 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: Conclusion

You've already calculated the two Degrees of Freedom that you'll need for this analysis when you completed the ANOVA Summary Table (Table 12.8.5), but you can calculate each again using the formulas on the the Critical Values of F Table by going to the page.

Based on the completed ANOVA Summary Table (Table 12.8.5), our calculated F-score is 4.71. Using the Degrees of Freedom in the ANOVA Summary Table (Table 12.8.5), we found our critical F-score to be 3.35 from the p=.05 table on the Critical Values of F page.

$$\textbf{(Critical} < \textbf{Calculated)} = \textbf{Reject null} = \textbf{At least one mean is different from at least one other mean.} \ = p < .05$$

$$\textbf{(Critical} > \textbf{Calculated)} = \textbf{Retain null} = \textbf{All of the means are similar.} \ = p > .05$$

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. Step 3 of the Null Hypothesis Significance Testing process can be pretty long with ANOVAs!

Pairwise Comparisons

Let's start with finding the differences between each pair of means.

Example 12.8.5

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 12.8.6

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 12.8.5).

$$HSD = q * \sqrt{\frac{MSw}{n_{group}}}$$

$$HSD = 3.506 * \sqrt{\frac{41.90}{10}}$$

$$HSD = 3.506 * \sqrt{4.19}$$

$$HSD = 3.506 * 2.05$$

$$HSD = 7.18$$

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 12.8.7

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*.

Step 5: Write-Up

Okay, here's the big finish! As a reminder, the Note includes the necessary components for a write-up.

Note

- Research Hypothesis, including:
  - Re-state the RH (including the groups being compared, the DV, and the direction word), and
  - If the RH was fully supported, partially supported, or not supported.
- Calculation results:
  - Calculated statistical (statistical sentence): $F(df_{BG}, df_{WG})$ = calculated omnibus F, p <.05 [or p>.05]
  - Means, with units, of each group (name the groups)

Exercise 12.8.5

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 available at the end of this section.

Did Dr. MO's write-up include all of the components?

You did it! Take a break and reward yourself!

Answers

Exercise 12.8.1

- 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.

Exercise 12.8.2

- 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:
  - $\overline{X}_N = \overline{X}_M$
  - $\overline{X}_N = \overline{X}_S$
  - $\overline{X}_M = \overline{X}_S$

Exercise 12.8.3

1. The sample is 30 community college students.
2. The population could be community college students, or maybe community college students with women professors?
3. The IV is the type of intervention. The intervention levels are: None, Minimal, Super
4. The DV is a Mindset Quiz.

Exercise 12.8.4

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..

Exercise 12.8.5

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 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.

This page titled 12.8: Practice BG ANOVA and Pairwise Comparisons on Mindset Data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

12.9: ANOVA versus 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 12.9: ANOVA versus t-test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

12.10: 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 $12.10.1$- *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!

$$H = \left[\left(\frac{12}{(N * (N = 1)} \right) * \left(\sum \frac{R_{group}^2}{n_{group}} \right) \right] - (3 * (N = 1))$$

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 12.10: 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.

13: RM ANOVA

13.1: Introduction to Repeated Measures ANOVA
13.2: RM ANOVA Summary Table
13.3: Repeated Measures ANOVA Sum of Squares Formulas
13.4: Practice RM ANOVA Summary Table on Typing Speed
13.5: Practice RM ANOVA on Mindset Scores
13.6: Non-Parametric RM ANOVA

This page titled 13: RM ANOVA is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

13.1: 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 13.1.1

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 13.1.1

Is this more like an independent t-test or a dependent t-test?

Answer available at the end of this section.

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 $13.1.1$ uses the outdated name for the within-participant variation as "Subjects".

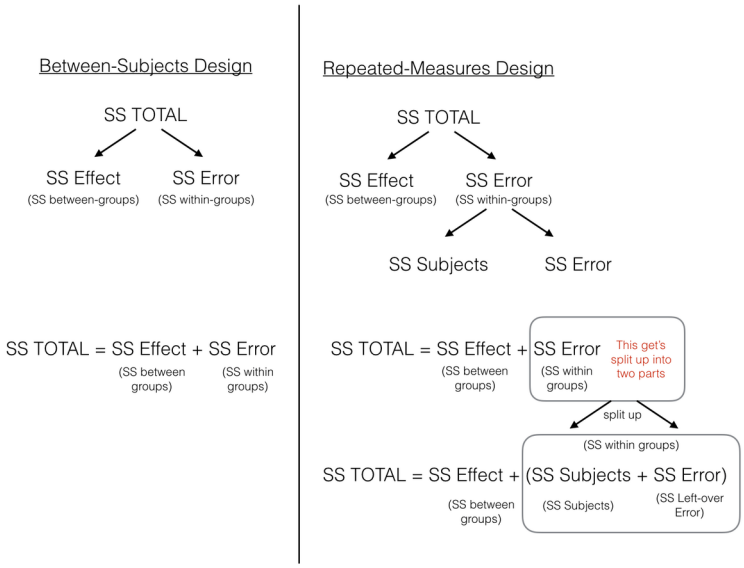


Figure 13.1.1: Illustration showing how the total sums of squares are partitioned differently for a between versus repeated-measures design. (CC-BY-SA Matthew J. C. Crump via Answering Questions with Data)

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 13.1.2:

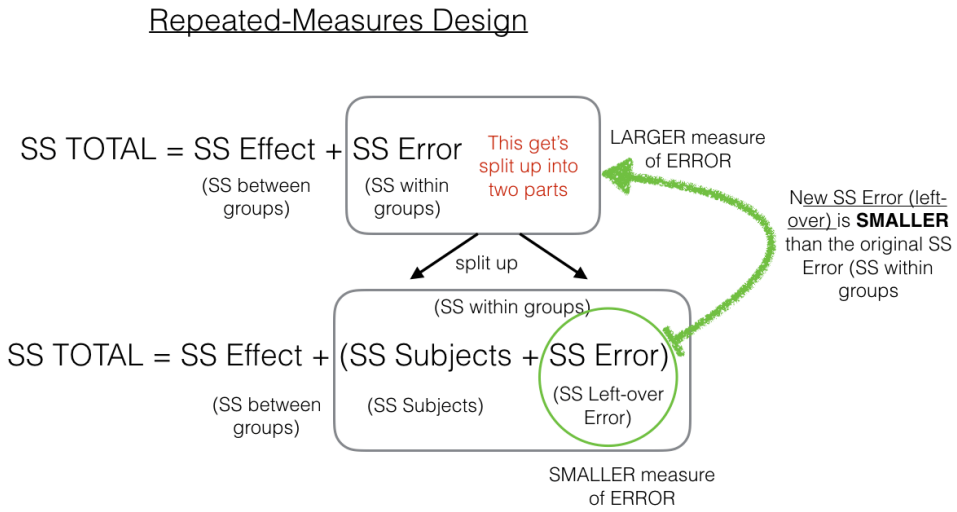


Figure 13.1.2: Close-up showing that the Error term is split into two parts in the repeated measures design. (CC-BY-SA Matthew J. C. Crump via Answering Questions with Data)

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)

Answers

Exercise 13.1.1

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).

This page titled 13.1: 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.

13.2: RM 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 13.2.1 presents an empty Repeated Measures ANOVA Summary Table.

Table 13.2.1- *RM ANOVA Summary Table*

Source	SS	df	MS	F
Between Groups				
Participants				
Within Groups (Error)				
Total				

Exercise 13.2.1

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 available at the end of this section.

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: $SS_{WGerror} = SS_{tot} - SS_b - SS_{Ps}$

Table 13.2.2 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 13.2.2- *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 13.2.3- *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:

$$\textbf{(Critical} < \textbf{Calculated)} = \textbf{Reject null} = \textbf{At least one mean is different from at least one other mean.} \ = p < .05$$

$$\textbf{(Critical} > \textbf{Calculated)} = \textbf{Retain null} = \textbf{All of the means are similar.} \ = p > .05$$

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)

Answers

Exercise 13.2.1

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).

This page titled 13.2: RM ANOVA Summary Table is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 8.3: 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.

13.3: 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 the sum of squares for every score:
    1. Add
    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 13.3: 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.

13.4: Practice RM ANOVA Summary Table on Typing Speed

Let's use a real scenario to practice with the Repeated Measures ANOVA Summary Table.

Scenario

The data are taken from a 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).

Step 1: Hypotheses

The means from Table 13.4.1 seem to look pretty different from each other. What could be a directional research hypothesis?

Table 13.4.1- *Descriptive Statistics for Reaction Time by Word Conditions*

Word Type:	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 13.4.1

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 available at the end of this section.

What about the null hypothesis? What might that look like?

Exercise 13.4.2

State the null hypothesis in words and symbols. .

Answer available at the end of this section.

Step 2: Collect Data

This study was designed and conducted by Behmer and Crump (2017). From the description of the scenario and Table 13.4.1, can you identify the sample, population, IV levels, and DV?

Example 13.4.1

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)?

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 3: Analyze Data

Using the Sum of Squares provided in the following ANOVA Summary Table (Table 13.4.2) and the information from the scenario or Table 13.4.1, fill in the ANOVA Summary Table to determine the calculated F-value.

Table 13.4.2- 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 13.4.2

Complete the ANOVA Summary Table in Table 13.4.2 to determine the calculated F-score.

Solution

Table 13.4.3- 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	$SS_{Total}–SS_{BG}–SS_{Ps}$ = 4,101,175.30 – 1,424,914 – 2,452,611.90 = 223,649.40	$(k-1) \times (P-1) = 2 \times 37 = 74$	$\frac{SS_E}{df_E} = \frac{223649.40}{74} = 3022.29$	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 13.4.4:

Table 13.4.4- 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: Conclusion

To determine whether or not the research hypothesis should be fully supported, partially supported, or not supported, we compare the calculated omnibus F from Step 3 to the critical value. You can use the sample size information included in Table 13.4.1 to calculate the Degrees of Freedom to use on the Critical Values of F Table, but you already found the relevant Degrees of Freedom when you completed the ANOVA Summary Table (Table 13.4.4)!

As shown on the bottom of the critical values page, the two Degrees of Freedom that you'll use are still from the numerator (Between Groups) and the denominator (Within Group or Error), but the denominator's df is calculated slightly differently than for Between Groups ANOVAs. This is a good reminder to check how to calculate each Degree of Freedom rather than relying on your memory or insight.

Example 13.4.3

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.

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 13.4.5- 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 13.4.5:

(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 13.4.3

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 pair of means are different from each other?

Answer available at the end of this section.

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 13.4.4

What is one problem with using t-tests to check for multiple sets of mean differences in post-hoc analyses?

Answer available at the end of this section.

Exercise 13.4.5

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 available at the end of this section.

Step 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:

- Research Hypothesis, including:
  - Re-state the RH (including the groups being compared, the DV, and the direction word), and
  - If the RH was fully supported, partially supported, or not supported.
- Calculation results:
  - Calculated statistical (statistical sentence): $F(df_{BG}, df_{WG})$ = calculated omnibus F, p <.05 [or p>.05]
  - Means, with units, of each group (name the groups).

Exercise 13.4.6

What could the conclusion look like for this scenario?

Answer available at the end of this section.

That's it! Let's try one more example; this time, we'll calculate the Sum of Squares and the pairwise comparison.

Reference

Behmer, L. P., Jr., & Crump, M. J. C. (2017). Spacial knowledge during skilled action sequencing: Hierarchical versus nonhierarchical representations. *Attention, Perception, and Psychophysics, 79*, 2435-2448. https://doi.org/10.3758/s13414-017-1389-3.

Answers

Exercise 13.4.1

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.

Exercise 13.4.2

The reaction time will be similar for the Normal condition, the Bigram condition, and the Random condition.

$\overline{X}_N = \overline{X}_B = \overline{X}_R$

Exercise 13.4.3

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.

Exercise 13.4.4

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 13.4.5

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%.

Exercise 13.4.6

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).

This page titled 13.4: Practice RM ANOVA Summary Table on Typing Speed 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/.

13.5: Practice RM ANOVA on Mindset Scores

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 developmental English class, at the end the semester in the developmental 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).

Step 1: Hypotheses

Using the descriptive statistics in Table 13.5.1, what could be the research hypothesis?

Table 13.5.1- Descriptive Statistics for each IV level and the Total

Type of English Class:	N:	Mean:	SD:
Beginning of Developmental English (D1)	12	39.17	6.46
End of Developmental English (D2)	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 13.5.2 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 13.5.1

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:

- $\overline{X}_{D1} < \overline{X}_{D2}$
- $\overline{X}_{D1} < \overline{X}_{N1}$
- $\overline{X}_{D1} < \overline{X}_{N2}$
- $\overline{X}_{D2} < \overline{X}_{N1}$
- $\overline{X}_{D2} < \overline{X}_{N2}$
- $\overline{X}_{N1} < \overline{X}_{N2}$

Despite having so many combinations, the null hypothesis for this scenario is still relatively simple, just like all null hypotheses.

Exercise 13.5.1

State the null hypothesis in words and symbols.

Answer available at the end of this section.

Step 2: Collect Data

The data has already been collected. The data was anonymized (names removed to make it anonymous) before it was provided to you, the researchers. That's why Table 13.5.2 has letters instead of names in the first column that identifies the participants. After reading the scenario, can you describe important variables from the scenario?

Example 13.5.2

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 developmental English.
2. The population could be students taking remedial English at community colleges?
3. IV: Time- Beginning of Developmental English, End of Developmental English, Beginning of Next English, End of Next English
4. DV: Mindset Quiz

Take a breath; now we will start with the messy calculations!

Step 3: Analyze Data

If your professor won't be asking you to calculate the Sums of Squares by hand, skip this part and just fill in the ANOVA Summary Table (Table 13.5.5) 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 13.5.2- Raw Mindset Quiz Scores for Four Conditions

Participant	Pretest in Developmental English	Posttest in Developmental 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 13.5.3

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 developmental English class (R1).

$$\text{R1} = \left[\left(\overline{X}_{group} - \overline{X}_T \right)^2 \times (n_{group}) \right]$$

$\overline{X}_{group}$ 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 13.5.1. 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!

$$\text{D1} = \left[(39.17–42.54)^2 \times (12) \right]$$

$$\mathrm{D1} = \left[(-3.37)^2 \times (12)\right]$$

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:

$$\mathrm{D1} = [(11.36) \times (12)]$$

$$\mathrm{D1} = [136.28]$$

And let's do that process three more times, once for the End of the Developmental 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.

$$\mathrm{D2} = \left[\left(\overline{X}_{group} - \overline{X}_T\right)^2 \times (n_{group})\right]$$

$$\mathrm{D2} = \left[(43.33–42.54)^2 \times (12)\right]$$

$$\mathrm{D2} = \left[(0.79)^2 \times (12)\right]$$

$$\mathrm{D2} = [(0.62) \times (12)]$$

$$\mathrm{D2} = [7.44]$$

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?

$$\left[\left(\overline{X}_{group} - \overline{X}_T\right)^2 \times (n_{group})\right]$$

$$\mathrm{N1} = [0.96]$$

$$\mathrm{N2} = [99.48]$$

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 13.5.4

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 13.5.3.

Table 13.5.3- Raw Mindset Quiz Scores and Calculations for Participant Sum of Squares

Participant	Pretest in Developmental English	Posttest in Developmental 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 13.5.3? 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 13.5.2 ($\sum X = 2042$). 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 13.5.5

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 13.5.1). This is again easiest to compute in a table.

Table 13.5.4 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 13.5.4- Total Mean Subtracted from Each Score, Squared, then Summed

IV Levels	Mindset Quiz Scores	minus Total mean	squared
D1	30	-12.54	157.25
D1	31	-11.54	133.17
D1	33	-9.54	91.01
D1	35	-7.54	56.85

IV Levels	Mindset Quiz Scores	minus Total mean	squared
D1	36	-6.54	42.77
D1	38	-4.54	20.61
D1	39	-3.54	12.53
D1	42	-0.54	0.29
D1	43	0.46	0.21
D1	47	4.46	19.89
D1	47	4.46	19.89
D1	49	6.46	41.73
D2	48	5.46	29.81
D2	39	-3.54	12.53
D2	35	-7.54	56.85
D2	32	-10.54	111.09
D2	37	-5.54	30.69
D2	33	-9.54	91.01
D2	45	2.46	6.05
D2	52	9.46	89.49
D2	39	-3.54	12.53
D2	51	8.46	71.57
D2	54	11.46	131.33
D2	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
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

IV Levels	Mindset Quiz Scores	minus Total mean	squared
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 13.5.6

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 13.5.5) so that you can fill out the rest of the table to calculate the final ANOVA F-value.

Table 13.5.5- ANOVA Summary Table with SS only

Source	SS	df	MS	F
Between	144.72			
Participants	1296.40			
Error	699.40			
Total	2240.00			

Exercise 13.5.2

Fill out the rest of the ANOVA Summary Table.

Answer available at the end of this section.

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 13.5.7):

Table 13.5.7- 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: Conclusion

To make a conclusion about whether the Research Hypothesis is fully supported, partially supported, or not supported, you now compare the calculated omnibus F from Table 13.5.7 to a critical value that you find in a Critical Values of F table. We'll be using p=.05. To use the Critical Value of F tables, you need the Degrees of Freedom of the omnibus F's numerator and denominator. Happily, you've already found these when you were completing Table 13.5.7. If you want to practice, you can complete Example 13.5.7.

Example 13.5.7

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 α = 0.05?

Solution

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.

$$(k-1) \times (Ps-1) = (4-1) \times (12-1) = 3 * 11 = 33$$

Based on the completed ANOVA Summary Table (Table 13.5.7), our calculated F-score is 3.84, and 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 13.5.3

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 available at the end of this section.

Example 13.5.8

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 13.5.9

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 13.5.10

What is the difference between each pair of means?

Solution

Using the means from Table 13.5.1, I subtract each mean from each other mean:

$$\overline{X}_{D1} - \overline{X}_{D2} = 39.17 - 43.33 = -4.16$$

$$\overline{X}_{D1} - \overline{X}_{N1} = 39.17 - 42.25 = -3.08$$

$$\overline{X}_{D1} - \overline{X}_{N2} = 39.17 - 45.42 = -6.25$$

$$\overline{X}_{D2} - \overline{X}_{N1} = 43.33 - 42.25 = 1.08$$

$$\overline{X}_{D2} - \overline{X}_{N2} = 43.33 - 45.42 = -2.09$$

$$\overline{X}_{N1} - \overline{X}_{N2} = 42.25 - 45.42 = -3.17$$

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 Developmental English ($\overline{X}_{D1} = 39.17$) is statistically lower than the mean of the End of the Next English Class ($\overline{X}_{N2} = 45.42$).

Write-Up

Okay, here's the big finish!

Exercise 13.5.4

Write a conclusion to describe the results of the analysis. Don't forget to include the components necessary in any report of results.

Answer available at the end of this section.

One addition to this write-up that isn't part of the four required components is the last sentence, which is 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 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...

Answers

Exercise 13.5.1

Null Hypothesis: Students will have a similar average scores on the Mindset Quiz in all four conditions: Beginning of Developmental English, End of Developmental English, Beginning of Next English Class, and End of Next English Class

Symbols: $\overline{X}_{R1} = \overline{X}_{R2} = \overline{X}_{N1} = \overline{X}_{N2}$

Exercise 13.5.2

Table 13.5.6- *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.

Exercise 13.5.3

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.

Exercise 13.5.4

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 (Developmental) class. This research hypothesis was *partially* supported (F(3,33)=3.84, p<0.05) because the mean of the Beginning of Developmental 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 ($M_{D2}=43.33, M_{N1}=42.25$). This suggests that it takes a long time to change mindset, or that the journaling activity was not a strong intervention.

This page titled 13.5: Practice RM ANOVA on Mindset Scores is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

13.6: 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
- *ANOVA*s
- Pearson's correlation

Alternatives

So far, we've talked about non-parametric analyses for t-tests and Between Groups ANOVAs.

Example 13.6.1

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 13.6: Non-Parametric RM ANOVA is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

14: FACTORIAL ANOVA (TWO-WAY)

14.1: Introduction to Factorial Designs

14.2: Factorial Notations and Square Tables

14.3: Introduction to Main Effects and Interactions

14.4: Example with Main Effects and Interactions

14.5: Graphing Main Effects and Interactions

14.6: Interpreting Main Effects and Interactions in Graphs

14.7: Do Main Effects Matter?

14.8: Beyond 2x2 in Graphs

14.9: Two-Way ANOVA Summary Table

14.9.1: Calculating Sum of Squares for the Factorial ANOVA Summary Table

14.10: When Should You Conduct Post-Hoc Pairwise Comparisons?

14.11: Practice Factorial on Attention

14.12: Practice Factorial on Mindset Scores

14.13: Choosing the Correct Analysis- Mean Comparison Edition

This page titled 14: Factorial ANOVA (Two-Way) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

14.1: 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 14.1.1

What would be the IV and IV levels? What would be the DV?

Answer available at the end of this section.

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 14.1.1 is an example:



Figure 14.1.1: 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)

Answers

Exercise 14.1.1

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.

This page titled 14.1: 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/.

14.2: 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 14.2.1 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 14.2.1- *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 14.2.1

Create a "Punnett's Square" for the IVs and DV of the Distraction scenario.

Solution

Table 14.2.2- *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 14.2.1

What is the factorial design notation for a study with two IVs, one has 2 levels and the other has 3 levels?

Answer available at the end of this section.

Exercise 14.2.2

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 available at the end of this section.

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 14.2.5

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 available at the end of this section.

Last one! This one has a few more questions to better understand the scenario.

Exercise 14.2.6

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 available at the end of this section.

Just for fun, let's illustrate a 2x3 design using the same kinds of tables we looked at before for the 2x2 design.

Table 14.2.3- *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 14.2.2

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 14.2.4- *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 14.2.5- *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 14.2.5- *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)

Answers

Exercise 14.2.1

2x3

There are two IVs, so there are two numbers. Each number represents the number of levels for each IV.

Exercise 14.2.2

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.

Exercise 14.2.5

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!

Exercise 14.2.6

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.

This page titled 14.2: 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/.

14.3: 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 14.3.1). Remember, the bar graphs in Figure 14.3.1 show the mean for each group; they are not frequency charts.

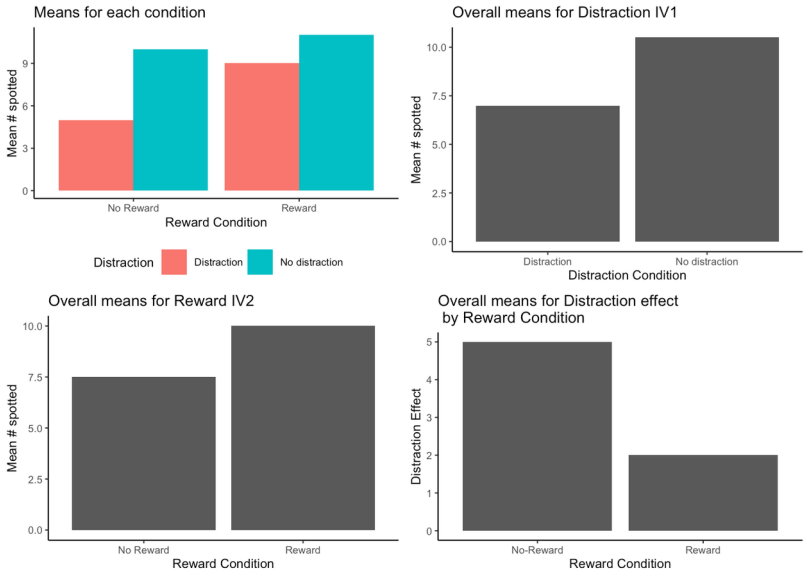


Figure 14.3.1: Each panel shows the mean for different effects in the design. (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

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 14.3: 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/.

14.4: 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 14.4.1

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 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 14.4.2

Complete a grid showing the factorial design IVs and DVS.

Solution

Table 14.4.1- *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 14.4.3

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 14.4.2, 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 14.4.2- *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 14.4.2.

Example 14.4.1

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:

$$\textbf{Critical } < |Calculated| = \textbf{Reject null} = \textbf{means are different} = \textbf{main effects } = p < .05$$

$$\textbf{Critical } > |Calculated| = \textbf{Retain null} = \textbf{means are similar} = \textbf{no main effects } = p > .05$$

Let's move on to the other independent variable.

Exercise 14.4.1

Based on Table 14.4.2, what are the marginal means for department? Which department seemed to have a higher Difference score?

Answer available at the end of this section.

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 14.4.3 has the means for each combination of each IV level in the individual cells. Looking at Table 14.4.3, how do the inside cells seem relate to each other?

Table 14.4.3- 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 14.4.5

Answer the following questions related the cell means in Table 14.4.3,

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.

Answers

Exercise 14.4.1

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.

This page titled 14.4: 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.

14.5: 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 14.5.1 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.

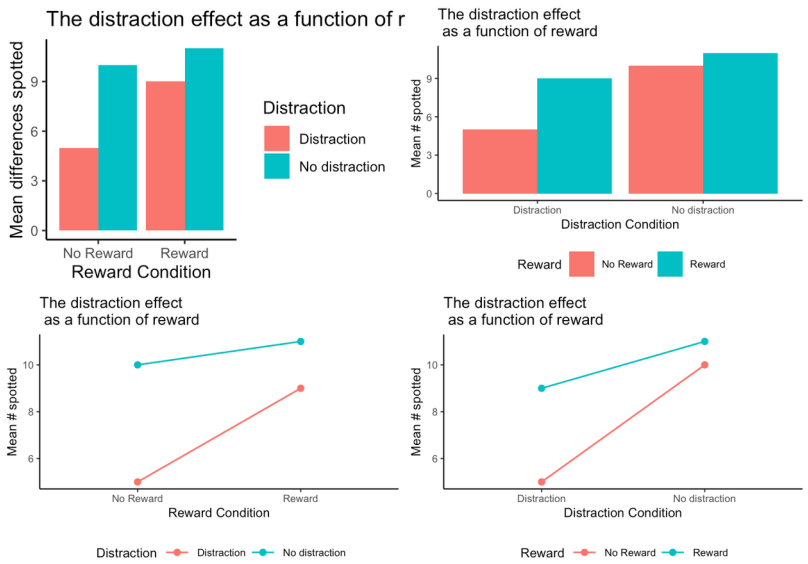


Figure 14.5.1: The same example means plotted using bar graphs or line graphs, and with Distraction or Reward on the x-axis. (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

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 14.5.1, 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 14.5.1? 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 14.5.2 and determine the factorial notations. Figure 14.5.2 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.

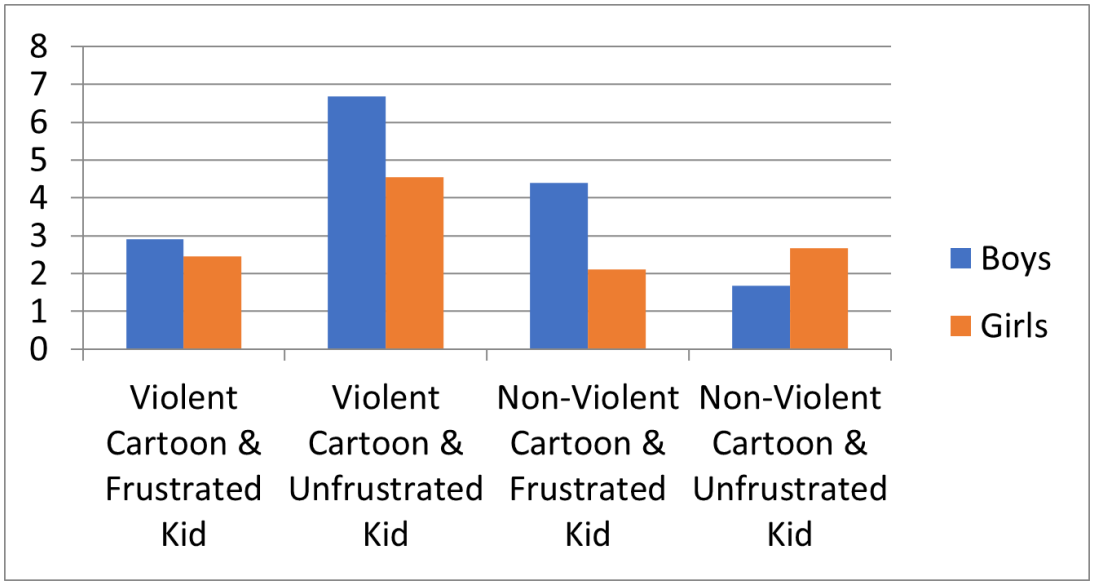


Figure 14.5.2: Example of Factorial Design in a Bar Chart. (CC-BY-SA; graph created by Michelle Oja)

Let's answer the questions that we talked about when identifying factorial notations about Figure 14.5.2 to see if we can find the main effects and interactions.

Exercise 14.5.3

For Figure 14.5.2:

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 available at the end of this section.

Now that we see the main effects, let's determine what kind of factorial design is shown in Figure 14.5.1.

Exercise 14.5.4

What kind of factorial designs is shown in Figure 14.5.2? 2x2? 2x3? 2x2x2? Other?

Answer available at the end of this section.

Contributors and Attributions

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

Answers

Exercise 14.5.3

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".

Exercise 14.5.4

Figure 14.5.2 shows a 2x2x2

2 (Gender: Boy or Girl) by 2 (Cartoon: Violent or Not) by 2 (Mood: Frustrated or Not)

This page titled 14.5: 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/.

14.6: 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 14.6.1.

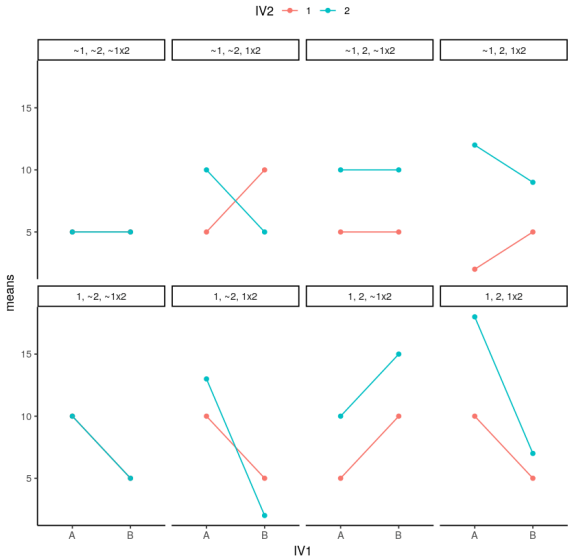


Figure 14.6.1: Line graphs showing 8 possible general outcomes for a 2x2 design. (CC-BY-SA Matthew J. C. Crump from 10.1 of Answering Questions with Data)

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).

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Figure 14.6.2: 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 14.6: 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/.

14.7: 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.

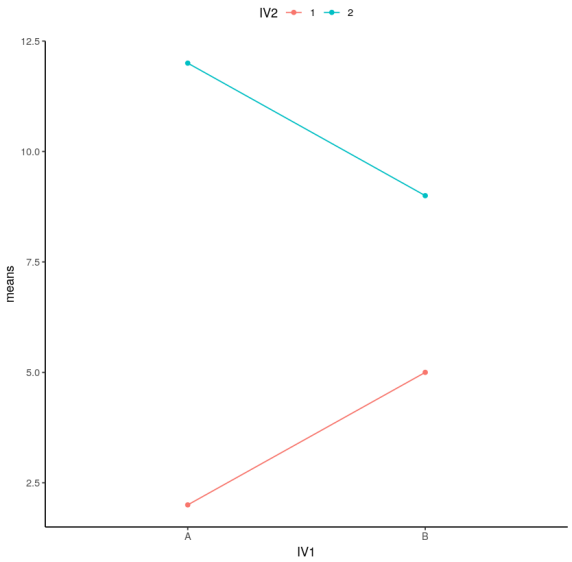


Figure 14.7.1: Example means showing a generally consistent main effect along with an interaction. (CC-BY-SA Matthew J. C. Crump from 10.2 of Answering Questions with Data)

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 14.7.1- *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 14.7.1, 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 14.7.2. 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.

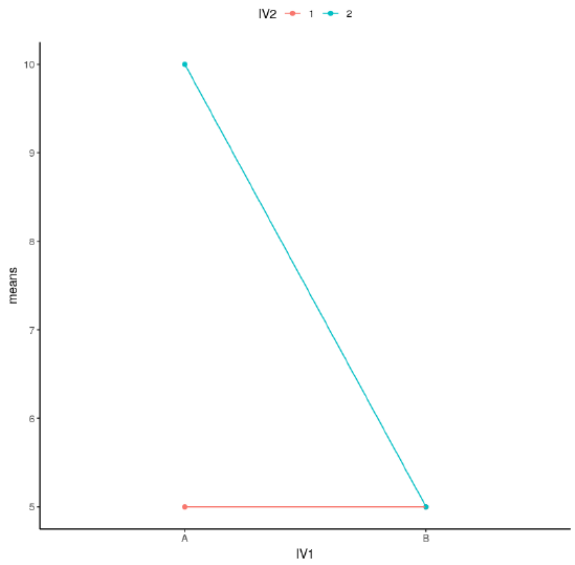


Figure 14.7.2: Example data showing how an interaction exists, and a main effect does not, even though the means for the main effect may show a difference. (CC-BY-SA Matthew J. C. Crump from 10.2 of Answering Questions with Data)

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 14.7.2- *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 14.7.2, 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 14.7.1

In Table 14.7.2, 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 14.7.3- *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 14.7.4 and Figure 14.7.3, and decide if you think that the main effects are consistent or inconsistent with the interaction.

Table 14.7.4 shows the marginal means and the cell means for our two IVs (Intervention: Yes or No; Department: English or Psychology).

Table 14.7.4- *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 14.7.4. 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 14.7.2

Does one cell in the middle of Table 14.7.4 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 14.7.3, 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.

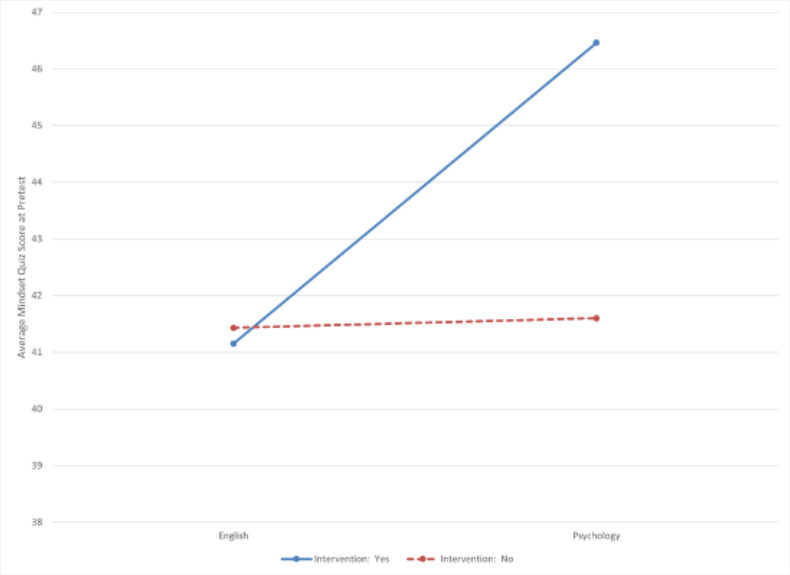


Figure 14.7.3: Mindset Quiz Pretest Means. (CC-BY-SA; Michelle Oja)

Figure 14.7.3 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 14.7.3

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 14.7.4

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 14.7: 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/.

14.8: 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 14.8.1. 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.

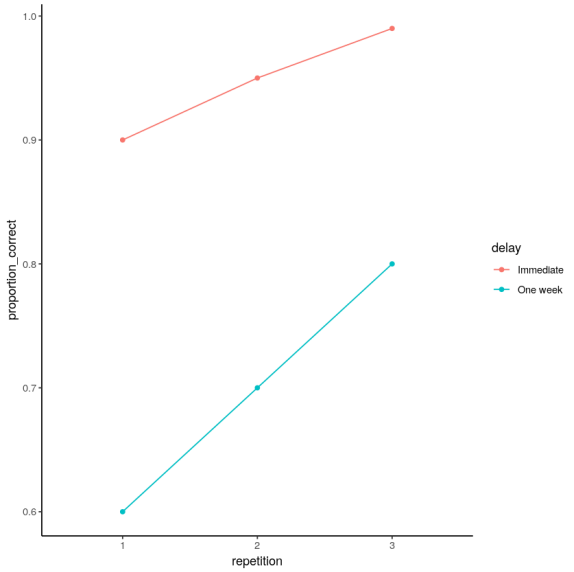


Figure 14.8.1: Example means for a 2x3 factorial design. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

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 14.8.2.

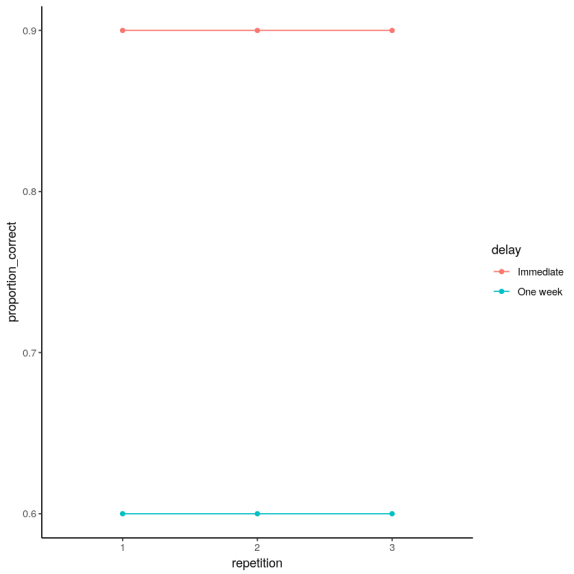


Figure 14.8.2: Example means for a 2x3 design when there is only one main effect. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

What would you say about the interaction if you saw something like Figure 14.8.3? 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.

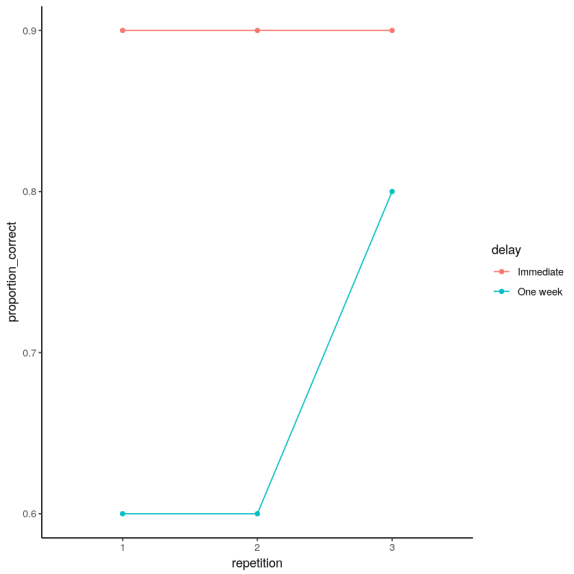


Figure 14.8.3: Example means for a 2x3 design showing another pattern that produces an interaction. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

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 14.8.4 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).

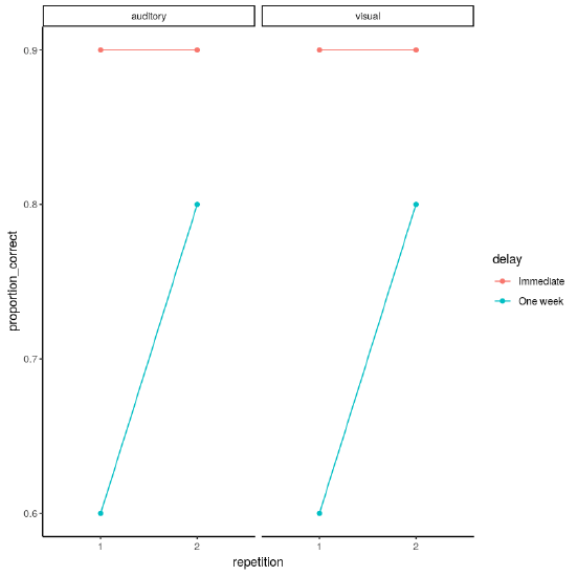


Figure 14.8.4: Example means from a 2x2x2 design with no three-way interaction. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

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 14.8.5).

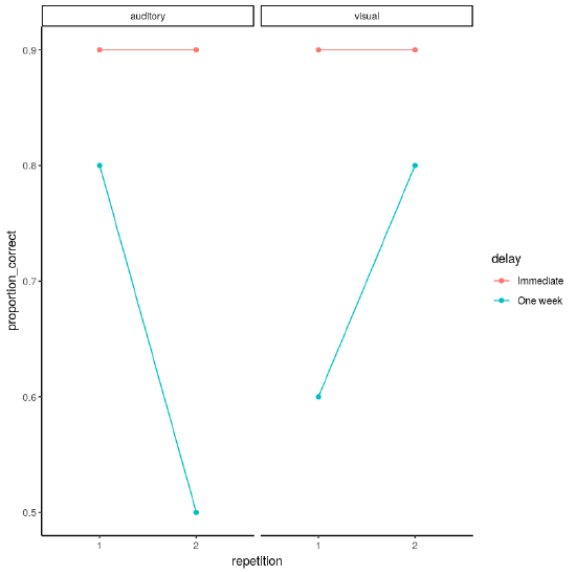


Figure 14.8.5: Example means from a 2x2x2 design with a three-way interaction. (CC-BY-SA Matthew J. C. Crump via 10.4 in Answering Questions with Data)

We are looking at a 3-way interaction between modality, repetition and delay in Figure 14.8.5. 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 14.8: 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/.

14.9: 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 14.9.1?

Table 14.9.1*: 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 14.9.2)?

Table 14.9.2*- 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 14.9.3:

Table 14.9.3- *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 14.9.1

What were the names of the three rows called? In other words, what were the names of the three Sums of Squares?

Answer available at the end of this section.

With dependent sample Repeated Measures ANOVAs, we needed four Sums of Squares and four degrees of freedom.

Exercise 14.9.2

What were the names of the four rows called? In other words, what were the names of the four Sums of Squares?

Answer available at the end of this section.

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} / $MS_{WG\text{ or }Error}$

For a two-way ANOVA, that looks like:

Exercise 14.9.3

What is similar to past F-score calculations? What is similar amongst all three of these calculations?

Answer available at the end of this section.

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 14.9.4.

Table 14.9.4- *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 14.9.1

Example 1 Complete the following ANOVA Summary Table with the information provided. The Sums of Squares are already in Table 14.9.5.

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 14.9.5- 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 14.9.1

Complete a grid showing the factorial design IVs and DVS for the scenario in Example 14.9.1.

Solution

Add text here.

Table 14.9.6- 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 14.9.1

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.

Answers

Exercise 14.9.1

1. Between Groups
2. Within-Groups (or Error)
3. Total

Exercise 14.9.2

1. Between Groups
2. Participants
3. Within-Groups (or Error)
4. Total

Exercise 14.9.3

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.

This page titled 14.9: Two-Way ANOVA Summary Table is shared under a CC BY-NC-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.

14.9.1: 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:

$$\overline{X}_{\text{condition}} - \overline{X}_{\text{IV1}} - \overline{X}_{\text{IV2}} + \overline{X}_{\text{Grand Mean}}$$

Or we could write it this way:

$$\overline{X}_{\text{condition}} - \overline{X}_{\text{Grand Mean}} + \overline{X}_{\text{IV1}} + \overline{X}_{\text{IV2}}$$

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;">0.0000011
Reward	1	110.45	110.45	69.03125	F)" style="vertical-align:middle;">0.0000003
Distraction:Reward	1	14.45	14.45	9.03125	F)" style="vertical-align:middle;">0.0083879
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 14.9.1: 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/.

14.10: 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 14.10.1

Did we conduct pairwise comparisons when we *retained* the null hypothesis when comparing two groups with a t-test? Why or why not?

Answer available at the end of this section.

Exercise 14.10.2

Did we conduct pairwise comparisons when we *rejected* the null hypothesis when comparing two groups with a t-test? Why or why not?

Answer available at the end of this section.

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 14.10.3

Did we conduct pairwise comparisons when we *retained* the null hypothesis when comparing three groups with an ANOVA? Why or why not?

Answer available at the end of this section.

Exercise 14.10.4

Did we conduct pairwise comparisons when we *rejected* the null hypothesis when comparing three groups with an ANOVA? Why or why not?

Answer available at the end of this section.

Finally, on to factorial designs! If you've been answering the Exercises as you go, these should be pretty easy.

Exercise 14.10.5

Do we conduct pairwise comparisons when we *retain* the null hypothesis for main effects in a factorial design? Why or why not?

Answer available at the end of this section.

Okay, this one might be a little challenging, so we'll walk through it together.

Example 14.10.1

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 14.10.6

Do we conduct pairwise comparisons when we *retain* the null hypothesis for an interaction in a factorial design? Why or why not?

Answer available at the end of this section.

This one should be clear if you understand the reasoning for when we do and do not conduct post-hoc pairwise comparisons.

Example 14.10.2

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 14.10.1- *Short Answer for When to Conduct Post-Hoc Pairwise Comparisons*

When to Conduct 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!

Answers

Exercise 14.10.1

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 14.10.2

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.

Exercise 14.10.3

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 14.10.4

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.

Exercise 14.10.5

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.

Exercise 14.10.6

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 page titled 14.10: 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.

14.11: Practice Factorial on Attention

Let's go through the process of looking at a 2x2 factorial design about selective attention.

Scenario

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 that we are basing our calculations on is from a paper called "Stand by your Stroop: Standing up enhances selective attention and cognitive control" (Rosenbaum et al., 2017), although Dr. MO tweaked the experiment to simplify ANOVA Summary Table calculations a tiny bit. Rosenbaum et al. (2017) asked whether sitting versus standing would influence a measure of selective attention, the ability to ignore distracting information.

Rosenbaum et al. (2017) 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 14.11.1. 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!

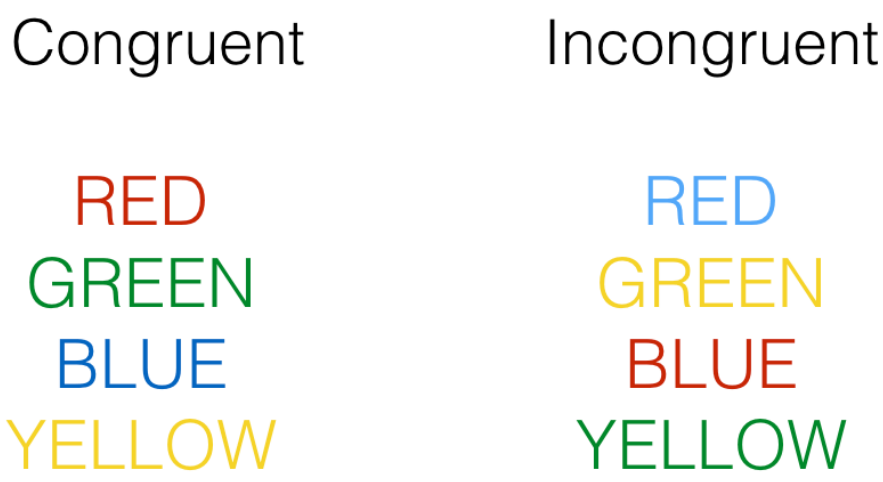


Figure 14.11.1: Examples of congruent and incongruent Stroop stimuli. The task is to name the color, not the word. (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

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 Rosenbaum et al.'s (2017) study was 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.

Step 1. 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 14.11.2 is a bar graph of the means.

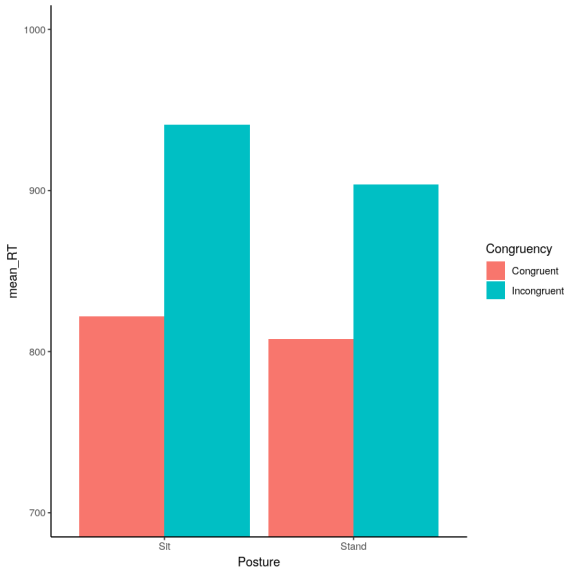


Figure 14.11.2: Means from Rosenbaum et al (2017). (CC-BY-SA Matthew J. C. Crump from Answering Questions with Data- Introductory Statistics for Psychology Students)

We can see in Figure 14.11.2 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 14.11.1

Use the information in Figure 14.11.2 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 14.11.1

What are the null hypotheses for this scenario (in words)?

Answer available at the end of this section.

Step 2. Collect Data

Rosenbaum et al. (2017) used the Stroop effect to collect data on selective attention. Can you identify the variables to show that you understand the scenario?

Exercise 14.11.2

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 available at the end of this section.

Step 3: Analyze Data

We'll complete an ANOVA Summary to see 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 14.11.1- *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 14.11.2

Use the Sum of Squares provided in Table 14.11.1 to fill in the rest of the ANOVA Summary Table in Table 14.11.2.

Solution

Table 14.11.2- *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:

$$\begin{aligned} df_{cells} &= (k_1 * k_2) - 1 \\ df_{cells} &= (2 * 2) - 1 \\ df_{cells} &= 4 - 1 \\ df_{cells} &= 3 \end{aligned}$$

That wasn't so hard! Now, let's figure out what this all means.

Step 4. Conclusion

With factorial designs, there's a lot of means to compare to other means, so it's likely that some of the mean pairs are statistically

significantly different. To start to figure out which means are different from each other, we must compare each calculated F-score to a critical score. The first step to that comparison, is finding the critical scores in the Critical Value of F table with each F-scores' Degrees of Freedom. You've actually found all of these when you 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 practice with the Degrees of Freedom formulas.

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 14.11.3

What are the Degrees of Freedom for each of the six sources if N = 50?

Answer available at the end of this section.

Well, that was a lot!

Now, let's use these to find the critical values.

Exercise 14.11.4

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 available at the end of this section.

To make the decision, we need to compare each calculated F-score to the critical values that we found in Step 2.

Exercise 14.11.5

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 available at the end of this section.

That wasn't so hard, but now what?

Example 14.11.3

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 14.11.3.

Table 14.11.3-*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 14.11.4

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 14.11.4- *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

Pairs Being Compared	Mean Difference	Statistically Significant?
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.

Step 5: Write-Up

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 14.11.3, 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 14.11.3, 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 14.11.3 and the mean differences in Table 14.11.4, 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 14.11.5

Are all of the required components when Reporting Results included?

Solution

- Research Hypothesis, including:
  - Re-state the RH (including the groups being compared, the DV, and the direction word), and
  - If the RH was fully supported, partially supported, or not supported.
- Calculation results:
  - Calculated statistical (statistical sentence): $F(df_{BG}, df_{WG})$ = calculated omnibus F, p <.05 [or p>.05]
  - Means, with units, of each group (name the groups)

Summary

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)

Answers

Exercise 14.11.1

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 interaction is that the average reaction for Congruency and Posture combined will be similar. [You could list out all of the combinations here again, too.]

Exercise 14.11.2

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. Incongruent color while sitting

Exercise 14.11.3

1. Cells: $(k_1 \times k_2)–1 = (2 \times 2) - 1 = 4 - 1 = 3$
2. Between group for one variable (IV1): $k_1–1 = 2 - 1 = 1$
3. Between group for the other variable (IV2) $k_2)–1 = 2 - 1 = 1$
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: $N - 1 = 50 - 1 = 49$

Exercise 14.11.4

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).

Exercise 14.11.5

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.

This page titled 14.11: Practice Factorial on 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/.

14.12: Practice Factorial on Mindset Scores

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 a conference posting session of prior studies in this series, the student researchers had a discussion with a 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.

Step 1. Hypotheses

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 14.12.1

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 available at the end of this section.

What about the null hypotheses?

Exercise 14.12.2

What are the null hypotheses for this scenario (in words)?

Answer available at the end of this section.

Step 2. Collect Data

Researchers in the student club collected this data, but let's see how well you understand what's going on in this scenario.

Exercise 14.12.3

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 available at the end of this section.

Step 3. Analyze Data

We'll skip calculating the Sums of Squares to focus on practicing the other parts of NHST of factorial designs; you can find the Sums of Squares already in the ANOVA Summary Table (Table 14.12.1).

Table 14.12.1- 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 14.12.1

Use the Sum of Squares provided in Table 14.12.1 to fill in the rest of the ANOVA Summary Table in Table 14.12.2.

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:

$$\begin{aligned} df_{cells} &= (k_1 \times k_2)–1 \\ df_{cells} &= (2 \times 3) - 1 \\ df_{cells} &= 6 - 1 \\ df_{cells} &= 5 \end{aligned}$$

Table 14.12.2-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 14.12.3

Table 14.12.3-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. Conclusion

You've found the Degrees of Freedom when you filled in the ANOVA Summary Table (Table 14.12.2), so you can move on to finding the critical values for each calculated F-score.

Exercise 14.12.4

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 (Gender)?
2. Main Effect of IV 2 (Mindset Intervention)?
3. Interaction?

Answer available at the end of this section.

To make the decision, we need to compare each calculated F-score to the critical values that we found.

Note

Remember:

$$\textbf{(Critical} < \textbf{Calculated)} = \textbf{Reject null} = \textbf{At least one mean is different from at least one other mean.} = p < .05$$

$$\textbf{(Critical} > \textbf{Calculated)} = \textbf{Retain null} = \textbf{All of the means are similar.} = p > .05$$

Exercise 14.12.5

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.

Answer available at the end of this section.

Factorial designs are more complex, but it's the same basic process that we've been working through this whole time.

Step 5: Write-Up

Before we move on to the full write-up of results, let's just make sure that we know how to write the "statistical sentences".

Exercise 14.12.6

What would the statistical sentences look like for each effect?

Answer available at the end of this section.

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 14.12.4. 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 14.12.4-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 14.12.2

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 14.12.3

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 14.12.4

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 ($F(2,400)=0.99$, $p>.05$). The means by gender and mindset intervention did not statistically significantly interact. The means for each combination can be seen in Table 14.12.4.

Before you move on, look at back at your own write-ups to makes sure that each of them includes all of the required components of write-ups.

Summary

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 14.12.5

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!

Answers

Exercise 14.12.1

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...

Exercise 14.12.2

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.]

Exercise 14.12.3

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

Exercise 14.12.4

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), $F_{Crit}(2, 40) = 3.23$.

Exercise 14.12.5

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.

Exercise 14.12.6

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).

This page titled 14.12: Practice Factorial on Mindset Scores is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

14.13: 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 14.13.1

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 14.13: 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/.

15: CORRELATIONS

15.1: Refresh to Prepare
15.2: What do Two Quantitative Variables Look Like?
15.3: Introduction to Pearson's r
15.4: Correlation versus Causation
15.4.1: Correlation versus Causation in Graphs
15.5: Strength, Direction, and Linearity
15.6: Hypotheses
15.7: Correlation Formula- Covariance Divided by Variability
15.8: Practice Correlation on Anxiety and Depression
15.9: Practice Correlation on Nutrition
15.10: Alternatives to Pearson's Correlation
15.11: Final Considerations

This page titled 15: Correlations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

15.1: Refresh to Prepare

We are heading into a whole new territory now! In the first part of this book, we looked at how to describe data with the measures of central tendency (mean, median, and mode), the measure of variability (standard deviation), and graphing. Then in the second part of this book, we looked at using inferential statistics to compare means with t-tests and ANOVAs. That's a short sentence to describe all that you learned in that second section!

Exercise 15.1.1

What kind of variables can we calculate a mean from: Quantitative or Qualitative?

Answer available at the end of this section.

Now, we are leaving means behind, and focusing on *relationships* between variables.

But before we do that, let's make sure that we remember the difference between qualitative and quantitative variables.

Exercise 15.1.2

What type of variable is each of the following: Quantitative or Qualitative?

1. Major in college
2. Age
3. Gender
4. Test scores (in points)
5. Ounces of vodka

Answer available at the end of this section.

In this section of the textbook, we are going to focus on health. If we wanted to compare the nutritional value of different food items at different fast food restaurants, we could run a variety of statistical analyses.

As a reminder, a t-test has one quantitative variable (DV) and one qualitative variable with two levels (IV).

Example 15.1.1

If we wanted to compare the average calories for cheeseburgers from different two different fast food restaurants:

1. What is the quantitative DV? In other words, what is being measured?
2. What is the IV and its two levels? The IV levels are the qualitative groups.

Solution

1. The DV would be the average calories for the cheeseburgers. Calories are numeric, so they are quantitative.
2. The IV is the two different restaurants.

An ANOVA also has one quantitative variable and one (or more) qualitative variables.

Example 15.1.2

If we wanted to compare the average calories for three or more types of food (cheeseburger, chicken sandwich, salad) from one fast food restaurant:

1. What is the quantitative DV? In other words, what is being measured?
2. What is the IV and its levels? The IV levels are the qualitative groups.

Solution

1. The DV would be the average calories for the cheeseburgers. Calories are numeric, so they are quantitative.
2. The IV is the different types of food, with the levels being cheeseburger, chicken sandwich, or salad.

(Or more...)

Exercise 15.1.3

If we wanted to compare the average calories for three or more types of food (cheeseburger, chicken sandwich, salad) from two different fast food restaurants:

1. What is the quantitative DV? In other words, what is being measured?
2. What is one of the IVs and its levels? The IV levels are the qualitative groups.
3. What is one of the IVs and its levels? The IV levels are the qualitative groups.
4. What kind of factorial design would that be? 2x2? 2x3? 2x2x2? Something else?

Answer available at the end of this section.

In this chapter, we will analyze scenarios in which we have only two quantitative variables. Although we could look at the means of each different variable, it doesn't really make sense to compare differences in means of two different variables. For example, let's say that our two variables were cheeseburger calories and cheeseburger price. We could find the average calories of all of the cheeseburger was 350 calories, and the average prices was $4.49. Without doing any statistical analyses, I can say that 350 is higher than 4.49, so comparing means doesn't give us any useful information when we have two quantitative variables. Remember way back in the section on IVs and DVs when we talked about IVs being predictors and DVs being outcomes? The IV is the predictor that we think is what we causes the outcome, which is the DV.

This is still true with two quantitative variables. We can predict that the IV of calories predicts the outcome (DV) of cost. We can use inferential statistics to test is if prices go up when calories go up. Thus, the research question changes from "Are these means from different populations?" to something more like, "Do these numbers vary together?" We follow the Null Hypothesis Significance Testing process that we've been doing:

1. Step 1: State the Hypotheses
2. Step 2: Find the Critical Values
3. Step 3: Compute the Test Statistic
4. Step 4: Make the Decision, then report the results!

Now that you've been refreshed, let's move on to our next statistical analysis: Pearson's Correlation

Answers

Exercise 15.1.1

Means can only be calculated for quantitative variables.

Exercise 15.1.2

1. Major is qualitative (named categories or groups)
2. Age is quantitative (number)
3. Gender is qualitative (named categories or groups)
4. Test scores (in points) is quantitative (number)

5. Ounces of vodka is quantitative (number)

Exercise 15.1.3

1. The DV would be the average calories for the cheeseburgers. Calories are numeric, so they are quantitative.
2. One IV is type of food, and the levels are cheeseburger, chicken sandwich, or salad. These are all qualitatively different foods, which are different categories.
3. Another IV is the restaurant, with two different restaurants being the two levels.
4. 2x3: Restaurant (2 levels) by Food Type (Cheeseburger, Chicken Sandwich, or Salad)

This page titled 15.1: Refresh to Prepare is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

15.2: What do Two Quantitative Variables Look Like?

Let's begin learning about two quantative variables with a scenario about chocolate and happiness.

Chocolate and Happiness Scenario

Let's imagine that a person's supply of chocolate has a *causal influence* on their level of happiness. This means that having chocolate is the reason why someone is happy. Let's further imagine that the more chocolate you have the more happy you will be, and the less chocolate you have, the less happy you will be. (What tends to blow my mind is that many causal relationships can also go the other way; what if I eat chocolate every time that I'm happy? That means that being happy is the reason why I might eat chocolate!) 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 today?, and
2. How happy are you today?

For convenience, both the scales will go from 0 to 100. For the chocolate scale, 0 means no chocolate, 100 means eating chocolate constantly from the moment they woke up. 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.

We asked each participants our two questions so there are two scores for each participant, one for their chocolate supply and one for their level of happiness. Although they both are on a 0-10 scale, they are different questions, so finding the means of each won't necessarily help us if chocolate predicts happiness. Table 15.2.1 show data from the first 10 imaginary participants, who all had very low ratings on our chocolate scale and on our happiness scale.

Table 15.2.1- *Chocolate and Happiness Scores*

Participant	Chocolate	Happiness
A	1	1
B	1	1
C	2	2
D	2	4
E	4	5
F	4	5
G	7	5
H	8	5
I	8	6
J	9	6

Look at Table 15.2.1. Do you notice any relationship between amount of chocolate and level of happiness? We can see that there is variance in chocolate supply across the 10 participants. We can see that there is variance in happiness across the 10 participants. Variance means that the numbers have some change in them, they are not all the same, some of them are big, some are small.

In particular, does it look like participants with more chocolate tend to score higher on the happiness scale? 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 seem to vary together. When we have two measures that vary together, they are like a happy couple

who share their variance. Chocolate and happiness seem to co-vary (meaning that the two variables seem to vary together). 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.

To make this co-variance even more clear, let's plot all of the data in a graph. Think back to the chapter on graphs, then answer the following question:

Exercise 15.2.1

Which type of graph is used when we have two quantitative variables and want to see them both combined?

Answer available at the end of this section.

Graphing Two Quantitative Variables

The scatter plot in Figure 15.2.1 shows 100 dots for each participant. 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 participant, there are 100 participant, so there are 100 dots. Each dot is placed so that it represents *both* measurements for each participant. In other words, each dot has two coordinates, an x-coordinate for chocolate, and a y-coordinate for happiness.

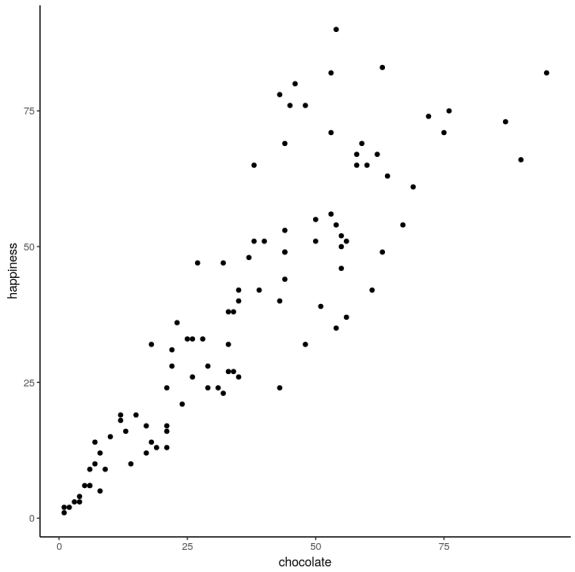


Figure 15.2.1: 100 Participants' Amount of Chocolate and Amount Happiness. (CC-BY-SA, Matthew J. C. Crump via Answer Questions with Data)

What do the dots mean? The dot all the way on the bottom left is a participant 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 participant has. You can draw a straight line left to the y-axis: that will tell you how much happiness the participant has.

Positive, Negative, Curvilinear, and No Relationship

Now that we are looking at the scatter plot, we can see many things. It looks like 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*. In this sense, "positive" doesn't mean "good." Instead, it means "going in the same direction."

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. What do you imagine the scatter plot would look like if the relationship was reversed, and larger chocolate supplies decreased happiness? This means that the more chocolate you might have, the less happy

you are. This is called a *negative correlation* because as one variable increase, the other variable decreases. Again, "negative" isn't saying that anything is bad, but that the variables are going in the opposite direction. Instead of a general trend going up and to the right, a negative correlation goes starts at the upper left of the graph and has a trend down to the lower right.

What do you imagine the scatter plot would look like if people were least happy with no chocolate and with lots of chocolate? This would be a curvilinear (curved line) relationship, and is described as a reverse-U because the general trend looks like an upside down U.

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? That kind of scatter plot just lots like a bunch of dots, scattered randomly.

Correlations

Now that we know that two quantitative variables can be related, can we infer a statistical relationship? Yes! That's what a Pearson's r (or Pearson's correlation) measures. A positive Pearson's r means that there is a positive linear relationship; as one variable increase, the other variable also increases. If you calculated a correlation with the sample of 10 participants from Table 15.2.1, you'd get r(8)=0.86, p<.05. The calculated r is positive and the p-value is less than 0.05, so we can say that these 10 participants show a positive linear relationship between chocolate and happiness.

A negative Pearson's r means that there is a negative linear relationship; as one variable increases, the other variable decreases. For example, the more you miss class (number of absences), the lower your final grade percentage; as absences increase, grades decrease.

The tricky part is when Pearson's r is close to zero. This could show that there is no relationship between the two quantitative variables, *OR* it could mean that there is a curvilinear relationship. Pearson's r only looks for linear relationships (how close to a straight line all of the dots form), so it can't tell the difference between a random splatter of dots and clearly curved line. That's why looking at the scatter plot is so important.

Cause and Effect

We are wading into the idea that one variable causes the change in another variable (having more chocolate causes more happiness), but Pearson's r only measures linear relationships, it can't tell you what causes happiness. Maybe something else causes both happiness and chocolate supply to increase? Dr. MO always have more chocolate around the house at the end of October, but she also loves costumes, so maybe Halloween causes an increase in chocolate supply and an increase in happiness. Fair warning: we will find patterns that look like one thing is causing another, even when that one thing DOES NOT CAUSE the other thing. Hang in there.

Contributors and Attributions

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

Answers

Exercise 15.2.1

A scatterplot plots one quantitative variable on the x-axis and the other quantitative variable on the y-axis. This makes a "scatter" of dots in which scores that are low on both variables are towards the bottom left, and scores that are high on both variables are towards the top right.

This page titled 15.2: What do Two Quantitative Variables 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/.

15.3: Introduction to Pearson's r

There are several different types of correlation coefficients, but we will only focus on the most common: Pearson's r. r is a very popular correlation coefficient for assessing *linear* relations, and it serves as both a descriptive statistic (like $\overline{X}$) and as an inferential test statistic (like t). It is descriptive because it describes what is happening in the scatterplot; r will have both a sign (plus for positive or minus for negative) for the direction and a number (from -1.00 to 1.00) for the magnitude (strength). As noted above, Pearson's r assumes a linear relation, so nothing about r will suggest what the shape the dots tend towards; the correlation statistic will only tell what the direction and magnitude would be if the form is linear. Always make a scatterplot!

r also works as a test statistic because the magnitude of r will correspond directly to a t value as the specific degrees of freedom, which can then be compared to a critical value. We will again have a table of r critical values that we can compare our r directly to those to decide if we retain or reject the null hypothesis.

The formula for r is very simple: it is just the covariance (defined above) divided by the standard deviations of X and Y:

$$r = \frac{\text{cov}_{XY}}{s_X s_Y}$$

The first formula gives a direct sense of what a correlation is: a covariance standardized onto the scale of X and Y; the second formula is computationally simpler and faster. Both of these equations will give the same value, and as we saw at the beginning of the chapter, all of these values are easily computed if you use a sum of products table (which will be discussed later). When we do this calculation, we will find that our answer is always between -1.00 and 1.00 (if it's not, check the math again), which gives us a standard, interpretable metric.

This page titled 15.3: Introduction to Pearson's r is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 12.4: Pearson's r by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

15.4: Correlation versus Causation

We cover a great deal of material in introductory statistics, including how to interpret statistical information. Hopefully you've seen how the statistical results from quality studies can be used in your career, and day to day life, to help you make better decisions. We now come to what may be the most important lesson in introductory statistics: the difference between variables that are related (correlated) and a variable that causes a change in another variable (causation).

It is **very, very tempting** to look at variables that are correlated (have a significant Pearson's r) and assume that this means they are *causally* related; that is, it gives the impression that what we call the IV is *causing* changes in what we call the DV.

However, in reality, Pearson's correlational analysis doe not – and cannot – do this. Correlations DO NOT show causation. No matter how logical or how obvious or how convenient it may seem, no correlational analysis can demonstrate causality. The ONLY way to demonstrate a causal relation is with a properly designed and controlled experiment that rules out all of the other things that could have affected the DV.

Many times, we have prior information that suggests that one variable causes changes in the other. Thus, when we run our Pearson's r and find a strong, statistically significant results, it is very tempting to say that we found the causal relation that we are looking for. The reason we cannot do this is that, without an experimental design that includes random assignment and control variables, is that the relation we observe between the two variables may be caused by something else that we failed to measure. These “third variables” are lurking variables or confound variables, and they are impossible to detect and control for without an experiment.

Note: TW for drowning

A common example of this is the strong, positive correlation between ice cream sales and drowning; as ice cream sales increase, so does death by drowning. Does eating ice cream cause drowning? Probably not. Does drowning cause people to have eaten ice cream? Definitely not. Could there be a third variable? Yes! There is also a strong, positive correlation between outside temperatures and both ice cream sales and drowning. When it gets hotter outside, more people buy ice cream. And when it's hotter, more people go swimming (which just, statistically, leads to more drowning). Remember this example when you start thinking that a significant correlation means that one variable *caused* the change in the other!

Confound variables, which we will represent with C, can cause two variables (X and Y) to appear related when in fact they are not. They do this by being the hidden cause of each variable independently. That is, if C causes X and Z causes Y, the X and Y will appear to be related . However, if we control for the effect of C (the method for doing this is beyond the scope of this text), then the relation between X and Y will disappear. As another example, consider shoe size and spelling ability in elementary school children. Although there should clearly be no causal relation here (why would bigger feet lead to better spelling? Or why would better spelling lead to bigger feet?), the variables are nonetheless consistently correlated. The confound in this case? Age. Older children spell better than younger children and are also bigger, so they have larger shoes.

When there is the possibility of confounding variables being the hidden cause of our observed correlation, we will often collect data on C as well and control for it in our analysis. This is good practice and a wise thing for researchers to do. Thus, it would seem that it is easy to demonstrate causation with a correlation that controls for C. However, the number of variables that could potentially cause a correlation between X and Y is functionally limitless, so it would be impossible to control for everything. That is why we use experimental designs; by randomly assigning people to groups and manipulating variables in those groups, we can balance out individual differences in any variable that may be our cause.

It is not always possible to do an experiment, however, so there are certain situations in which we will have to be satisfied with our observed relation and do the best we can to control for known confounds. However, in these situations, even if we do an excellent job of controlling for many extraneous (a statistical and research term for “outside”) variables, we must be very careful not to use causal language. That is because, even after controls, sometimes variables are related just by chance.

Sometimes, variables will end up being related simply due to random chance, and we call these correlation *spurious*. Spurious just means random, so what we are seeing is random correlations because, given enough time, enough variables, and enough data, sampling error will eventually cause some variables to be related when they should not. Sometimes, this even results in incredibly strong, but completely nonsensical, correlations. This becomes more and more of a problem as our ability to collect

massive datasets and dig through them improves, so it is very important to think critically about any significant correlation that you encounter.

The next section talks more about spurious correlations and other issues with trying to say that one variable causes changes in the other.

This page titled 15.4: Correlation versus Causation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 12.7: Correlation versus Causation by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

15.4.1: Correlation versus Causation in Graphs

What does the presence or the absence of a correlation between two measures mean? How should correlations be interpreted? As discussed previously, and will be discussed in more detail soon, a correlational analysis can only show the strength and direction of a linear relations. Let's use graphs to show that correlation does not equal causation.

Sometimes there's no correlation, but there is causation.

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. But, like all plants, snake plants do need some water to stay alive. Unfortunately, they 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. The amount of water given each snake plant per day is one of our variables, probably the IV because it is part of the causal process that allows snake plants to grow. Every week the experimenter measures snake plant growth, which will be the second measurement. Plant growth is proably our DV, because we think that water will cause growth. Now, can you imagine for yourself what a scatter plot of weekly snake plant growth by tablespoons of water would look like?

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, perhaps even negative 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 Dr. Crump's 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. The scatter plot could look something like Figure 15.4.1.1:

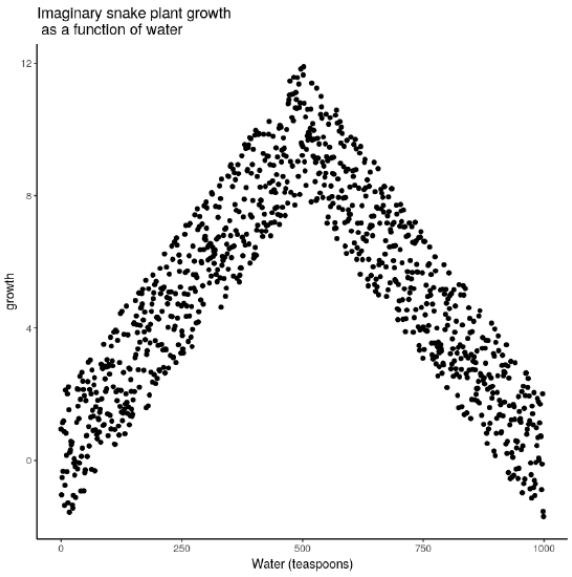


Figure 15.4.1.1: Illustration of a possible relationship between amount of water and snake plant growth. (CC-BY-SA Matthew J. C. Crump via Answering Questions with Data)

Granted this looks more like an inverted V, than an inverted U, but you get the idea right? Growth goes up with water, but eventually goes back down as too much water makes snake plants die.

Computing Pearson's r for data like this can give you r values close to zero. Based on this statistical analysis, there is no relationship between water and growth, but the scatterplot clearly shows that there is a relations. It's just not a linear relationship (a straight line), so Pearson's r won't find it. As a result, when we compute the correlation in terms of Pearson's r, we get a value suggesting no relationship. What this really means is there is no *linear* relationship (there is no relationship between the two variables that can be described by a single straight line). When we need lines or curves going in more than one direction, we have a non-linear, or cuvilinear, 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 (as water goes up, growth goes up), if we look at the last half of the data we see a negative correlation (as water goes up, growth goes down), 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.

This is one reason why plotting your data is so important. If you see a U-shaped or reverse-U shape pattern, then a correlation analysis is probably not the best analysis for your data. There are statistical analyses that will work with these curvilinear relationships, but they are beyond the scope of an introductory statistics textbook.

Confound It: Sometimes there's a correlation, but something else causes both variables.

We discussed this "third variable" issue previously. Can you think of two quantitative variables that are related, but only because they are both caused by something else? A statistically significant 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.

For example, in one of Dr. MO's courses, she found a positive correlation between the number of pens in students' bags and their final grade percentage. Does having more pens actually cause students to learn more and earn more points? Probably not. It's mroe likely that students are maybe over-achievers or want to be totally prepared have more pens in their bags, and also study in ways that result in learning.

Correlation and Random Chance: Sometimes there's a correlation that's really a Type I Error

Another very important aspect of correlations is the fact that they can be produced by random chance. This means that you can find a statistically significant 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 get back to the final grades and pens-in-the-bag example. Once Dr. MO found that correlation, she's tried to replicate it with other classes for several years, and has never been able to. It appears that we had one wonky sample that produced a signficant correlation, but that there is actually no real correlation between pens and grades in the population of students.

Watching Random Correlations

In Figure 15.4.1.2, Dr. Crump wrote code to 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 (N), from 10, 50, 100, to 1,000 in each sample.

Remember, because these are randomly sampled numbers, *there should be no relationship* between the two variables. But, as we have been discussing, because of chance, we can sometimes observe a correlation due to chance alone, a Type I Error. The important thing to watch is how the line behaves across the four panels when you see these online. This line shows the best-fit line for all of the data. The closer that the dots are to the line, the stronger the correlation. As you can see, 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 1,000, but much less. In all cases we expect that the line should be parallel with the horizon (x-axis),

but every time there's a new sample, sometimes the line shows us pseudo patterns. The best fit line is not very stable for small sample-sizes, but becomes more reliably flat as sample-size increases.

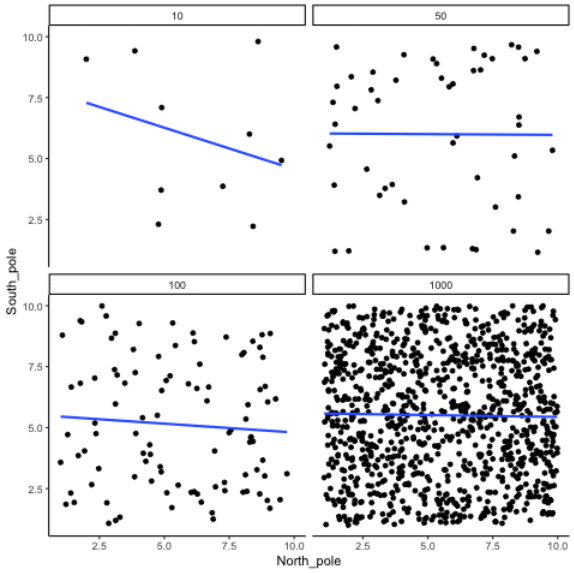


Figure 15.4.1.2: Animation of how correlation behaves for completely random X and Y variables as a function of sample size. (CC-BY-SA Matthew J. C. Crump via Answering Questions with Data)

Which line should you trust? Well, hopefully you can see that the line for 1000 samples is the most stable. It tends to be parallel to the horizon 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.

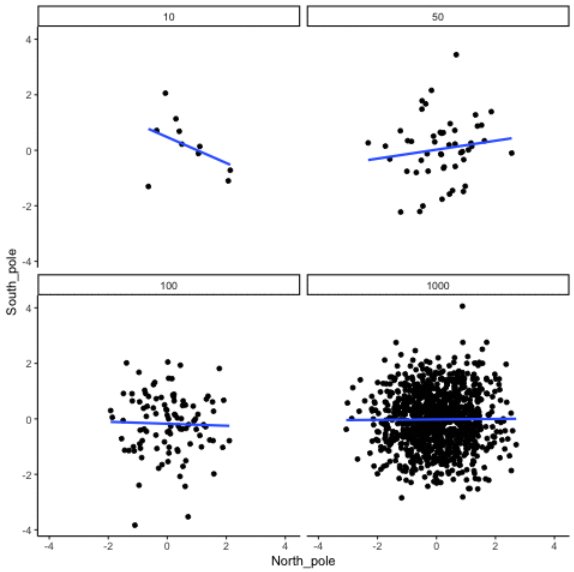


Figure 15.4.1.3: Animation of correlation for random values sampled from a normal distribution, rather than a uniform distribution. (CC-BY-SA Matthew J. C. Crump via Answering Questions with Data)

OK, so what do things look like when there actually is a correlation between variables?

Watching Real Correlations

Sometimes there really are correlations between two variables that are not caused by chance. Figure 15.4.1.4 has four more animated scatter plots. Each shows the correlation between two variables. Again, we change the sample-size in steps of 10, 50 100, and 1,000. The data have been programmed to contain a real positive correlation (as the scores on the x-axis variable increase, scores on the y-axis variable should also increase). Positive correlations have a trend that goes up and to the right. 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.

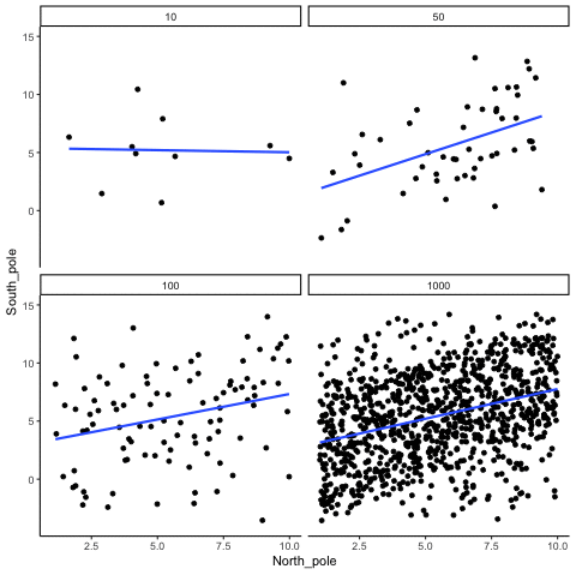


Figure 15.4.1.4: How correlation behaves as a function of sample-size when there is a true correlation between X and Y variables. (CC-BY-SA Matthew J. C. Crump via Answering Questions with Data)

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, like the sample that Dr. MO found with her pen data. 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.

Contributors and Attributions

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

This page titled 15.4.1: Correlation versus Causation in Graphs is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 3.6: Interpreting Correlations by Matthew J. C. Crump is licensed CC BY-SA 4.0. Original source: https://www.crumplab.com/statistics/.

15.5: Strength, Direction, and Linearity

Strength, Direction, and Linearity

We've talk about how Pearson's r can be from -1.00 to 1.00. How close the result is to $|1.00|$ (the absolute value of 1.00) shows how strong the correlation is; the closer to $|1.00|$, the more the dots follow a straight line (linearity). With a strong relationship, it's easier to predict the value of one variable from the value of the other variable. The negative or positive part of the result shows whether the variables vary in the same way; one the scores on one variable increase, do the scores on the other variable also increase (positive correlation) or do they decrease (negative correlation).

Scenario

Let's turn to a topic close to every parent's heart: sleep. The following data set is fictitious, but based on real events experienced by one of the authors, Dr. Danielle Navarro (a.k.a., Dani).

Suppose Dr. Navarro was curious to find out how much her infant son's sleeping habits affected her mood. Let's say that she rated her grumpiness very precisely, on a scale from 0 (not at all grumpy) to 100 (as grumpy as it get). And, lets also assume that she measured her grumpiness, her sleeping patterns, and her son's sleeping patterns for quite some time. Let's say, for 100 days. And, being a nerd, Dr. Navarro calculated some basic descriptive statistics, shown in Table 15.5.1:

Table 15.5.1- *Descriptive Statistics of Sleeping & Grumpiness Study*

Variable	N	Mean	SD	Median	Minimum	Maximum
Dani's Sleep	100	6.97	1.02	7.03	4.84	9.00
Baby's Sleep	100	8.05	2.07	7.95	3.25	12.07
Dani's Grumpiness	100	63.71	10.05	62.00	41.00	91.00

When you look at Table 15.5.1, what do you see?

Exercise 15.5.1

What can you say about each of the variables based on information in Table 15.5.1?

Answer available at the end of this section.

To give a graphical depiction of what each of the three interesting variables looks like, Figure 15.5.1 plots histograms.

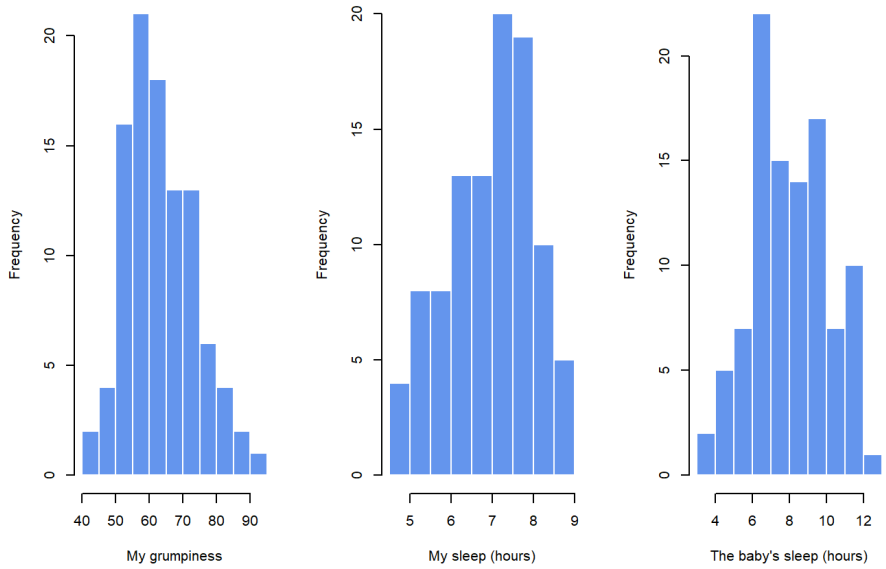


Figure 15.5.1- Histograms for Dani's Grumpiness, Dani's Sleep, and Baby's Sleep (CC-BY-SA Danielle Navarro from Learning Statistics with R)

Do our answer in Exercise 15.5.1 match what is shown in Figure 15.5.1? The histograms aren't quite as symmetrical as Dr. MO interpreted the means and medians to show! Do you see anything new in Figure 15.5.1?

Strength and Direction of a Relationship

Let's use a different kind of graph to look at these variable. Figure 15.5.2 shows a scatterplot of Dani's Sleep (x-axis) and Dani's Grumpiness (y-axis), and Figure 15.5.3 shows Dani's grumpiness on the y-axis again, but now the Baby's sleep is on the x-axis.

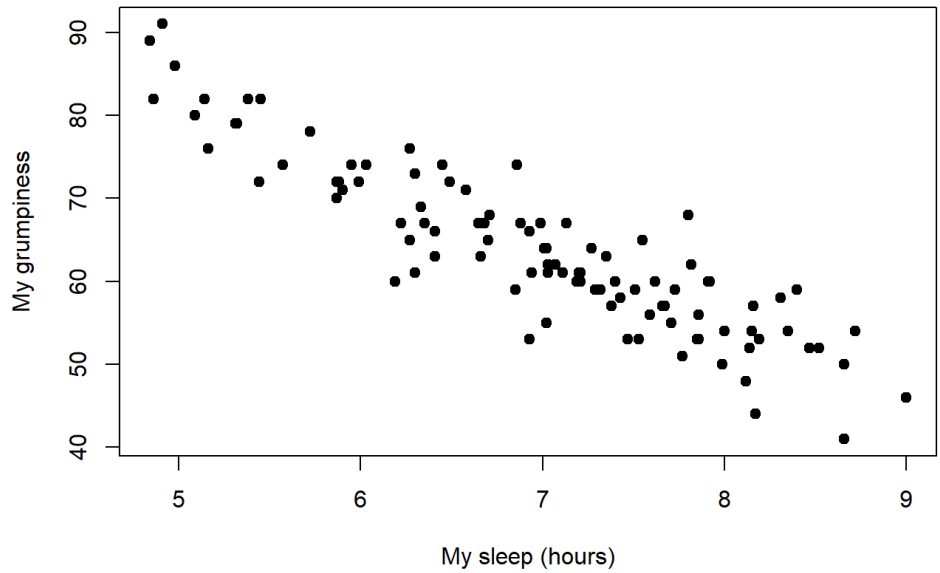


Figure 15.5.2- Scatterplot showing the relationship between Dani's Sleep and Dani's Grumpiness (CC-BY-SA Danielle Navarro from Learning Statistics with R)

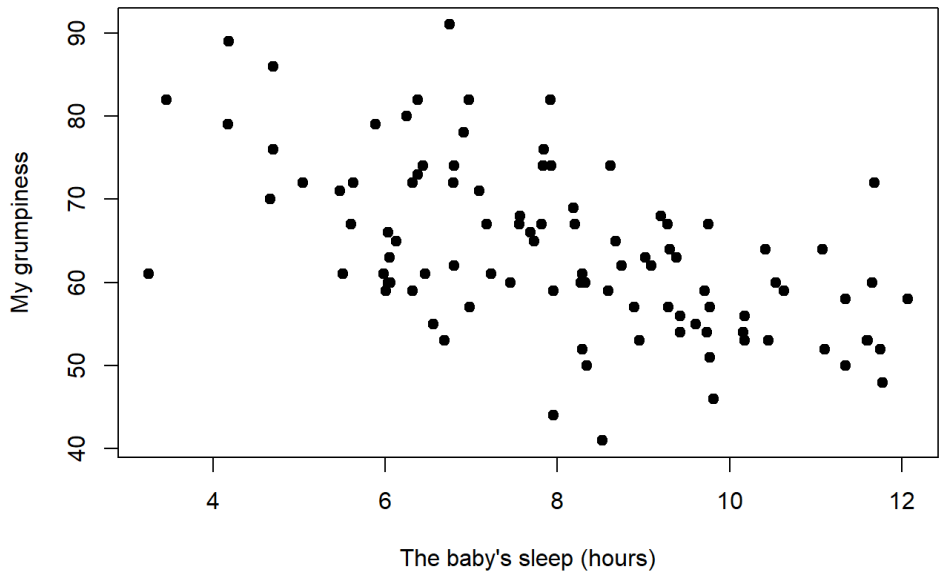


Figure 15.5.3- Scatterplot showing the relationship between Baby's Sleep and Dani's Grumpiness (CC-BY-SA Danielle Navarro from Learning Statistics with R)

When looking at these two plots side by side, it's clear that the relationship is *qualitatively* the same in both cases: more sleep equals less grump! We can see as the sleep hours (on the bottom, x-axis) get higher, the grumpiness gets lower (y-axis); this would be a negative correlation because the variable vary in *opposite* directions.

However, it's also pretty obvious that the relationship between Dani's Sleep and Dani's Grumpiness is *stronger* than the relationship between Baby's Sleep and Dani's Grumpiness. How is this clear? Because Figure 15.5.2 is "neater" than the one on the right. The dots are closer together, and it's easier to see how each dot might be closer to a straight line that we could draw through the data. What it feels like is that if you want to predict what Dani's mood is, it'd help you a little bit to know how many hours her son slept, but it'd be *more* helpful to know how many hours she slept.

In contrast, let's consider Figure 15.5.3 versus Figure 15.5.4. If we compare the scatterplot of Baby's Sleep & Dani's Grumpiness with the scatterplot of Baby's Sleep and Dani's Sleep, the overall *strength* of the relationship is the same (the dots seem about as messy and far from a hypothetical line through the middle), but the *direction* is different. That is, if Dani's son sleeps more, Dani seems to get *more* sleep (positive correlation), but if he sleeps more then she get *less* grumpy (negative correlation).

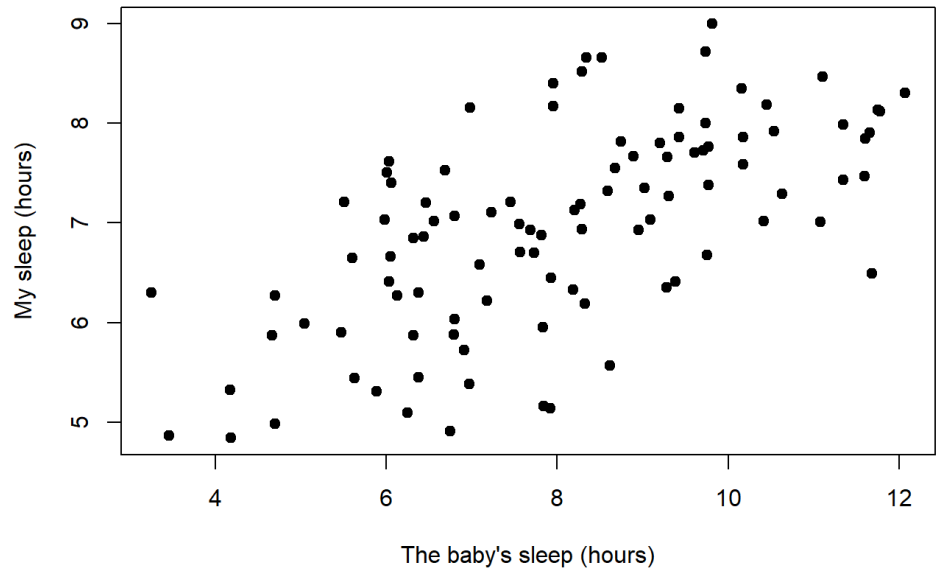


Figure 15.5.4- Scatterplot showing the relationship between Baby's Sleep and Dani's Sleep (CC-BY-SA Danielle Navarro from Learning Statistics with R)

The Correlation Coefficient

As a refresher, Pearson's correlation coefficient, which is traditionally denoted by r, shows the relationship betweentwo variables, and is a measure that varies from –1.00 to 1.00. When r=–1.00 it means that we have a perfect negative relationship, and when r=1.00 it means we have a perfect positive relationship. When r=0, there's no relationship at all. If you look at Figure 15.5.5, you can see several plots showing what different correlations look like.

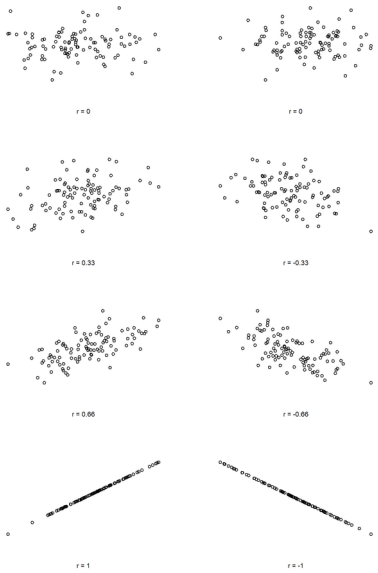


Figure 15.5.5- Illustration of the effect of varying the strength and direction of a correlation (CC-BY-SA Danielle Navarro from Learning Statistics with R)

As you can see, strong correlations (shown on the bottom, r-values close to -1.00 or 1.00) are straight lines of dots. The closer to zero, the less linear the dots are, until you get to a correlation of zero (r=0.00) and there's just a random splatter of dots (top scatterplots). You can also see that positive correlations "point" up and to the right; as one variable increases, the other also increases. Negative correlations "point" down and to the right; as one variable increases, the other variable decreases.

Interpreting a Correlation

Naturally, in real life you don't see many correlations of 1.00. So how should you interpret a correlation of, say r = .40? The honest answer is that it really depends on what you want to use the data for, and on how strong the correlations in your field tend to be. A friend of mine in engineering once argued that any correlation less than .95 is completely useless (I think he was exaggerating, even for engineering). On the other hand there are real cases – even in psychology – where you should really expect correlations that strong. For instance, one of the benchmark data sets used to test theories of how people judge similarities is so clean that any theory that can't achieve a correlation of at least .90 really isn't deemed to be successful. However, when looking for (say) elementary correlates of intelligence (e.g., inspection time, response time), if you get a correlation above .30 you're doing very very well. In short, the interpretation of a correlation depends a lot on the context. That said, the rough guide in Table 15.5.2 is pretty typical.

Table 15.5.2- General Correlation Interpretations

Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative

Correlation	Strength	Direction
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive

However, something that can never be stressed enough is that you should *always* look at the scatterplot before attaching any interpretation to the data. As was said in prior sections, a correlation might not mean what you think it means. The classic illustration of this is "Anscombe's Quartet" (Anscombe, 1973), which is a collection of four data sets. Each data set has two variables, an X and a Y. For all four data sets the mean value for X is 9 ($\overline{X}_{x-axis} = 9.00$) and the mean for Y is 7.5 ($\overline{X}_{y-axis} = 7.50$). The, standard deviations for all X variables are almost identical, as are those for the the Y variables. And in each case the correlation between X and Y is r=0.816. You'd think that these four data sets would look pretty similar to one another. They do not. If we draw scatterplots of X against Y for all four variables, as shown in Figure $15.5.6$ we see that all four of these are *spectacularly* different to each other.

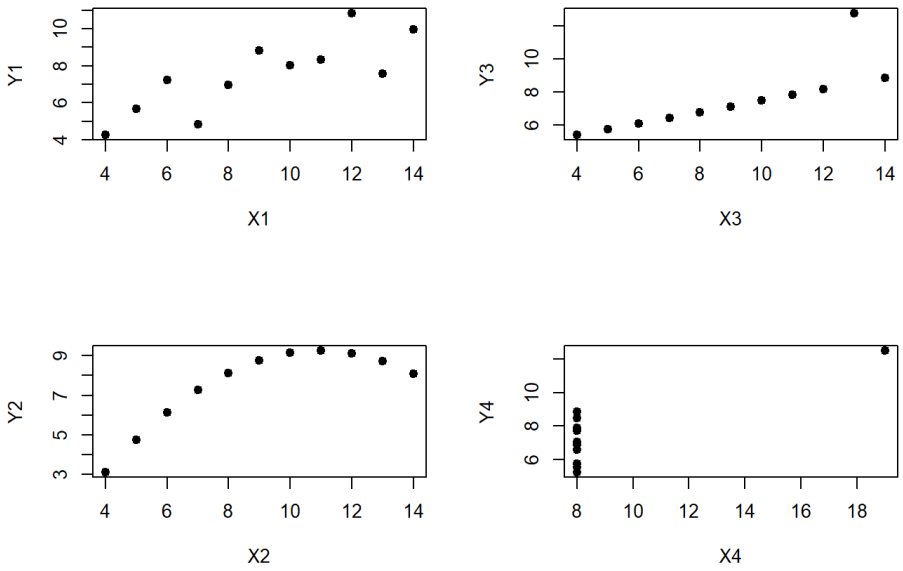


Figure $15.5.6$- Anscombe's quartet. All four of these data sets have a Pearson correlation of r=.816 (CC-BY-SA Danielle Navarro from Learning Statistics with R)

The lesson here, which so very many people seem to forget in real life is "*always graph your raw data*".

Now that you know a little more about what correlations can show, let's talk about hypotheses.

Reference

Anscombe, F. J. (1973). Graphs in statistical analysis, *The American Statistician, 27*(1), 17-21).

Contributors and Attributions

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

Answers

Exercise 15.5.1

For Dani's Sleep, Dr. MO notices the low Minimum of 4.84 hours; at least once, Dani got less than 5 hours of sleep one night!

Looking at the Baby's sleep, I also see a very low Minimum. The standard deviation (SD) is also twice that of Dani's, suggesting that the baby's sleep is much more variable. And that is confirmed when seeing that on night the baby slept only 3.25 hours but one night he slept over 12 hours! Babies are wierd.

And finally, Dani's Grumpiness seems sorta high since the average and the median are higher than the middle of the scale of 50, suggesting that she tends towards very grumpy.

Because the medians and means are pretty close, Dr. MO thinks that the distributions are fairly symmetrical (not skewed).

This page titled 15.5: Strength, Direction, and Linearity is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

15.6: Hypotheses

As we've been learning, Pearson's correlation coefficient, r, tells us about the strength and direction of the linear relationship between two variables. This is the basis of our research hypothesis. We perform a hypothesis test of the "significance of the correlation coefficient" to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population; the null hypothesis is that there is no relationship between the variables.

Research Hypothesis

Pearson's r measures is the strength and direction of a linear relationship between two quantitative variables; what does this look like as a research hypothesis? As with all inferential statistics, the sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

- The symbol for the population correlation coefficient is ρ, the Greek letter "rho."
- $\rho =$ population correlation coefficient (unknown)
- $r =$ sample correlation coefficient (known; calculated from sample data)

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is "significant," and can state that there is sufficient evidence to conclude that there is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population. We'll talk about the regression line much more in the next chapter. For now, just know that it's a line that tries to be as close as possible to all of the data points. The closer the dots are to this straight line, the stronger the correlation. The conclusion means that there is a significant linear relationship between x and y. This leads to the research hypothesis of:

> Pearson's Correlation research hypothesis: There is a (positive *or* negative) **linear** relationship between one quantitative variable and another quantitative variable (name the variables).

With research hypotheses for mean differences, we specified which mean(s) would be bigger than which other means. With a correlation, we specify whether the two variables vary together (positive correlation: both go up *or* both down down) or vary in opposite directions (negative correlation: as one variable increases, the other variable decreases). With all hypothesis, it's important to include the names of the variables and what was measured. In Dr. Navarro's example the name of the variable was what was measured (Dani's Sleep, Baby's Sleep, Dani's Grumpiness). This is nice because in the prior examples comparing mean differences, we had to identify the IV, the levels of the IV (the groups), and the DV (what was measured). With correlations, the IV is usually one variable (sometimes called the predictor) and the DV is usually the other variable (sometimes called the outcome). But be careful! These names are starting to suggest that one variable (the predicting IV) causes the changes in the other variable (outcome DV), but we learned earlier (Correlation versus Causation Part 1 and Part 2) that correlations just show a linear relationship, not whether one variable *caused changes* in the other variable.

Null Hypothesis

If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant," and state that there is insufficient evidence to conclude that there is a significant linear relationship between x and y. This conclusion means that there is not a significant linear relationship between x and y. This leads to the null hypothesis of:

> Pearson's Correlation research hypothesis: There is **no** linear relationship between one quantitative variable and another quantitative variable (name the variables).

We CANNOT use the regression line to model a linear relationship between x and y in the population.

Making the Decision

Null hypothesis significance testing lets us decide whether the value of the population correlation coefficient ρ is "close to zero," meaning that there is no linear relationship between the two variable in the population; when one variable changes, we know nothing about the other variables changes. We reject this null hypothesis base on the sample correlation coefficient r and the sample size n.

The Table of Critical Values of r page (or found through the Common Critical Value Table page) is used to give you a good idea of whether the computed value of r is significant or not. Compare the absolute value of your calculated r to the appropriate critical value in the table. If calculated r is bigger than the critical value then the correlation coefficient is significant. If r is significant, then you may use the line for prediction.

With null hypothesis significance testing, we either retain the null hypothesis (we don't think that there is a linear relationship between the two variables) or we reject the null hypothesis (we think that there is a linear relationship between the two variables). Again, rejecting the null hypothesis does not mean that the data automatically support the research hypothesis. To support the research hypothesis, the correlation has to be in the direction (positive or negative) that we predicted in our research hypothesis.

Example 15.6.1

What do you think the results sentence will look like?

Solution

The statistical sentence would be: r(df)=r-calc, p __ .05

In which everything underlined is replaced.

Now that we know a little bit about what correlational analyses shows, let's look at the actual formulas to calculate Pearson's r.

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 15.6: Hypotheses is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 12.5: Testing the Significance of the Correlation Coefficient by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

15.7: Correlation Formula- Covariance Divided by Variability

The best way to learn the formula for correlations is to learn about two ideas and what they look like mathematically. We'll start with co-variance, which will become the numerator (top). Then we'll talk about standard deviations AGAIN for the denominator (bottom). To preview where we're going, here's the formula when we already have the standard deviation computed:

$$r = \frac{\left(\frac{\sum((x_{Each} - \overline{X_x}) \times (y_{Each} - \overline{X_y}))}{(N-1)}\right)}{(s_x \times s_y)}$$

We will go through each part below!

Numerator: Co-Variation

Because we have two quantitative variables, we will have two characteristics or score on which people will vary. What we want to know is do people vary on the scores together? That is, as one score changes, does the other score also change in a predictable or consistent way? This notion of variables differing together is called covariance (the prefix “co” meaning “together”).

Standard Deviation Refresher

We'll talk about standard deviations again in the denominator, but for now, let's look at the formula for standard deviation:

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

We use X to represent each score on the variable at hand, and $\overline{X}$ to represent the mean of that variable. The numerator of this formula is the Sum of Squares, which we have seen several times for various uses. Recall that squaring a value is just multiplying that value by itself. Thus, we can write the same equation as:

$$s = \sqrt{\frac{\sum((X - \overline{X})(X - \overline{X}))}{N-1}}$$

This is the same formula and works the same way as before, where we multiply the deviation score by itself (we square it) and then sum across squared deviations. You can trust me, or you can try it yourself on any data that has all scores and the calculated standard deviation provided!

Now, let’s look at the formula for covariance. In this formula, we will still use X to represent the score on one variable (but called x_{Each} to make it clear that you *subtract the mean from each number*), and we will now use y_{Each} to represent the score on the second variable. Some statisticians use $\overline{Y}$ to represent the mean of the second variable, but we've used subscripts for this whole book so we're going to keep that up. Either option will confuse a portion of you, so Dr. MO is sorry if you're the portion that is confused by $\overline{X_y}$ the mean of the "y" variable. When we start having variables, we'll use those as the subscripts again.

$$\text{COV} = \frac{\sum((x_{Each} - \overline{X_x}) \times (y_{Each} - \overline{X_y}))}{(N-1)}$$

As we can see, this is the exact same structure as the second formula presented (showing how squaring numbers is just multiplying them by themselves). Now, instead of multiplying the deviation score on one variable by itself , we take the two deviation scores from a single person on *each variable* and multiply them together. We do this for each person (exactly the same as we did for standard deviations) and then sum them to get our numerator. You can use Table $15.7.1$ to help you calculate both the standard deviations and the correlation. If the standard deviations are provided, then you can skip the columns for $(x - \overline{X_x})^2$ and $(y - \overline{X_y})^2$.

Table 15.7.1*: Sum of Products table*

Participant	X	$(X - \overline{X_x})$	$(X - \overline{X_x})^2$ (skip this column if have SD)	Y	$(Y - \overline{X - y})$	$(Y - \overline{X_y})^2$ (skip this column if have SD)	$(X - \overline{X - x})(Y - \overline{X - y})$
A							
B							
C							
and so on...							
Sum each column:	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$	$\Sigma =$

The column headers tell you exactly what to do in that column. We list our raw data for the X and Y variables in the X and Y columns, respectively, then add them up so we can calculate the mean of each variable. We then take those means and subtract them from the appropriate raw score to get our deviation scores for each person on each variable, and the columns of deviation scores will both add up to zero. We will square our deviation scores for each variable to get the sum of squares for X and Y so that we can compute the variance and standard deviation of each (we will use the standard deviation in our equation below). Finally, we take the deviation score from each variable and multiply them together to get our product score. Summing this column will give us our sum of products. It is very important that you multiply the raw deviation scores from each variable, NOT the squared deviation scores.

Let's get back to the formula to see where we're at:

$$r = \frac{\left(\frac{\sum((x_{Each} - \overline{X_x}) \times (y_{Each} - \overline{X_y}))}{(N-1)}\right)}{(s_x \times s_y)}$$

We took each score from the first variable ("x") and subtracted that variables mean from it. Then subtracted the mean of the second variable ("y") from each score of that variable. Then we multiplied them all together, and finally added up all of those products.

When we add up all of the answers from the the last column in Table 15.7.1 to calculate find the numerator of the numerator, also known as the numerator of the covariation formula ($\text{COV} = \frac{\sum((x_{Each} - \overline{X_x}) \times (y_{Each} - \overline{X_y}))}{(N-1)}$) from the table, and then we only have to divide by N–1 to get our covariance (the numerator of the correlation formula). Note that N is the number of people, so the number the pairs. It is not the number of scores. For example, if we measured 10 participants' happiness scores and their chocolate supply, we would have 10 scores for happiness and 10 scores for chocolate, and N would also be 10 (n = 10) because we had 10 people.

Unlike the sum of squares, both our sum of products and our covariance can be positive, negative, or zero, and they will always match (e.g. if our sum of products is positive, our covariance will always be positive). A positive sum of products and covariance indicates that the two variables are related and move in the same direction. That is, as one variable goes up, the other will also go up, and vice versa. A negative sum of products and covariance means that the variables are related but move in opposite directions when they change, which is called an inverse relation. In an inverse relation, as one variable goes up, the other variable goes down. If the sum of products and covariance are zero, then that means that the variables are not related. As one variable goes up or down, the other variable does not change in a consistent or predictable way.

The previous paragraph brings us to an important definition about relations between variables. What we are looking for in a relation is a consistent or predictable pattern. That is, the variables change together, either in the same direction or opposite directions, in the same way each time. It doesn't matter if this relation is positive or negative, only that it is not zero. If there is

no consistency in how the variables change within a person, then the relation is zero and does not exist. We will revisit this notion of direction vs zero relation later on.

Denominator: Standard Deviations

As you can see in the formula for Pearson's correlation:

$$r = \frac{\left(\frac{\sum((x_{Each} - \overline{X_x}) \times (y_{Each} - \overline{X_y}))}{(N-1)}\right)}{(s_x \times s_y)}$$

that the denominator is just multiplying the two standard deviations together. It looks like this:

$$r_{denominator} = \left(\sqrt{\frac{\sum (x - \overline{X_x})^2}{N-1}}\right) \times \left(\sqrt{\frac{\sum (y - \overline{X_y})^2}{N-1}}\right)$$

Easy-peasy!

Full Formula to Calculate Standard Deviations

Okay, I don't want to scare you, but I do want you to be prepared. Although, firstly, you will probably never calculate a standard deviation or a correlation by hand outside of this class. Secondly, even if your professor asks you to calculate a correlation, it is sorta unlikely that they wouldn't just give you the standard deviations. That's a lot of calculations to do by hand! But just in case, here is the very fullest of formulas for Pearson's r:

$$\frac{\left(\frac{\sum((x - \overline{X_x}) * (y - \overline{X_y}))}{(N-1)}\right)}{\left(\sqrt{\frac{\sum (x - \overline{X_x})^2}{N-1}}\right) \times \left(\sqrt{\frac{\sum (y - \overline{X_y})^2}{N-1}}\right)}$$

Next up, let's practice using these formulas!

This page titled 15.7: Correlation Formula- Covariance Divided by Variability is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

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

15.8: Practice Correlation on Anxiety and Depression

Our first example will focus on variables related to health: mental health. Our hypothesis testing procedure follows the same four-step process as before, starting with our research and null hypotheses.

Scenario

Anxiety and depression are often reported to be highly linked (or “co-morbid”). We will test whether higher scores on an anxiety scale are related to higher scales on a depression scale, lower scores on the depression scale, or not related to scores on a depression scale (meaning that the two variables do not vary together). We have scores on an anxiety scale and a depression scale for a group of 10 people. Because both scales are measured by numbers, this means that we have two different quantitative variables for the same people; perfect for a correlational statistical analysis!

Step 1: Hypotheses

For a research hypothesis for correlation analyses, we are predicting that there will be a positive linear relationship or a negative linear relationship. A positive linear relationship means that the two variables vary in the same direction (when one goes up, the other also goes up), while a negative linear relationship means that the two variables vary in opposite directions (when one goes up, the other goes down). We keep the word "linear" in the hypotheses to keep reminding ourselves that Pearson's r can only detect linear relationships. If we expect something else, then Pearson's r is not the correct analysis.

Okay, let's start with a research hypothesis. If the two variables are co-morbid, do you think that the would vary in the same direction or in opposite directions? Use that to determine your research hypothesis.

Example 15.8.1

What is a research hypothesis for this scenario?

Solution

The research hypothesis should be: There is positive linear relationship between anxiety scores and depression scores.

Look back at the research hypothesis above in Example 15.8.1. Does it include the names of both variables (what was measured)? Yes. Does it state whether the relationship will be positive or negative? Yes. Does it include the phrase "linear relationship"? Yes. Yay, we have all of the important parts of a research hypothesis! Does it say anything about means? No, which is fine; correlations aren't comparing means. Although knowing the means and standard deviations of each variable might be interesting, that descriptive information is not really necessary to test whether the two variables are linearly related.

Let's move on to the null hypothesis. What do you think that will look like?

Example 15.8.2

What is a null hypothesis for this scenario?

Solution

The null hypothesis should be: There is no linear relationship between anxiety scores and depression scores..

Notice that the variables are included again, and the word "linear".

Step 2: Collect Data

You do not need to collect this data, but you still need to understand the variables in it. Let's start with whose data we're working with and who they could represent.

Exercise 15.8.1

1. Who is the sample?
2. Who do might be the population?

Answer available at the end of this section.

Note that neither the sample nor the population mentions people with anxiety or depression. This is not a mistake! The research collected data on 10 people, but it doesn't say 10 people who were diagnosed with anxiety and/or depression. Of these 10 people, we would expect that some would score low on the anxiety scale, some would score low on the depression scale, and some would score low on both; in other words, there will be people in the sample who are not anxious or depressed.

Typically, we would now try to determine the IV and the DV. This gets a little more complicated with correlational analyses because we're not testing if one variable caused changes in the other variable. Instead, we're measuring if the variables are related; when someone scores high on one variable, can we predict their score on the other variable?

Step 3: Analyze Data

We have laid out our hypotheses and the criteria we will use to assess them, so now we can move on to our test statistic. Before we do that, we should first create a scatterplot of the data to make sure that the most likely form of our relation is in fact linear. Figure 15.8.2 below shows our data plotted out. Dr. MO is not convinced that there's a strong linear relationship, but there doesn't seem to be a strong curvilinear relationship, either. Dr. Foster thinks that the dots are, in fact, linearly related, so Pearson's r is appropriate.

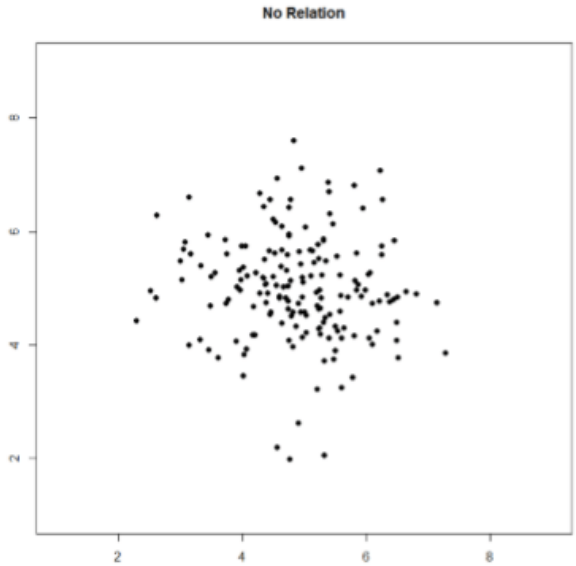


Figure 15.8.2: Scatterplot of Anxiety and Depression (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

The data we gather from our participants is shown in Table 15.8.1. The sum for each variable is also provided so that it's easy to calculate the mean.

Table 15.8.1: Depression and Anxiety Scores

Particpant	Depression Scores	Anxiety Scores
Angel	2.81	3.54
Danielle	1.96	3.05
Maritza	3.43	3.81
Christopher	3.40	3.43
Ariel	4.71	4.03
Jason	1.80	3.59
Guadalupe	4.27	4.17
Gilbert	3.68	3.46
Rosa	2.44	3.19
Mercedes	3.13	4.12
SUM:	$\sum$ = 31.63	$\sum$ = 36.39

It's easiest to calculate the formula that we're using for correlations by using a table to find the difference from the mean for each participant and multiplying them. Let's start by finding the means.

Exercise 15.8.2

What is the average depression score? What is the average anxiety score?

Answer available at the end of this section.

Now that you've done that, let's put the means and the standard deviations in Table 15.8.2. We can learn a littlebit about the two variables from this information alone. It looks like the participants might be more anxious than they are depressed. This conclusion is very tentative, however, for two reasons. First, we didn't conduct a t-test to see if the means are statistically significantly different from each other. Second, we don't actually know that the two scales have the same range! Maybe the Depression Scale ranges from 1-5, and the Anxiety Scale ranges from 0-10; we don't really have enough information to make any final determines about the means with only this table. However, it also appears that there is more variability in the Depression Scale than the Anxiety Scale, but again, we don't know that for sure unless we do some other statistical analyses.

Table 15.8.2- Descriptive Statistics for Depression and Anxiety

Scale	N	Mean	Standard Deviation
Depression Scale	10	3.16	0.94
Anxiety Scale	10	3.64	0.38

But since we're trying to conduct a correlational analysis:

$$r = \frac{\left(\frac{\sum((x_{Each} - \overline{X_x}) * (y_{Each} - \overline{X_y}))}{(N-1)}\right)}{(s_x * s_y)}$$

let's get going by filling in Table 15.8.3. Because we were provided the standard deviations, we do not need columns to square the differences from the means so those columns are not included in Table 15.8.3.

Example 15.8.3

Fill in Table 15.8.3. Keep the negative signs!

Table 15.8.3: Sum of Products Table

Depression Score	Depression Difference (Depression Score Minus Depression Mean)	Anxiety Score	Anxiety Difference (Anxiety Score Minus Anxiety Mean)	Depression Difference Times Anxiety Difference
2.81		3.54		
1.96		3.05		
3.43		3.81		
3.40		3.43		
4.71		4.03		
1.80		3.59		
4.27		4.17		
3.68		3.46		
2.44		3.19		
3.13		4.12		
$\sum$ = 31.63	$\sum$ = ?	$\sum$ = 36.39	$\sum$ = ?	$\sum$ = ?

Solution

Now it's Table 15.8.4:

Table 15.8.4: Completed Sum of Products Table

Depression Score	Depression Difference (Depression Score Minus Depression Mean)	Anxiety Score	Anxiety Difference (Anxiety Score Minus Anxiety Mean)	Depression Difference Times Anxiety Difference
2.81	-0.35	3.54	-0.10	0.04
1.96	-1.20	3.05	-0.59	0.71
3.43	0.27	3.81	0.17	0.05
3.40	0.24	3.43	-0.21	-0.05
4.71	1.55	4.03	0.39	0.60
1.80	-1.36	3.59	-0.05	0.07
4.27	1.11	4.17	0.53	0.59
3.68	0.52	3.46	-0.18	-0.09
2.44	-0.72	3.19	-0.45	0.32
3.13	-0.03	4.12	0.48	-0.01
$\sum$ = 31.63	$\sum$ = 0.03	$\sum$ = 36.39	$\sum$ = -0.01	$\sum$ = 2.22

The bottom row is the sum of each column. The difference scores for Depression sum to 0.03 and the differences scores fo Anxiety sum to -0.1; both of these are very close to 0 so, given rounding error, everything looks right so far. If you had a spreadsheet conduct all of your computations, your sums of the differences will be slightly different. However, it all sorta washes out because the final sum of the products is 2.22 no matter how many decimals you save.

Okay, let's look at the formula again to see what information we might still need:

$$r = \frac{\left(\frac{\sum((x_{Each} - \overline{X_x}) * (y_{Each} - \overline{X_y}))}{(N-1)}\right)}{(s_x \times s_y)}$$

Hey, it looks like we have all of the numbers that we need! But let's replace the "x" and "y" with D for Depression and A for Anxiety. That gives us:

$$r = \frac{\left(\frac{\sum((D_{Each} - \overline{X_D}) * (A_{Each} - \overline{X_A}))}{(N-1)}\right)}{(s_D \times s_A)}$$

Example 15.8.4

Use the information provided and that we've calculated to complete the correlational analysis:

$$r = \frac{\left(\frac{\sum((D_{Each} - \overline{X_D}) * (A_{Each} - \overline{X_A}))}{(N-1)}\right)}{(s_D \times s_A)}$$

Solution

$$r_{FilledIn} = \frac{\left(\frac{2.22}{(10-1)}\right)}{(0.94 \times 0.38)}$$

It's pretty crazy, but all of that mess at the very top of the formula (the numerator of the numerator) is what we did in the table, so it boils down to 2.22. After that, it's pretty easy!

$$r_{Parentheses} = \frac{\left(\frac{2.22}{9}\right)}{0.36}$$

$$r_{Divide} = \frac{0.25}{0.36}$$

$$r = 0.69$$

.

If you use a spreadsheet to keep all of the decimal points, you might end up with r=0.68 or if you round differently you could get r=0.70. Since we're learning the process right now, these are all close enough.

So our calculated correlation between anxiety and depression is $r = 0.69$, which is a strong, positive correlation. Now we need to compare it to our critical value to see if it is also statistically significant.

Step 4: Conclusion

To conclude whether the research hypothesis is supported or not, we need to find the critical value, then compare that to our calculated value to see if we can reject the null hypothesis or not. If the null hypothesis is retained, then we did not support our

research hypothesis because we're saying that the variables are not linearly related. If the null hypothesis is rejected, we may still not support the research hypothesis! We only support a directional research hypothesis if we reject the null hypothesis (that there is no linear relationship) and we find that linear relationship is in the direction that we hypothesized (positive or negative).

To start, let's find the critical values! The critical values for correlations come from the Table of Critical Values of r, found in the back of this book in the Common Critical Value Tables sections. For correlations, we have $N-2$ degrees of freedom, rather than $N-1$. For our example, we have 10 people.

We were not given any information about the level of significance at which we should test our hypothesis, so we will assume $\alpha = 0.05$. From the table, we can see that at p = .05 level, the critical value of $r_{Critical}$ = 0.576. Thus, if our calculated correlation is greater than 0.576, it will be statistically significant and we would reject the null hypothesis. This is a rather high bar (remember, the guideline for a strong relation is r = 0.50); this is because we have so few people. Larger samples make it easier to find significant relations.

Our calculated value was \(r_{Calc} = 0.69\), so our calculated value was larger than our critical value, so we can reject the null hypothesis because this is still true:

$$\textbf{(Critical} < \textbf{Calculated)} = \textbf{Reject null} = \textbf{There is a linear relationship.} \ = p < .05$$

$$\textbf{(Critical} > \textbf{Calculated)} = \textbf{Retain null} = \textbf{There is not a linear relationship.} \ = p > .05$$

The statistical sentence is similar to what we've been doing, so let's try that now:

Example 15.8.5

What is the statistical sentence for the results for our correlational analysis?

Solution

The statistical sentence is: r(8)=0.69, p<.05

Because the Degrees of Freedom are N-2 for Pearson's r.

Step 5: Write-Up

Okay, we're ready for the final write-up! Because our r = 0.69, a positive number, we can say that the linear relationship is positive.

Example 15.8.6

Write up your conclusion. Don't forget to include the requirements for reporting results!

Solution

"We hypothesized that there would be a positive linear relationship between depression scores (M=3.16) and anxiety scores (M=3.64). This hypothesis was supported (r(8)=0.69, p < .05). As depression scores increase, anxiety scores also increase. With our sample, depression and anxiety do appear to be co-morbid."

As we will discover in the next chapter, we can also add a sentence about predictions: "Based on these results, we can use a participant's depression score to predict their anxiety score."

Even though we are dealing with a very different type of data, our process of hypothesis testing has remained unchanged.

Let's try that again, but use the full Sum of Products table to calculate the standard deviations, too.

Answers

Exercise 15.8.1

1. The sample is 10 participants.
2. The population could be all people.

Exercise 15.8.2

$$\overline{X_D} = \frac{\sum X}{N} = \frac{31.63}{10} = 3.163 = 3.16$$

$$\overline{X_A} = \frac{\sum X}{N} = \frac{36.39}{10} = 3.639 = 3.64$$

This page titled 15.8: Practice Correlation on Anxiety and Depression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 12.5: Anxiety and Depression by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

15.9: Practice Correlation on Nutrition

This second example will use the same formula, more or less. Instead of providing the standard deviations, you will need to calculate them. This example uses nutrition data from fast food restaurants from www.OpenIntro.org. You should probably try the previous practice problem because it gives more hints than this one will.

Scenario

When Dr. MO was on a low-carb diet, she noticed that things that were low in carbohydrates tended to be high in fat, and vice versa. Now that she has the skills and the data set, we can test to see if there is a statistical relationship! Our analysis will look at the chicken dishes from one fast food restaurant using the nutritional data from OpenIntro.org.

Step 1: Hypotheses

What kind of linear relationship is it if it looks like as fat goes up, carbs go down?

Exercise 15.9.1

What is a research hypothesis for this scenario if Dr. MO expects that as one goes up, the other will go down?

Answer available at the end of this section.

Sometimes, people will add a description of what the negative linear relationship would look like to their research hypothesis: There is negative linear relationship between fat and carbohydrates in the sample of chicken dishes from one fast food restaurant such that foods with more fat tend to have fewer carbohydrates.

Let's look at the scatterplot to see if it looks like a negative linear relationships (as one goes up, the other goes down), which should have a general linear trend from the top left to the bottom right.

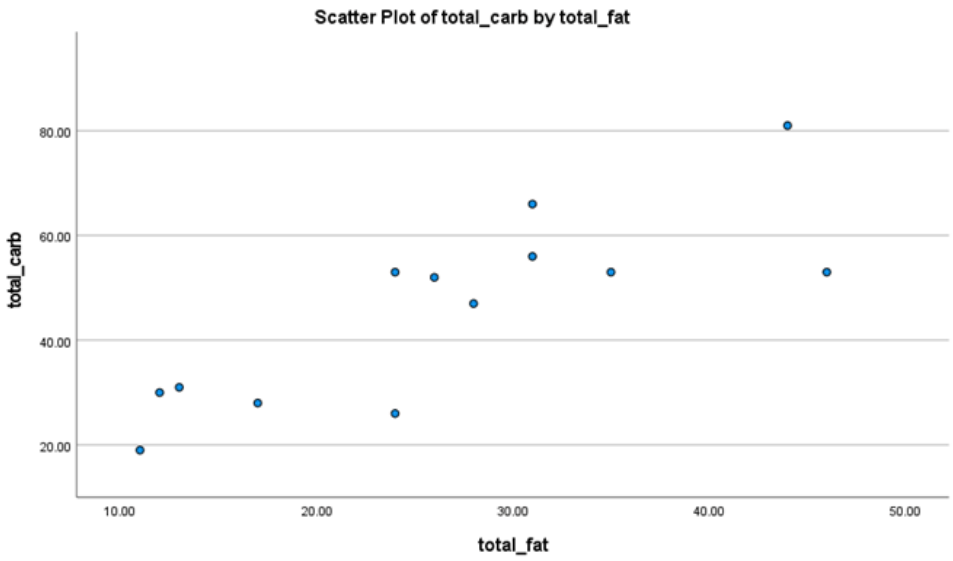


Figure 15.9.1: Scatterplot of Carbohydrates by Fat for 13 Chicken Meals. (CC-BY; Michelle Oja via OpenIntro.org)

Uh-oh, already it looks like Dr. MO's research hypothesis is not working out. There does seem to be a general linear trend, but it does not look like a negative correlation; it actually looks like a positive correlation!

Let's state our null hypothesis so that we can get to the analysis to see what's going on statistically.

Exercise 15.9.2

What is a null hypothesis for this scenario?

Answer available at the end of this section.

Step 2: Collect Data

Before we get too far into the weeds, let's consider why a correlational analysis is the appropriate analysis. Your first clue is the research hypothesis. We are not comparing whether the average fat content is higher than the average carbohydrate content; in other words, we aren't looking at mean differences. Another clue is that we have means of two different variables, not means of two different groups. So that leads to recognizing that we have two quantitative variables, not a quantitative variable that we measured for two or more groups (qualitative variables). Whenever you're not sure, check out the Decision Tree handout in the first chapter.

To understand the scenario, let's identify what those quantitative variable are, as well as the sample and population.

Exercise 15.9.3

1. Who is the sample?
2. Who do might be the population?
3. What two different (quantitative) variables were measured?

Answer available at the end of this section.

Step 3: Analyze Data

It's easiest to calculate the standard deviation and Pearson's r by using a table to do most of the calculations. If we sum each column in the table, we also can easily find the mean.

Table 15.9.1- Raw Data for Sum of Products Table for Fat and Carbs

Item	Total Fat	Fat Difference: Fat Minus Mean	Squared	Total Carbs	Carb Differences: Carb Minus Mean	Squared	Fat Diff * Carb Diff
2 piece Prime-Cut Chicken Tenders	11			19			
Chicken Tender 'n Cheese Slider	12			30			
Buffalo Chicken Slider	13			31			
3 piece Prime-Cut Chicken Tenders	17			28			
Buttermilk Buffalo Chicken Sandwich	24			53			
Crispy Chicken Farmhouse Salad	24			26			
Buttermilk Crispy Chicken Sandwich	26			52			

Item	Total Fat	Fat Difference: Fat Minus Mean	Squared	Total Carbs	Carb Differences: Carb Minus Mean	Squared	Fat Diff * Carb Diff
5 piece Prime-Cut Chicken Tenders	28			47			
Bourbon BBQ Chicken Sandwich	31			66			
Buttermilk Chicken Bacon & Swiss	31			56			
Buttermilk Chicken Cordon Bleu Sandwich	35			53			
Pecan Chicken Salad Sandwich	44			81			
Pecan Chicken Salad Flatbread	46			53			
SUM:	$\sum$ = **342**			$\sum$ **=595**			

Let's go through this table. The Item (first column) is what kind of chicken dish we are talking about. Although it's not necessary to include this information, Dr. MO likes to show that the numbers in the second column (Total Fat) are connected to the numbers in the Total Carbs column because they are associated with the same dish. The data area ordered so that it goes from items with the least fat to items with the most fat. That means that the Total Carbs cannot ASLO be ordered from least to most carbs because they would lose their connection to the meal that they are associated with. We can put the data in any order that we'd like, but, for example, the Prime-Cut Chicken Tenders must always have Total Fat of 11 and Total Carbs of 19.

Okay, so we talked about the two columns that have data in them now. The empty columns are what we need to fill with our calculations to finish both the formula for standard deviations:

$$s = \sqrt{\frac{\sum\left((X-\overline{X})^2\right)}{N-1}}$$

and the formula for correlations:

$$r = \frac{\left(\frac{\sum((x-\overline{X_x})\times(y-\overline{X_y}))}{(N-1)}\right)}{\left(\sqrt{\frac{\sum\left((x-\overline{X_x})^2\right)}{N-1}}\right)\times\left(\sqrt{\frac{\sum\left((y-\overline{X_y})^2\right)}{N-1}}\right)}$$

First things first, though. If you look at the table, the first empty columns says "Fat Difference: Fat Minus Mean". This means that we must find the mean for the Total Fat variable, then subtract that number from each of the scores for fat.

Exercise 15.9.4

What is the average amount of fat for chicken dishes from this fast-food restaurant? What is the average amount of carbohydrates for chicken dishes from this fast-food restaurant?

Answer available at the end of this section.

Hopefully these data makes it a little more clear why we are more interested in finding a linear relationship between these two variables, rather than seeing if the means are different. We could conduct a t-test to see if the means are different, but what question would that answer? Dr. MO's prediction was that items with more carbs had less fat, so finding that chicken dishes have more carbohydrates than fat doesn't help us answer that question at all.

Now that we have the means for our two quantitative variables, we can start filling in the table.

Example 15.9.1

Fill in Table 15.9.1 with the differences for each variable, square those differences, and multiply those differences (not the squares, though). At the end, you should have each column summed.

Solution

Table 15.9.2- Raw Data with Completed Sum of Products Table for Fat and Carbs

Item	Total Fat	Fat Difference: Fat Minus Mean	Squared	Total Carbs	Carb Differences: Carb Minus Mean	Squared	Fat Diff * Carb Diff
2 piece Prime-Cut Chicken Tenders	11	-15.31	234.40	19	-26.77	716.63	409.85
Chicken Tender 'n Cheese Slider	12	-14.31	204.78	30	-15.77	248.69	225.67
Buffalo Chicken Slider	13	-13.31	177.16	31	-14.77	218.15	196.59
3 piece Prime-Cut Chicken Tenders	17	-9.31	86.68	28	-17.77	315.77	165.44
Buttermilk Buffalo Chicken Sandwich	24	-2.31	5.34	53	7.23	52.27	-16.70
Crispy Chicken Farmhouse Salad	24	-2.31	5.34	26	-19.77	390.85	45.67
Buttermilk Crispy Chicken Sandwich	26	-0.31	0.10	52	6.23	38.81	-1.93
5 piece Prime-Cut Chicken Tenders	28	1.69	2.86	47	1.23	1.51	2.08
Bourbon BBQ Chicken Sandwich	31	4.69	22.00	66	20.23	409.25	94.88
Buttermilk Chicken Bacon & Swiss	31	4.69	22.00	56	10.23	104.65	47.98
Buttermilk Chicken Cordon Bleu Sandwich	35	8.69	75.52	53	7.23	52.27	62.83
Pecan Chicken Salad Sandwich	44	17.69	312.94	81	35.23	1241.15	623.22
Pecan Chicken Salad Flatbread	46	19.69	387.70	53	7.23	52.27	142.36
SUM:	$\sum$ = **342**	$\sum$ = **-0.03**	$\sum$ = **1536.77**	$\sum$ = **595**	$\sum$ = **-0.01**	$\sum$ = **3842.31**	$\sum$ = **1997.92**

What you should see when you look at the table is that the sum of the differences from the mean are really close to zero, which means that we probably the calculations correctly.

From here, there are two different options that you can take. One option is to calculate the standard deviations separately using the same formula that we've used before:

$$s = \sqrt{\frac{\sum\left((X - \overline{X})^2\right)}{N-1}}$$

Then add those standard deviation results into the formula that we used in our mental health example.

$$r = \frac{\left(\frac{\sum((x_{Each} - \overline{X_x}) \times (y_{Each} - \overline{X_y}))}{(N-1)}\right)}{(s_x \times s_y)}$$

It is totally fine if you choose this option. At this point, it's whatever you feel more comfortable with. If you choose this option, meet us at the end to make sure that you got the same answer!

What the example is going to do is the second option, and that is to fill in this crazy formula with all of the numbers from the table:

$$r = \frac{\left(\frac{\sum((x - \overline{X_x}) \times (y - \overline{X_y}))}{(N-1)}\right)}{\left(\sqrt{\frac{\sum\left((x - \overline{X_x})^2\right)}{N-1}}\right) \times \left(\sqrt{\frac{\sum\left((y - \overline{X_y})^2\right)}{N-1}}\right)}$$

Hopefully you can see that they are the same formula, but replacing the standard deviation (s) of each variable with the formula for standard deviations.

This is going to be a doozy!

Example 15.9.2

Fill in the sums from Table 2 to complete the Pearson's correlation formula:

$$r = \frac{\left(\frac{\sum((x - \overline{X_x}) * (y - \overline{X_y}))}{(N-1)}\right)}{\left(\sqrt{\frac{\sum\left((x - \overline{X_x})^2\right)}{N-1}}\right) \times \left(\sqrt{\frac{\sum\left((y - \overline{X_y})^2\right)}{N-1}}\right)}$$

Solution

First, Dr. MO likes to replace the "x" and "y" with F for Fat and C for Carbs. That helps her follow the formula better. That gives us:

$$r = \frac{\left(\frac{\sum((X_F - \overline{X_F}) * (X_C - \overline{X_C}))}{(N-1)}\right)}{\left(\sqrt{\frac{\sum\left((X_F - \overline{X_F})^2\right)}{N-1}}\right) \times \left(\sqrt{\frac{\sum\left((X_C - \overline{X_C})^2\right)}{N-1}}\right)}$$

As long as you follow the order of operations, you can fill in the formula however you'd like. You can start with the numerator, and ignore the denominator for a bit, or fill in the two standard deviations first, then fill out the numerator. Again, it's whatever makes the most sense to you. Dr. MO likes to plug in all of the numbers at once, then do sets of parentheses to start making the formula smaller. So here it is with all of the numbers:

$$r = \frac{\left(\frac{1997.92}{(13-1)}\right)}{\left(\sqrt{\frac{1539.77}{13-1}}\right) \times \left(\sqrt{\frac{3842.31}{13-1}}\right)}$$

Remember for standard deviations that you square each difference score, then add up all of the squares (sum of squares!). We did that in the table, so we just plug those numbers in. We also already got the entire top part of the numerator from the table.

Next step is to start calculating what's in the parentheses. This first set will be the N-1's, then the first set of division.

$$r_{N-1's} = \frac{\left(\frac{1997.92}{12}\right)}{\left(\sqrt{\frac{1539.77}{12}}\right) \times \left(\sqrt{\frac{3842.31}{12}}\right)}$$

$$r_{Division} = \frac{166.49}{\left(\sqrt{128.06}\right) \times \left(\sqrt{320.19}\right)}$$

A couple more steps to get down to one number!

$$r_{SquareRoot} = \frac{166.49}{(11.32) \times (17.89)}$$

$$r_{Multiply} = \frac{166.49}{(202.49)}$$

$$r = \frac{166.49}{202.49} = 0.82$$

So our calculated correlation between fat and cars is $r = 0.82$, which is a strong, positive correlation. Now we need to compare it to our critical value to see if it is also statistically significant.

Step 4: Conclusion

As always, to decide if our data can support our research hypothesis, we first have to decide if the null hypothesis is retained or rejected. To do that, we compare our calculated r-value (that you should just found in Step 3) to a critical r that you find in the Table of Critical Values of r found in the back of this textbook. We were not given any information about the level of significance at which we should test our hypothesis, so we will assume $\alpha = 0.05$ as always. From the page, we can see that we need the Degrees of Freedom, which is N–2. The data is provided shows that we have 13 chicken dishes that we are looking at. With p=.05 level and N –2 (13–2=11) , the critical r-value is 0. 0.553.

Our critical value was is $r_{Critical} = 0.0.553$ and our calculated value was $r_{Calc} = 0.82$. Our calculated value was larger than our critical value, so we can reject the null hypothesis because this is still true:

$$\textbf{(Critical} < \textbf{Calculated)} = \textbf{Reject null} = \textbf{There is a linear relationship.} \ = p < .05$$

$$\textbf{(Critical} > \textbf{Calculated)} = \textbf{Retain null} = \textbf{There is not a linear relationship.} \ = p > .05$$

What does the statistical sentence look like?

Exercise 15.9.5

What is the statistical sentence for the results for our correlational analysis?

Answer available at the end of this section.

This tells us that we are rejecting the null hypothesis. But are we supporting the research hypothesis? Our research hypothesis stated: There is negative linear relationship between fat and carbohydrates in the sample of chicken dishes from one fast food restaurant such that foods with more fat tend to have fewer carbohydrates. To support this research hypothesis, then we'd not only need to reject the null hypothesis (which we did because p<.05), but the calculated Pearson's r should be a negative number (since we predicted a negative linear relationship). As our calculate r is a positive number (r=0.82), we do not support our research

hypothesis. Dr. MO predicted that when the amount of fat went up in a meal, the amount of carbohydrates would go down, but she instead found a strong positive correlation! As the amount of fast goes up in a meal, the amount of carbohydrates also increased.

Step 5: Write-Up

Okay, we're ready for the final write-up! Because our r=0.82, a positive number, the linear relationship is also positive. (Hint: This is not what Dr. MO hypothesized, but it does match the scatter plot!)

Example 15.9.3

Write up your conclusion. The expected information in a write-up would be:

- Research Hypothesis, including:
  - Re-state the RH (including the groups being compared, the DV, and the direction word), and
  - If the RH was fully supported or not supported.
- Calculation results:
  - Calculated statistical (statistical sentence): r(df) = calculated r, p <.05 [or p>.05]
  - Means, with units, of each group (name the groups)

You can include the means to be consistent, but they are less relevant when we are testing for linear relationships.

Solution

"We hypothesized that there would be a negative linear relationship between fat scores (M=26.31) and carbs scores (M=45.77) such that the chicken meals with more fat would have less carbs. Although there is a strong correlation between fat and carbs (r(11)=0.82, p<.05), the relationship is the opposite of the research hypothesis; chicken meals with more fat also tended to have more carbs."

And that's how you calculate and interpret a Pearson's correlation when you're given the raw data!

Next, we will learn about some options when we don't think that our distributions are normally distributed.

Contributors and Attributions

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

Answers

Exercise 15.9.1

The research hypothesis should be: There is negative linear relationship between fat and carbohydrates in the sample of chicken dishes from one fast food restaurant.

Exercise 15.9.2

The null hypothesis should be: There is no linear relationship between fat and carbohydrates in the sample of chicken dishes from one fast food restaurant.

Exercise 15.9.3

1. The sample is 13 chicken meals from one fast food chain.
2. The population could be all fast food chicken meals, or maybe all fast food chicken meals from that one fast food chain.
3. The two quantitative variables are the amount of fat in a chicken meal, and the amount of carbohydrates in that same meal.

Exercise 15.9.4

$$\overline{X_F} = \frac{\sum X}{N} = \frac{342}{13} = 26.31$$

$$\overline{X_C} = \frac{\sum X}{N} = \frac{595}{13} = 45.77$$

Exercise 15.9.5

The statistical sentence is: r(11)=0.82, p<.05

This page titled 15.9: Practice Correlation on Nutrition is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

15.10: Alternatives to Pearson's Correlation

The Pearson correlation coefficient is useful for a lot of things, but it does have shortcomings. One issue in particular stands out: what it actually measures is the strength of the *linear* relationship between two variables. In other words, what it gives you is a measure of the extent to which the data all tend to fall on a single, perfectly straight line. Often, this is a pretty good approximation to what we mean when we say “relationship”, and so the Pearson correlation is a good thing to calculation. Sometimes, it isn’t. In this section, Dr. Navarro and Dr. MO will be reviewing the correlational analysis you could use if you're a linear relationship isn't quite what you are looking for.

Spearman’s Rank Correlations

One very common situation where the Pearson correlation isn’t quite the right thing to use arises when an increase in one variable X really is reflected in an increase in another variable Y, but the nature of the relationship is more than linear (a straight line). An example of this might be the relationship between effort and reward when studying for an exam. If you put in zero effort (x-axis) into learning a subject, then you should expect a grade of 0% (y-axis). However, a little bit of effort will cause a *massive* improvement: just turning up to lectures means that you learn a fair bit, and if you just turn up to classes, and scribble a few things down so your grade might rise to 35%, all without a lot of effort. However, you just don’t get the same effect at the other end of the scale. As everyone knows, it takes *a lot* more effort to get a grade of 90% than it takes to get a grade of 55%. What this means is that, if I’ve got data looking at study effort and grades, there’s a pretty good chance that Pearson correlations will be misleading.

To illustrate, consider the data in Table 15.10.1, plotted in Figure 15.10.1, showing the relationship between hours worked and grade received for 10 students taking some class.

Table 15.10.1- *Hours Worked and Percentage Grade*

Student	Hours Worked	Percentage Grade
A	2	13
B	76	91
C	40	79
D	6	14
E	16	21
F	28	74
G	27	47
H	59	85
I	46	84
J	68	88

The curious thing about this highly fictitious data set is that increasing your effort *always* increases your grade. This produces a strong Pearson correlation of r=.91; the dashed line through the middle shows this linear relationship between the two variables. However, the interesting thing to note here is that there’s actually a perfect monotonic relationship between the two variables: in this example at least, increasing the hours worked always increases the grade received, as illustrated by the solid line.

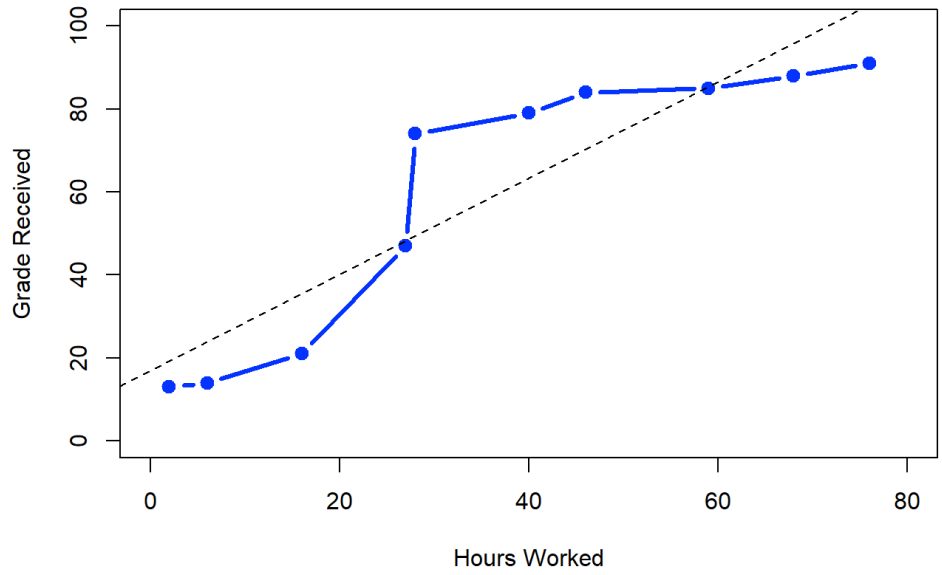


Figure 15.10.1-The relationship between hours worked and grade received, for a data set consisting of only 10 students (each circle corresponds to one student). The dashed line through the middle shows the linear relationship between the two variables. (CC-BY-SA- Danielle Navarro from Learning Statistics with R)

If we run a standard Pearson correlation, it shows a strong relationship between hours worked and grade received ($r(8) = 0.91$, $p < .05$), but this doesn't actually capture the observation that increasing hours worked *always* increases the grade. There's a sense here in which we want to be able to say that the correlation is *perfect* but for a somewhat different notion of what a "relationship" is. What we're looking for is something that captures the fact that there is a perfect ***ordinal relationship*** here. That is, if Student A works more hours than Student B, then we can *guarantee* that Student A will get the better grade than Student b. That's not what a correlation of r=.91 says at all; Pearson's r says that there's strong, positive linear relationship; as one variable goes up, the other variable goes up. It doesn't say anything about how much each variable goes up.

How should we address this? Actually, it's really easy: if we're looking for ordinal relationships, all we have to do is treat the data as if it were ordinal scale! So, instead of measuring effort in terms of "hours worked", lets *rank* all 10 of our students in order of hours worked. That is, the student who did the least work out of anyone (2 hours) so they get the lowest rank (rank = 1). The student who was the next most distracted, putting in only 6 hours of work in over the whole semester get the next lowest rank (rank = 2). Notice that Dr. Navarro is using "rank =1" to mean "low rank". Sometimes in everyday language we talk about "rank = 1" to mean "top rank" rather than "bottom rank". So be careful: you can rank "from smallest value to largest value" (i.e., small equals rank 1) or you can rank "from largest value to smallest value" (i.e., large equals rank 1). In this case, I'm ranking from smallest to largest, because that's the default way that some statistical programs run things. But in real life, it's really easy to forget which way you set things up, so you have to put a bit of effort into remembering!

Okay, so let's have a look at our students when we rank them from worst to best in terms of effort and reward:

Table 15.10.1- Ranking of Students by Hours Worked and Grade Percentage

Student	Hours Worked	Rank of Hours Worked	Percentage Grade	Rank of Percentage Grade
A	2	1	13	1
D	6	2	14	2
E	16	3	21	3
G	27	4	47	4
F	28	5	74	5
C	40	6	79	6
I	46	7	84	7
H	59	8	85	8

Student	Hours Worked	Rank of Hours Worked	Percentage Grade	Rank of Percentage Grade
J	68	9	88	9
B	76	10	91	10

Hm. The rankings are *identical*. The student who put in the most effort got the best grade, the student with the least effort got the worst grade, etc. If we run a Pearson's correlation on the rankings, we get a perfect relationship: r(8) = 1.00, p<.05. What we've just re-invented is *Spearman's rank order correlation*, usually denoted ρ or rho to distinguish it from the Pearson r. If we analyzed this data, we'd get a Spearman correlation of rho=1. We aren't going to get into the formulas for this one; if you have ranked or ordinal data, but you can find the formulas online or use statistical software.

For this data set, which analysis should you run? With such a small data set, it's an open question as to which version better describes the actual relationship involved. Is it linear? Is it ordinal? We're not sure, but we can tell that increasing effort will never *decrease* your grade.

Curvilinear Relationships

Another situation in which Pearson's correlation isn't quite right is when we expect a non-linear relationship. We've seen this in scatter plots that are U-shaped or reverse-U shaped (the section of the graphing chapter on scatterplots, the section of this chapter on correlation versus causation in graphs, and the section in this chapter on strength, direction, and linearity) . In that section on correlation versus causation in graphs, we talked about how plants die if they don't have enough water but also if they have too much water. There is not a linear relationship between water and plant growth, it is a curvilinear relationship. There are specific statistical analyses that look for specific curvilinear relationships (based on how curved it appears to be) that we aren't going to get into here, but they are available on advanced statistical software programs.

Summary

The key thing to remember, other than that these analyses exist, is that looking at the graph is key to helping you know what kind of analysis to conduct.

Contributors and Attributions

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

This page titled 15.10: Alternatives to Pearson's Correlation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

15.11: Final Considerations

Correlations, although simple to calculate, and be very complex, and there are many additional issues we should consider. We will look at two of the most common issues that affect our correlations, as well as discuss reporting methods you may encounter.

Range Restriction

The strength of a correlation depends on how much variability is in each of the variables. This is evident in the formula for Pearson's r, which uses both covariance (based on the sum of products, which comes from deviation scores) and the standard deviation of both variables (which are based on the sums of squares, which also come from deviation scores). Thus, if we reduce the amount of variability in one or both variables, our correlation will go down (less variability = weaker correlation). Range restriction is one case when the full variability of a variable is missed.

Take a look at Figures $15.11.1$ and $15.11.2$ below. The first shows a strong relation (r = 0.67) between two variables. An oval is overlain on top of it to make the relation even more distinct. The second figure shows the same data, but the bottom half of the X variable (all scores below 5) have been removed, which causes our relation (again represented by a red oval) to become much weaker (r = 0.38). Thus range restriction has truncated (made smaller) our observed correlation.

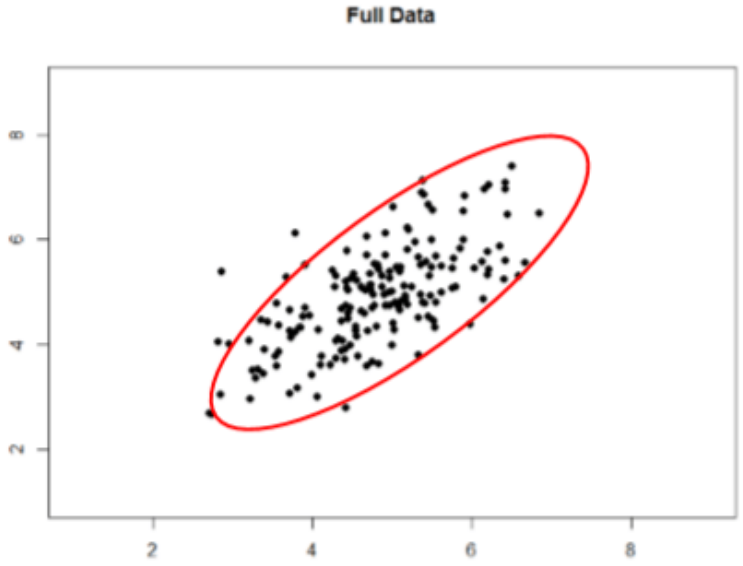


Figure $15.11.1$: Strong, positive correlation. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

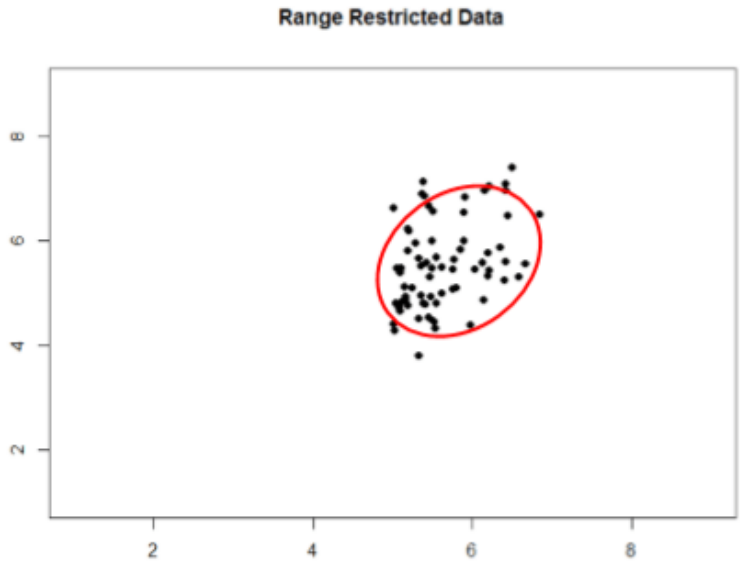


Figure 15.11.2: Effect of range restriction. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Sometimes range restriction happens by design. For example, we rarely hire people who do poorly on job applications, so we would not have the lower range of those predictor variables. Dr. MO's students researchers have asked for the ages of participants but offer a "30 years old or older" category or "under 18 years old" category. This restricts what the range of ages looks like because we have individual data points for everyone who is 18 to 29 years old, but then a clump of data points for those younger or older than that. Other times, we inadvertently cause range restriction by not properly sampling our population. Although there are ways to correct for range restriction, they are complicated and require much information that may not be known, so it is best to be very careful during the data collection process to avoid it.

Outliers

Another issue that can cause the observed size of our correlation to be inappropriately large or small is the presence of outliers. An outlier is a data point that falls far away from the rest of the observations in the dataset. Sometimes outliers are the result of incorrect data entry, poor or intentionally misleading responses, or simple random chance. Other times, however, they represent real people with meaningful values on our variables. The distinction between meaningful and accidental outliers is a difficult one that is based on the expert judgment of the researcher. That's you! When you collect and analyze your down data, you are the researcher and the most expert person on your own data.

When might you remove an outlier? Sometimes, we will remove the outlier (if we think it is an accident) or we may decide to keep it (if we find the scores to still be meaningful even though they are different). In one of her research studies, Dr. MO asked participants to read a brief story online (the IV was in the story), then click to the next page to answer questions on the story (DV). If participants clicked through the story in milliseconds, then Dr. MO removed their scores on the dependent variable from further analysis because they didn't actually experience the IV without actually reading the story.

The plots below in Figure 15.11.3 show the effects that an outlier can have on data. In the first, we have our raw dataset. You can see in the upper right corner that there is an outlier observation that is very far from the rest of our observations on both the X and Y variables. In the middle, we see the correlation computed when we include the outlier, along with a straight line representing the relation; here, it is a positive relation. In the third image, we see the correlation after removing the outlier, along with a line showing the direction once again. Not only did the correlation get stronger, it completely changed direction!

In general, there are three effects that an outlier can have on a correlation: it can change the magnitude (make it stronger or weaker), it can change the significance (make a non-significant correlation significant or vice versa), and/or it can change the direction (make a positive relation negative or vice versa). Outliers are a big issue in small datasets where a single observation can have a strong weight compared to the rest. However, as our samples sizes get very large (into the hundreds), the effects of outliers diminishes because they are outweighed by the rest of the data. Nevertheless, no matter how large a dataset you have, it is always a good idea to screen for outliers, both statistically (using analyses that we do not cover here) and/or visually (using scatterplots).

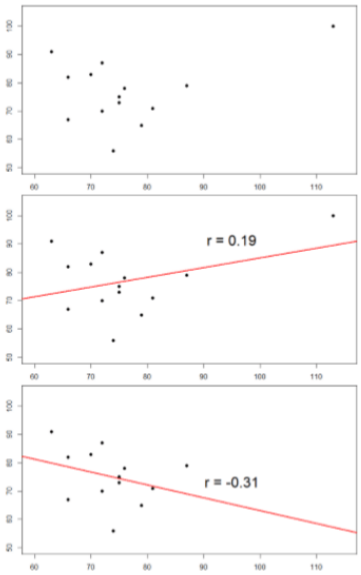


Figure 15.11.3: Three plots showing correlations with and without outliers. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

Correlation Matrices

Many research studies look at the relation between more than two quantitative variables. In such situations, we could simply list all of our correlations, but that would take up a lot of space and make it difficult to quickly find the relation we are looking for. Instead, we create correlation matrices so that we can quickly and simply display our results. A matrix is like a grid that contains our values. There is one row and one column for each of our variables, and the intersections of the rows and columns for different variables contain the correlation for those two variables.

We can create a correlation matrix to quickly display the correlations between job satisfaction, well-being, burnout, and job performance. each. Such a matrix is shown below in Table 15.11.1.

Table 15.11.1: Correlation matrix to display the numerical values

Measurement	Satisfaction	Well-Being	Burnout	Performance
Satisfaction	1.00			
Well-Being	0.41	1.00		
Burnout	-0.54	-0.87	1.00	
Performance	0.08	0.21	-0.33	1.00

Notice that there are values of 1.00 where each row and column of the same variable intersect. This is because a variable correlates perfectly with itself, so the value is always exactly 1.00. Also notice that the upper cells are left blank and only the cells below the diagonal of 1s are filled in. This is because correlation matrices are symmetrical: they have the same values above the diagonal as below it. Filling in both sides would provide redundant information and make it a bit harder to read the matrix, so we leave the upper triangle blank.

Correlation matrices are a very condensed way of presenting many results quickly, so they appear in almost all research studies that correkate several quantitative variables. Many matrices also include columns that show the variable means and standard deviations, as well as asterisks showing whether or not each correlation is statistically significant.

Summary

The principles of correlations underlie many other advanced analyses. In the next chapter, we will learn about regression, which is a formal way of running and analyzing a correlation that can be extended to more than two variables. Regression is a very powerful

technique that serves as the basis for even our most advanced statistical models, so what we have learned in this chapter will open the door to an entire world of possibilities in data analysis.

This page titled 15.11: Final Considerations is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 12.8: Final Considerations by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

16: REGRESSION

16.1: Introduction- Line of Best Fit
16.2: Regression Line Equation
16.3: Using Linear Equations
16.4: Hypothesis Testing- Slope to ANOVAs
16.5: Practice Regression on Health and Happiness
16.6: Practice Regression on Nutrition
16.7: Multiple Regression

This page titled 16: Regression is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

16.1: Introduction- Line of Best Fit

In correlations, we referred to a linear trend in the data. That is, we assumed that there was a straight line we could draw through the middle of our scatterplot that would represent the relation between our two variables. Regression involves an equation of that line, which is called the Line of Best Fit.

The line of best fit can be thought of as the central tendency of our scatterplot. The term “best fit” means that the line is as close to all points (with each point representing both variables for a single person) in the scatterplot as possible, with a balance of scores above and below the line. This is the same idea as the mean, which has an equal weighting of scores above and below it and is the best singular descriptor of all our data points for a single variable.

We have already seen many scatterplots in the chapter on graphs and the chapter on correlations, so we know that the dots on a scatterplot never form a perfectly straight line. Because of this, when we plot a straight line through a scatterplot, it will not touch all of the points, and it may not even touch any! This will result in some distance between the line and each of the points it is supposed to represent, just like a mean has some distance between it and all of the individual scores in the dataset.

The distances between the line of best fit and each individual data point go by two different names that mean the same thing: errors or residuals. The term “error” in regression is closely aligned with the meaning of error in ANOVAs (standard error); it does not mean that we did anything wrong! In statistics, "error" means that there was some discrepancy or difference between what our analysis produced and the true value we are trying to get at it. The term “residual” is new to our study of statistics, and it takes on a very similar meaning in regression to what it means in everyday parlance: there is something left over. In regression, what is “left over” – that is, what makes up the residual – is an imperfection in our ability to predict values of the Y variable using our line. This definition brings us to one of the primary purposes of regression and the line of best fit: predicting scores.

Predicting Y from X

If we know that there is a linear relationship between two variables, we can use one variable to predict the other. We use this regression line to “predict” one score/variable from one other score/variable. Remember, we can't say that one variable causes the other variable to change (see Correlation versus Causation)! We are merely saying that because of the statistically significant correlation, we can use a simple equation to use one variable to predict the other.

The most common reasons to predict scores:

- Time: You want an estimate of an event that hasn’t happened yet.
  - You base the prediction on a variable that is available now.
- Expense: You want an estimate of an expensive measure.
  - You base the prediction on a variable that is cheaper.

An example related to time is the SAT. SAT scores were designed to predict successfully finishing the first year of college. A bunch of students were given a bunch of questions, then followed for a year to see who were able to pass their first year of college. The questions that were correlated with finishing the first year of college were refined to create the SAT, then used for future students to predict finishing the first year of college.

An example related to expense are the personality tests that you might complete for certain jobs. High-quality personality assessments are expensive; most companies can't pay that kind of money for each and every applicant. However, what they can do is find a cheaper personality test that is statistically significantly correlated with the expensive test. That way, the company can use the cheaper test to predict scores on the high-quality (but expensive) test. Sure, it's not as good of an assessment of personality and fit with the company or job, but the company can afford to test lots of candidates!

This page titled 16.1: Introduction- Line of Best Fit is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 13.1: Line of Best Fit by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

16.2: Regression Line Equation

The goal of regression is the same as the goal of ANOVA: to take what we know about one variable (X) and use it to explain our observed differences in another variable (Y). In ANOVA, we talked about – and tested for – group mean differences, but in regression we do not have groups for our explanatory variable; we have a continuous variable, like in correlation. Because of this, our vocabulary will be a little bit different, but the process, logic, and end result are all the same.

Regression Equation

In regression, we most frequently talk about prediction, specifically predicting our outcome variable Y from our explanatory variable X, and we use the line of best fit to make our predictions. Let's take a look at the equation for the line, which is quite simple:

$$\widehat{\mathrm{Y}} = \mathrm{a} + \mathrm{bX}$$

In which bX mean beta times X.

The terms in the equation are defined as:

- $\widehat{\mathrm{Y}}$: the predicted value of Y for an individual person
- a: the intercept of the line
- b: the slope of the line
- X: the observed value of X for an individual person

What this shows us is that we will use our known value of X for each person to predict the value of Y for that person. The predicted value, $\widehat{\mathrm{Y}}$, is called "y-hat" and is our best guess for what a person's score on the outcome is. The equation has only two parameter estimates: an intercept (where the line crosses the Y-axis) and a slope (how steep – and the direction, positive or negative – the line is). I bet when you took geometry in high school, you never thought that you'd use that again! But here we are, talking about slopes. The intercept and slope are parameter estimates because, like everything else in statistics, we are interested in approximating the true value of the relation in the population but can only ever estimate it using sample data. We will soon see that one of these parameters, the slope, is the focus of our hypothesis tests (the intercept is only there to make the math work out properly and is rarely interpretable). The formulae for these parameter estimates use very familiar values.

First, we'll start with the intercept (a):

$$\mathrm{a} = \overline{X_y} - (\mathrm{b} \times \overline{X_x})$$

We have seen each of these before. $\overline{X_y}$ and $\overline{X_x}$ are the means of Y and X, respectively.

Next, the slope:

$$\frac{\sum(Diff_x \times Diff_y)}{\sum(Diff_X{}^2)}$$

In which "Diff" means the difference of each score from the mean. Thus, just like with standard deviations, we must subtract the mean from each score. The numerator of this formula means that, for each person, we multiply their score on the first variable minus the means of that variable times their score on the second variable minus the mean of that variable. When we do that for each participant, then we sum all of those products (scores that were multiplied). The denominator is the squaring of each of those differences score (each person's score minus the mean) for the first variable, then summing all of those squared into one number,

It is very important to point out that the Y values in the equations for a and b are our observed Y values in the dataset, NOT the predicted Y values ($\widehat{\mathrm{Y}}$) from our equation for the line of best fit. You may be asking why we would try to predict Y if we have an observed value of Y, and that is a very reasonable question. The answer is that we first need a group with all of the values

(observed X and observed Y) to create the intercept and slope. Then, we can use that on samples in which we don't have values for the Y variable. In other words, we need to use known values of Y to calculate the parameter estimates in our equation, and we use the difference between our observed values and predicted values ($Y-\widehat{Y}$) to see how accurate our equation is. Then, we can use regression to create a predictive model that we can then use to predict values of Y for other people for whom we only have information on X.

Example

Let's look at this from an applied example. Businesses often have more applicants for a job than they have openings available, so they want to know who among the applicants is most likely to be the best employee. There are many criteria that can be used, but one is a personality test for conscientiousness, with the belief being that more conscientious (more responsible) employees are better than less conscientious employees. A business might give their employees a personality inventory to assess conscientiousness and existing performance data to look for a relation. In this example, we have known values of the predictor (X, conscientiousness) and outcome (Y, job performance), so we can estimate an equation for a line of best fit and see how accurately conscientious predicts job performance, then use this equation to predict future job performance of applicants based only on their known values of conscientiousness from personality inventories given during the application process.

The key to assessing whether a linear regression works well is the difference between our observed and known Y values and our predicted $\widehat{Y}$ values. We use subtraction to find the difference between them ($Y-\widehat{Y}$) in the same way we use subtraction for deviation scores and sums of squares. The value ($Y-\widehat{Y}$) is our residual, which, as defined before, is how close our line of best fit is to our actual values. We can visualize residuals to get a better sense of what they are by creating a scatterplot and overlaying a line of best fit on it, as shown in Figure 16.2.1.

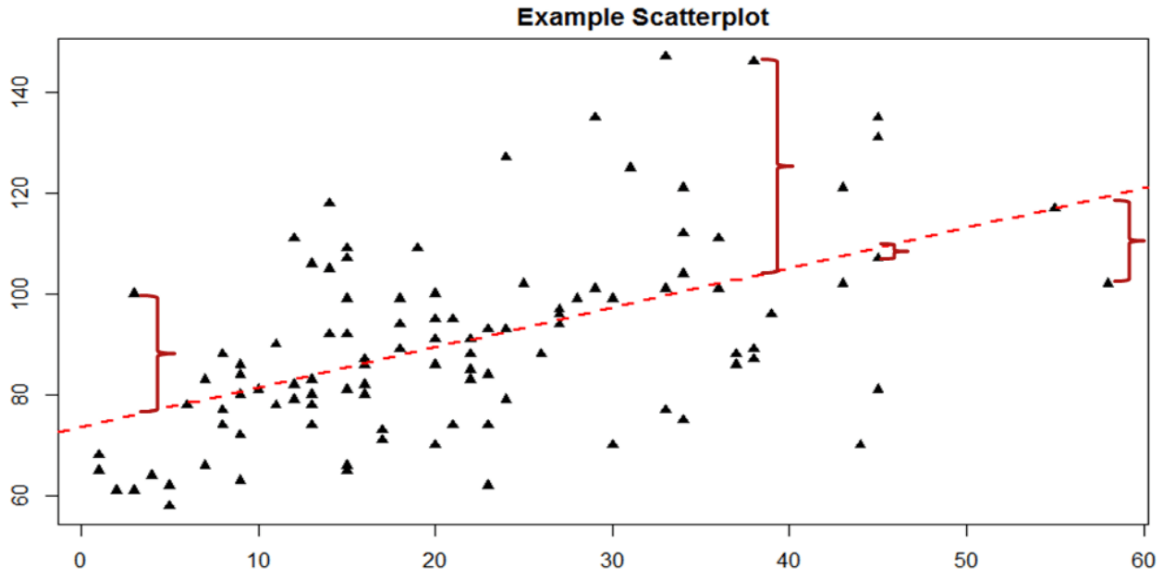


Figure 16.2.1: Scatterplot with Residuals. (CC-BY-NC-SA Foster et al. from from An Introduction to Psychological Statistics)

In Figure 16.2.1, the triangular dots represent observations from each person on both X (Conscientiousness scores) and Y (job performance ratings). The dashed red line is the line of best fit estimated by the equation $\widehat{Y} = a + (b * X)$. For every person in the dataset, the line represents their predicted score. The red bracket between a few of the triangular dots and the predicted scores on the line of best fit are our residuals (they are only drawn for four observations for ease of viewing, but in reality there is one for every observation); you can see that some residuals are positive and some are negative, and that some are very large and some are very small. This means that some predictions are very accurate and some are very inaccurate, and the some predictions overestimated values and some underestimated values. Across the entire dataset, the line of best fit is the one that minimizes the total (sum) value of all residuals. That is, although predictions at an individual level might be somewhat inaccurate, across our full sample and (theoretically) in future samples our total amount of error is as small as possible. We call this property of the line of best fit the Least Squares Error Solution. This term means that the solution – or equation – of the line is the one that provides the smallest possible value of the squared errors (squared so that they can be summed, just like in standard deviation) relative to any other straight line we could draw through the data.

Ack, that was a lot. It's okay if you don't totally understand what's going on here. This is the foundation of the equation and

the interpretation, but, just like with null hypothesis significance testing, you can figure out how to interpret the results without completely understanding the conceptual underpinnings.

Predicting Scores and Explaining Variance

The purpose of regression is twofold: we want to predict scores based on our line and, as stated earlier, explain variance in our observed Y variable just like in ANOVA. These two purposes go hand in hand, and our ability to predict scores is literally our ability to explain variance. That is, if we cannot account for the variance in Y based on X, then we have no reason to use X to predict future values of Y.

We know that the overall variance in Y (job performance in this example) is a function of each score deviating from the mean of Y (as in our calculation of variance and standard deviation). So, just like the red brackets in Figure $16.2.1$ representing residuals, given as $(Y-\widehat{Y})$, we can visualize the overall variance as each score's distance from the overall mean of Y, given as $(Y-\overline{X_y})$. This is shown in Figure $16.2.2$.

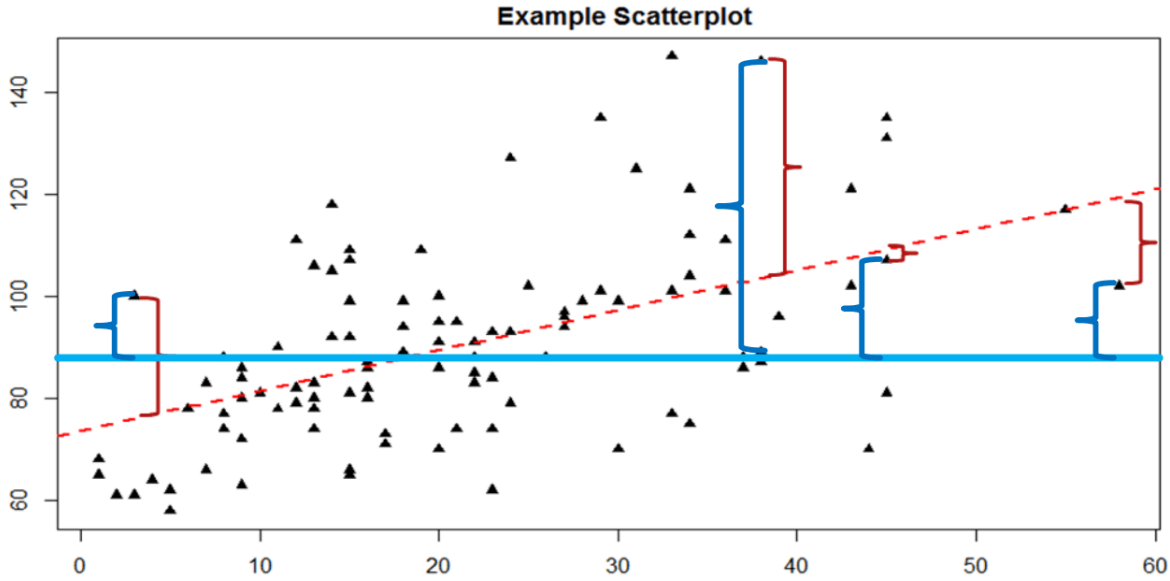


Figure $16.2.2$: Scatterplot with Residuals and Deviation Scores. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

In Figure $16.2.2$, the solid blue line is the mean of Y (job performanc), and the blue brackets are the deviation scores between our observed values of Y and the mean of Y. This represents the overall variance that we are trying to explain. Thus, the residuals and the deviation scores are the same type of idea: the distance between an observed score and a given line, *either* the line of best fit that gives predictions *or* the line representing the mean that serves as a baseline. Here's why math is so cool: The difference between these two values, which is the distance between the lines themselves, is our model's ability to predict scores above and beyond the baseline mean; that is, it is our model's ability to explain the variance we observe in Y (job performance) based on values of X (conscientiousness). If we have no ability to explain variance, then our line will be flat (the slope will be 0.00) and will be the same as the line representing the mean, and the distance between the lines will be 0.00 as well.

Summary

We now have three pieces of information: the distance from the observed score to the mean, the distance from the observed score to the prediction line, and the distance from the prediction line to the mean. These are our three pieces of information needed to test our hypotheses about regression. They are our three Sums of Squares, just like in ANOVA. Our distance from the observed score to the mean is the Sum of Squares Total, which we are trying to explain. Our distance from the observed score to the prediction line is our Sum of Squares Error, or residual, which we are trying to minimize. Our distance from the prediction line to the mean is our Sum of Squares Model, which is our observed effect and our ability to explain variance. Each of these will go into the ANOVA table to calculate our test statistic.

This page titled 16.2: Regression Line Equation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 13.2: Prediction by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

16.3: Using Linear Equations

Before we start practicing calculating all of the variables in a regression line equation, let's work a little with just the equation on it's own.

Regression Line Equations

As we just learned, linear regression for two variables is based on a linear equation:

$$\widehat{\mathrm{Y}} = \mathrm{a} + (\mathrm{b} * X)$$

where a and b are constant numbers. What this means is that for every sample, the intercept (a) and the slope (b) will be the same for every score. The X score will change, and that affects Y (or predicted Y, or $\widehat{\mathrm{Y}}$). Some consider the predictor variable (X) as an IV and the outcome variable (Y) as the DV, but be careful that you aren't confusing prediction with causation!

We also just learned that the graph of a linear equation of the form $\widehat{\mathrm{Y}} = \mathrm{a} + (\mathrm{b} * X)$ is a straight line.

Exercise 16.3.1

Is the following an example of a linear equation? Why or why not?

This is a graph of an equation. The x-axis is labeled in intervals of 2 from 0 - 14; the y-axis is labeled in intervals of 2 from 0 - 12. The equation's graph is a curve that crosses the y-axis at 2 and curves upward and to the right.

Figure 16.3.1. Sample Plotted Line (CC-BY by Barbara Illowsky & Susan Dean (De Anza College) from OpenStax)

Answer available at the end of this section.

The minimum criterion for using a linear regression formula is that there be a linear relationship between the predictor and the criterion (outcome) variables.

Exercise 16.3.2

What statistic shows us whether two variables are linearly related?

Answer available at the end of this section.

If two variables aren't linearly related, then you can't use linear regression to predict one from the other! The stronger the linear relationship (larger the Pearson's correlation), the more accurate will be the predictions based on linear regression.

Slope and Y-Intercept of a Linear Equation

As we learned previously, $b =$ slope and $a = y$-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis. Figure 16.3.2 shows three possible graphs of the regression equation ($y = a + b\mathrm{x}$). Panel (a) shows what the regression line looks like if the slope is positive ($b > 0$), the line slopes upward to the right. Panel (b) shows what the regression line looks like if there's no slope ($b = 0$); the line is horizontal. Finally, Panel (c) shows what the regression line looks like if the slope is negative ($b < 0$), the line slopes downward to the right.

Three plots with different regression lines. The first line is going up and to the right (positive correlation), the middle plot has a flat line, and the third plot is going down and to the right (negative correlation).

Figure 16.3.2: Three possible graphs of $y = a + b\mathrm{x}$. (CC-BY by Barbara Illowsky & Susan Dean (De Anza College) from OpenStax)

I get it, everything has been pretty theoretical so far. So let's get practical. Let's try constructing the regression line equation even when you don't have the scores for either of the variables. First, we'll start by identifying the variables in the examples.

Example 16.3.1

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the predictor and criterion (outcome) variables? What is the y-intercept and what is the slope? Answer using complete sentences.

Answer

The predictor variable, x, is the number of hours Svetlana tutors each session. The criterion (outcome) variable, y, is the amount, in dollars, Svetlana earns for each session.

The y-intercept is the constant, the one time fee of $25 ($a = 25$). The slope is 15 ($b = 15$) because Svetlana earns $15 for each hour she tutors.

Although it doesn't make sense in these examples, the y-intercept (a) is determined when $x = 0$. I guess with Svetlana, you could say that she gets $25 for any sessions that you miss or don't cancel ahead of time. But geometrically and mathematically, the y-intercept is based on when the predictor variable (x) has a value of zero.

Exercise 16.3.3

Jamal repairs household appliances like dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Jamal earns per visit is $y = 25 + 20x$.

What are the predictor and criterion (outcome) variables? What is the y-intercept and what is the slope? Answer using complete sentences.

Answer available at the end of this section.

Now, we can start constructing the regression line equations.

Example 16.3.2

Alejandra's Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the *equation* that expresses the total cost in terms of the number of hours required to complete the job. For this example,

- $x =$ the number of hours it takes to get the job done.
- $y =$ the total cost to the customer.

Answer

The $31.50 is a fixed cost. This is the number that you add after calculating the rest, so it must be the intercept (a).

If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only.

Thus, the total cost is: $y = 31.50 + 32x$

Let's try another example of constructing the regression line equation.

Exercise 16.3.4

Elektra's Extreme Sports hires hang-gliding instructors and pays them a fee of $50 per class as well as $20 per student in the class. The total cost Elektra pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Answer available at the end of this section.

You can also use the regression equation to graph the line if you input scores from your X variable and your Y variable into the equation. Let's see what that might look like in Figure 16.3.3 for the equation: $y = -1 + 2x$

Graph of the equation y = -1 + 2x. This is a straight line that crosses the y-axis at -1 and is sloped up and to the right, rising 2 units for every one unit of run.

Figure 16.3.3: Regression Line for $y = -1 + 2x$. (CC-BY by Barbara Illowsky & Susan Dean (De Anza College) from OpenStax)

In the example in Figure 16.3.3, the intercept (a) is replaced by -1 and the slope (b) is replaced by 2 to get the regression equation ($y = -1 + 2x$). Right now, you are being provided these constants. Soon, you'll be calculating them yourself!

Summary

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted. Algebraically, a linear equation typically takes the form $y = mx + b$, where m and b are constants, x is the independent variable, y is the dependent variable. In a statistical context, a linear equation is written in the form $y = a + bx$, where a and b are the constants. This form is used to help readers distinguish the statistical context from the algebraic context. In the equation $y = a + bx$, the constant b that multiplies the x variable (b is called a coefficient) is called the slope. The constant a is called the y-intercept.

The slope of a line is a value that describes the rate of change between the two quantitative variables. The slope tells us how the criterion variable (y) changes for every one unit increase in the predictor (x) variable, on average. The y-intercept is used to describe the criterion variable when the predictor variable equals zero.

Answers

Exercise 16.3.1

No, the graph is not a *straight* line; therefore, it is not a linear equation.

Exercise 16.3.2

Pearson's r (correlation).

Exercise 16.3.3

The predictor variable, x, is the number of hours Jamal works each visit. he criterion (outcome) variable, y, is the amount, in dollars, Jamal earns for each visit.

The *y*-intercept is 25 ($a = 25$). At the start of a visit, Jamal charges a one-time fee of $25 (this is when $x = 0$). The slope is 20 ($b = 20$). For each visit, Jamal earns $20 for each hour he works.

Exercise 16.3.4

For this example,

- $x =$ number of students in class
- $y =$ the total cost

The constant is $50 per class, so that must be the intercept (a).

So $20 per student is the slope (b).

The resulting regression equation is: $y = 50 + 20x$

This page titled 16.3: Using Linear Equations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 12.2: Linear Equations by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

16.4: Hypothesis Testing- Slope to ANOVAs

In regression, we are interested in predicting Y scores and explaining variance using a line, the slope of which is what allows us to get closer to our observed scores than the mean of Y can. Thus, our hypotheses can concern the slope of the line, which is estimated in the prediction equation by b.

Research Hypothesis

Specifically, we want to test that the slope is not zero. The research hypothesis will be that there is an explanatory relation between the variables.

- RH: $\beta > 0$
- RH: $\beta < 0$
- RH: $\beta \neq 0$

A non-zero slope indicates that we can explain values in Y based on X and therefore predict future values of Y based on X.

Null Hypothesis

Thus, the null hypothesis is that the slope *is* zero, that there is *no* explanatory relation between our variables

$$\text{Null Hypothesis} : \beta = 0$$

Regression Uses a ANOVA Summary Table

DId you notice that we don't have a test statistic yet (like t, F of ANOVA, or Pearson's r yet? To test the null hypothesis, we use the F statistic of ANOVA from an ANOVA Summary Table compared to a critical value from the F distribution table.

Our ANOVA table in regression follows the exact same format as it did for ANOVA (Table 16.4.1). Our top row is our observed effect, our middle row is our error, and our bottom row is our total. The columns take on the same interpretations as well: from left to right, we have our sums of squares, our degrees of freedom, our mean squares, and our F statistic.

Table 16.4.1: ANOVA Table for Regression

Source	SS	df	MS	F
Model	$\sum(\widehat{Y}-\overline{Y})^2$	1	SS_M/df_M	MS_M/MS_E
Error	$\sum(Y-\widehat{Y})^2$	$N-2$	SS_E/df_E	N/A
Total	$\sum(Y-\overline{Y})^2$	$N-1$	N/A	N/A

As with ANOVA, getting the values for the SS column is a straightforward but somewhat arduous process. First, you take the raw scores of X and Y and calculate the means, variances, and covariance using the sum of products table introduced in our chapter on correlations. Next, you use the variance of X and the covariance of X and Y to calculate the slope of the line, b, the formula for which is given above. After that, you use the means and the slope to find the intercept, a, which is given alongside b. After that, you use the full prediction equation for the line of best fit to get predicted Y scores ($\widehat{Y}$) for each person. Finally, you use the observed Y scores, predicted Y scores, and mean of Y to find the appropriate deviation scores for each person for each sum of squares source in the table and sum them to get the Sum of Squares Model, Sum of Squares Error, and Sum of Squares Total. As with ANOVA, you won't be required to compute the SS values by hand, but you will need to know what they represent and how they fit together.

The other columns in the ANOVA table are all familiar. The degrees of freedom column still has N–1 for our total, but now we have N–2 for our error degrees of freedom and 1 for our model degrees of freedom; this is because simple linear regression only has one predictor, so our degrees of freedom for the model is always 1 and does not change. The total degrees of freedom must still

be the sum of the other two, so our degrees of freedom error will always be $N–2$ for simple linear regression. The mean square columns are still the SS column divided by the df column, and the test statistic F is still the ratio of the mean squares. Based on this, it is now explicitly clear that not only do regression and ANOVA have the same goal but they are, in fact, the same analysis entirely. The only difference is the type of data we have for the IV (predictor): a quantitative variable for for regression and groups (qualitative) for ANOVA. The DV is quantitative for both ANOVAs and regressions/correlations.

With a completed ANOVA Table, we follow the same process of null hypothesis significance testing by comparing our calculated F-score to a critical F-score to determine if we retain or reject the null hypothesis. In ANOVAs, the null hypothesis was that all of the means would be similar, but with correlations (which are what regression is based on), the null hypothesis says that there is no linear relationship. However, what we are really testing is how much variability in the criterion variable (y) can be explained by variation in the predictor variable (y). So, for regression using ANOVA, the null hypothesis is saying that the predictor variable does not explain variation in the criterion variable.

This is a little confusing, so let's take a look at an example of regression in action.

This page titled 16.4: Hypothesis Testing- Slope to ANOVAs is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 13.4: Hypothesis Testing in Regression by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.
- 13.3: ANOVA Table by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

16.5: Practice Regression on Health and Happiness

This practice will include the whole process to calculate all of the equations and the ANOVA Summary Table.

Scenario

In this scenario, researchers are interested in explaining differences in how happy people are based on how healthy people are. They gather quantitative data on each of these variables from 18 people and fit a linear regression model (another way to say that they constructed a regression line equation) to explain the variability of one variable (say, happiness) based on the variability of the other variable (health). We will follow the five-step null hypothesis significance testing procedure to see if there is a relation between these variables that is statistically significant.

Step 1: Hypotheses

What do you think? Do you think that happiness and health vary in the same direction, in opposite directions, or do not vary together? This sounds like the set up for a research hypothesis for a correlation, and it is! The small change here is that we're focusing on the slope (b or beta) of the line, rather merely testing if there is a linear relationship.

Example 16.5.1

What could be a research hypothesis for this scenario? State the research hypothesis in words and symbols.

Solution

The research hypothesis should probably be something like, "There will be a positive slope in the regression line for happiness and health."

Symbols: $\beta > 0$

For right now, we haven't made it clear which might be our predictor variable and which might be our criterion variable, the outcome. If we decide that health is the predictor and happiness is the outcome, we could add that we think that changes in health predict changes in happiness. It seems reasonable in this situation to reverse the IV and DV here, and say that changes in happiness could also predict changes in health. There's data to support both of these ideas!

The null hypothesis in regression states that there is no relation between our variables so you can't use one variable to predict or explain the other variable.

Example 16.5.2

What is the null hypothesis for this scenario? State the null hypothesis in words and symbols.

Solution

The null hypothesis is that there is no relationship between health and happiness, so "The slope in the regression line for happiness and health will be zero." Neither variable can predict or explain the other variable.

Symbols: $\beta = 0$

These hypothesis are not as clear-cut as we've previously had because regression analyses can show a lot. If it's easier, you can fall back on the general hypothesis for correlations rather than focus on slopes. However, that misses the addition information that regressions also show: How much variability in one variable is due to the other variable.

Step 2: Collect Data

We can still identify information from the scenario even if we didn't collect the data ourselves. Regression analyses are built off of correlational analyses, so we again have two different quantitative variables that were gathered from a sample. Later, you'll see an ANOVA Summary Table; this is used more to test how strong one variable affects the other variable, so we won't treat it like we're comparing means of groups.

Example 16.5.3

Read the scenario and check out Table 16.5.1 to answer the following questions.

1. Who is the sample?
2. Who do might be the population?
3. What two different (quantitative) variables were measured?

Solution

1. The sample is the 18 people.
2. The population might be any person? The scenario didn't give us much information on who the data was collected from so we can't narrow our population down.
3. The two variables measured were happiness and health.

Step 3: Analyze Data

The process of calculating the test statistic for regression first involves computing the parameter estimates for the line of best fit. To do this, we first calculate the means, standard deviations, and sum of products for our variables, as shown in Table 16.5.1.

Table 16.5.1: Raw Scores in Empty Sum of Products Table

Health	Difference: Health - Mean	Health Difference Squared	Happiness	Difference: Happiness - Mean	Happiness Difference Squared	Health Diff * Happy Diff
16.99			16.38			
17.42			16.89			
17.65			10.36			
18.21			18.49			
18.28			14.26			
18.30			15.23			
18.39			12.08			
18.63			17.57			

Health	Difference: Health - Mean	Health Difference Squared	Happiness	Difference: Happiness - Mean	Happiness Difference Squared	Health Diff * Happy Diff
18.89			21.96			
19.45			21.17			
19.67			18.12			
19.91			17.86			
20.35			18.74			
21.89			17.71			
22.48			17.11			
22.61			16.47			
23.25			21.66			
23.65			22.13			
$\sum = 356.02$	$\sum = ?$	$\sum = ?$	$\sum = 314.18$	$\sum = ?$	$\sum = ?$	$\sum = ?$

This table should look pretty familiar, as it's the same as the one we used to calculate a correlation when we didn't have have the standard deviation provided.

First things first, let's find the means so that we can start filling in this table.

Exercise 16.5.1

What is the average score for the Health variable? What is the average score for the Happiness variable?

Answer available at the end of this section.

Now that we have the means, we can fill in the complete Sum of Products table.

Exercise 16.5.2

Fill in Table 16.5.1 by finding the differences of each score from that variable's mean, squaring the differences, multiplying them, then finding the sums of each of these.

Answer available at the end of this section.

In Table 16.5.2, the difference scores for each variable (each score minus the mean for that variable) sum to nearly zero, so all is well there. Let's use the sum of those squares to calculate the standard deviation for each variable.

Example 16.5.4

Calculate the standard deviation of the Health variable.

Solution

Using the standard deviation formula:

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

We can fill in the numbers from the table and N to get:

$$s_{Hth} = \sqrt{\frac{76.07}{18 - 1}} = \sqrt{\frac{76.07}{17}}$$

Some division, then square rooting to get:

$$s_{Hth} = \sqrt{4.47}$$

$$s_{Hth} = 2.11$$

If you used spreadsheet, you might get 2.12. That's fine, but we'll use 2.11 for future calculations.

Your turn!

Exercise 16.5.3

Calculate the standard deviation of the Happiness variable.

Answer available at the end of this section.

Next up, we must calculate the slope of the line. There are easier ways to show this if we start substituting names for symbols, but let's stick with the names of our variables in this formula:

$$\mathrm{b} = \frac{(Diff_{Hth} \times Diff_{Hpp})}{Diff_{Hth}^2} = \frac{58.22}{76.07} = 0.77$$

What this equation is telling us to do is pretty simple once we have the Sum of Products table filled out. We just take the sum of the multiplication of each of the differences (that's the most bottom most right cell in the table), and divide that by the sum of the difference scores squared for the Health variable. The result means that as Health (X) changes by 1 unit, Happiness (Y) will change by 0.77. This is a positive relation.

Next, we use the slope, along with the means of each variable, to compute the intercept: $a = \overline{X_y} - b\overline{X_x}$

Example 16.5.5

Using the means and the slope (b) that we just calculated, calculate the intercept (a).

Solution

$$a = \overline{X_y} - b\overline{X_x}$$

$$a = \overline{X}_{Hpp} - (b \times \overline{X}_{Hth})$$

$$a = 17.46 - (0.77 * 19.78)$$

$$a = 17.46 - (15.23)$$

$$a = 2.23$$

This number varies widely based on how many decimal points you save while calculating. The numbers shown are when only two decimal points are used, which is the minimum that you should be writing down and saving.

For this particular problem (and most regressions), the intercept is not an important or interpretable value, so we will not read into it further.

Now that we have all of our parameters estimated, we can give the full equation for our line of best fit: $\hat{y} = \mathrm{a} + (\mathrm{b} \times X)$

Example 16.5.6

Construct the regression line equation for the predicted Happiness score (\(\widehat{y}).

Solution

$$\hat{y} = 2.23 + 0.77\mathrm{x}$$

It doesn't quite make sense yet, but you actually just answered many potential research questions! We can also plot this relation in a scatterplot and overlay our line onto it, as shown in Figure 16.5.1.

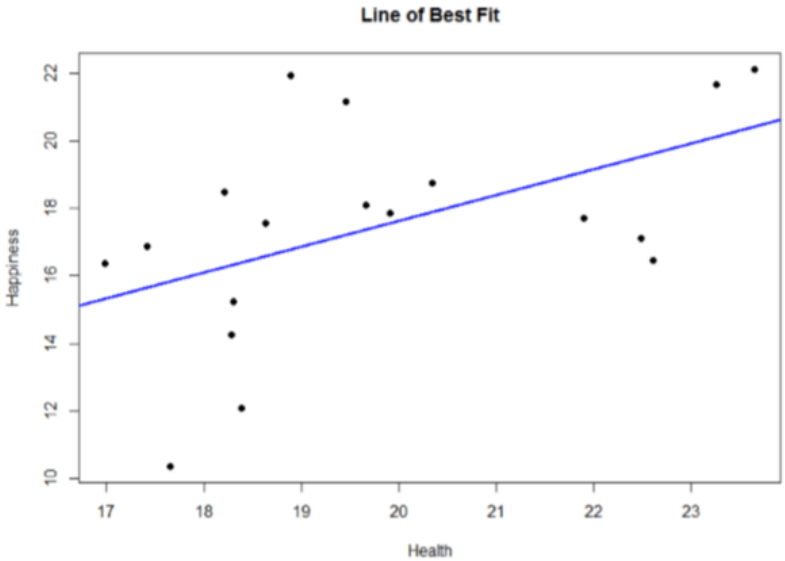


Figure 16.5.1: Health and Happiness Data and Regression Line. (CC-BY-NC-SA Foster et al. from An Introduction to Psychological Statistics)

We can use the regression line equation to find predicted values for each observation and use them to calculate our sums of

squares for the Model and the Error, but this is tedious to do by hand, so we will let the computer software do the heavy lifting in that column of our ANOVA Summary Table in Table 16.5.3.

Table 16.5.3: ANOVA Summary Table with SS

Source	SS	df	MS	F
Model	44.62			
Error	129.37			N/A
Total	173.99		N/A	N/A

Happily, the Total row is still the sum of the other two rows, so that was added, too. Now that we have these, we can fill in the rest of the ANOVA table.

Example 16.5.7

Fill in the rest of the ANOVA Summary Table from Table 16.5.3.

Solution

Table 16.5.4: ANOVA Summary Table with SS

Source	SS	df	MS	F
Model	44.62	1	$MS_M = \frac{SS_M}{df_M} = \frac{44.62}{1} = 44.62$	$F = \frac{MS_M}{MS_E} = \frac{44.62}{8.09} = 5.52$
Error	129.37	$N - 2 = 18 - 2 = 16$	$MS_E = \frac{SS_E}{df_E} = \frac{129.37}{16} = 8.09$	leave blank
Total	173.99	$N - 1 = 18 - 1 = 17$	leave blank	leave blank

Happily again, we can do a computation check to make sure that our Degrees of Freedom are correct since the sum of the Model's df and the Error's df should equal the Total df. And since $1 + 16 = 17$, we're doing well!

This gives us an obtained F statistic of 5.52, which we will now use to test our hypothesis.

Step 4: Conclusion

Let's compare our calculated value to a critical value to see if we retain or reject the null hypothesis. Our critical value for regression will come from the Table of Critical Values of F (found through the Common Critical Values page at the end of the book).

The ANOVA Summary Table used for regression uses two types of degrees of freedom. We previously saw that the Degrees of Freedom for our numerator, the Model line, is always 1 in a linear regression, and that the denominator degrees of freedom, from the Error line, is N–2.

In this instance, we have 18 people so our degrees of freedom for the denominator is 16. Going to our F table, we find that the appropriate critical value for 1 and 16 degrees of freedom is $F_{Critical}$ = 4.49. Isn't it nice to do the simple things that we learned about seemingly eons ago?

Our calculated test statistic was $F_{Calc} = 5.52$ and our critical value was $F_{Crit} = 4.49$. Since our calculated test statistic is greater than our critical value, we can reject the null hypothesis because this is *still* true:

Note

Critical $<$ **Calculated** $=$ **Reject null** $=$ **There is a linear relationship.** $= p < .05$

Critical $>$ **Calculated** $=$ **Retain null** $=$ **There is not a linear relationship.** $= p > .05$

Now that we know that we rejected the null hypothesis, we can check to see if we can support our research hypothesis. The research hypothesis was that here will be a positive slope in the regression line for happiness and health. The positive Sum or Products shows a positive slope so the research hypothesis is supported!

Step 5: Write-Up

We got here a sorta roundabout way, so it's hard to figure out what our write-up should look like. Let's start by including the components needed for reporting results.

Example 16.5.8

Use the four components for reporting results to start your concluding paragraph:

1. Re-state the research hypothesis.
2. State if the research hypothesis was supported or not.
3. Write out the statistical sentence.
4. If relevant, name the means for each variable.

Solution

1. The research hypothesis was that here will be a positive slope in the regression line for happiness and health.
2. The results are statistically significant, and the positive Sum or Products shows a positive slope so the research hypothesis is supported.
3. The ANOVA results were F(1,16)=5.52, p< .05.
4. Means aren't really relevant for this research question, but the average health score was 19.78 and the average happiness score was 17.46.

Let's add one more sentence for a nice little conclusion to our regression analysis: We can predict levels of happiness based on how healthy someone is. So, our final write-up could be something like:

> We can predict levels of happiness (M=17.46) based on how healthy someone is (M=19.78) (F(1,16)=5.52, p< .05). The average health score was 19.78. The average happiness score was 17.46. The research hypothesis was supported; there is a positive slope in the regression line for happiness and health.

Yep, that's pretty clunky. As you learn more about regression (in future classes), the research hypotheses will make more sense.

Accuracy in Prediction

We found a large, statistically significant relation between our variables, which is what we hoped for. However, if we want to use our estimated line of best fit for future prediction, we will also want to know how precise or accurate our predicted values are. What we want to know is the average distance from our predictions to our actual observed values, or the average size of the residual ($Y - \widehat{Y}$). The average size of the residual is known by a specific name: the standard error of the estimate ($S_{(Y-\widehat{Y})}$), which is given by the formula

$$S_{(Y-\widehat{Y})} = \sqrt{\frac{\sum (Y-\widehat{Y})^2}{N-2}}$$

This formula is almost identical to our standard deviation formula, and it follows the same logic. We square our residuals, add them up, then divide by the degrees of freedom. Although this sounds like a long process, we already have the sum of the squared residuals in our ANOVA table! In fact, the value under the square root sign is just the SS_E divided by the df_E, which we know is called the mean squared error, or MS_E:

$$s_{(Y-\widehat{Y})} = \sqrt{\frac{\sum (Y-\widehat{Y})^2}{N-2}} = \sqrt{MS_E}$$

For our example:

$$s_{(Y-\widehat{Y})} = \sqrt{\frac{129.37}{16}} = \sqrt{8.09} = 2.84$$

So on average, our predictions are just under 3 points away from our actual values.

There are no specific cutoffs or guidelines for how big our standard error of the estimate can or should be; it is highly dependent on both our sample size and the scale of our original Y variable, so expert judgment should be used. In this case, the estimate is not that far off and can be considered reasonably precise.

Clear as mud? Let's try another practice problem!

Answers

Exercise 16.5.1

$$\overline{X}_{Hth} = \frac{\sum X}{N} = \frac{356.02}{18} = 19.78$$
$$\overline{X}_{Hpp} = \frac{\sum X}{N} = \frac{314.19}{18} = 17.46$$

Exercise 16.5.2

Table 16.5.2- Completed Sum of Products Table

Health	Difference: Health - Mean	Health Difference Squared	Happiness	Difference: Happiness - Mean	Happiness Difference Squared	Health Diff * Happy Diff
16.99	-2.79	7.78	16.38	-1.08	1.17	3.01
17.42	-2.36	5.57	16.89	-0.57	0.32	1.35
17.65	-2.13	4.54	10.36	-7.10	50.41	15.12
18.21	-1.57	2.46	18.49	1.03	1.06	-1.62
18.28	-1.50	2.25	14.26	-3.20	10.24	4.80
18.30	-1.48	2.19	15.23	-2.23	4.97	3.30
18.39	-1.39	1.93	12.08	-5.38	28.94	7.48
18.63	-1.15	1.32	17.57	0.11	0.01	-0.13
18.89	-0.89	0.79	21.96	4.50	20.25	-4.01
19.45	-0.33	0.11	21.17	3.71	13.76	-1.22
19.67	-0.11	0.01	18.12	0.66	0.44	-0.07
19.91	0.13	0.02	17.86	0.40	0.16	0.05
20.35	0.57	0.32	18.74	1.28	1.64	0.73
21.89	2.11	4.45	17.71	0.25	0.06	0.53
22.48	2.70	7.29	17.11	-0.35	0.12	-0.95
22.61	2.83	8.01	16.47	-0.99	0.98	-2.80
23.25	3.47	12.04	21.66	4.20	17.64	14.57
23.65	3.87	14.98	22.13	4.67	21.81	18.07
$\sum = 356.02$	$\sum = -0.02$	$\sum = 76.07$	$\sum = 314.18$	$\sum = -0.09$	$\sum = 173.99$	$\sum = 58.22$

Exercise 16.5.3

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

$$s_{Hpp} = \sqrt{\frac{173.99}{18 - 1}} = \sqrt{\frac{173.99}{17}}$$

$$s_{Hpp} = \sqrt{10.23}$$

$$s_{Hpp} = 3.20$$

This page titled 16.5: Practice Regression on Health and Happiness is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 13.5: Happiness and Well-Being by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

16.6: Practice Regression on Nutrition

Let's try a second example using the same nutrition data set from fast food restaurants from OpenIntro.org as the chapter on correlations. If you grab your calculations and Sum of Products table from that practice with the nutrition data, you'll be ahead of the game!

Scenario

When Dr. MO was on a low-carb diet, she noticed that things that were low in carbohydrates tended to be high in fat, and vice versa. Our prior analysis found a strong correlation between fat and carbs, but not in the direction that Dr. MO expected! Now, we will use our knowledge of this strong correlation to see if we can use the Total Fat in our sample chicken dishes from one fast food restaurant to predict the Total Carbs in those same dishes.

Step 1: State the Hypotheses

To simplify regression, our hypotheses are still a little clunky, but hopefully you'll see the value of regression analyses in the end.

Example 16.6.1

What could be a research hypothesis for this scenario? State the research hypothesis in words and symbols.

Solution

The research hypothesis should probably be something like, "There will be a positive slope in the regression line for Total Fat and Total Carbs."

Symbols: $\beta > 0$

In other words, we're trying to construct a statistically significant (as determined by an ANOVA) regression line equation that could use to predict Total Carbs from Total Fat.

Example 16.6.2

What is the null hypothesis for this scenario? State the null hypothesis in words and symbols.

Solution

The null hypothesis is that there is no relationship between Total Fat and Total Carbs, so "The slope in the regression line for Total Fat and Total Carbs will be zero." Neither variable can predict or explain the other variable.

Symbols: $\beta = 0$

Step 2: Collect Data

Again, let's identify information from the scenario even if we didn't collect the data ourselves.

Example 16.6.1

Read the scenario to answer the following questions, although you might need Table 16.6.1 or Table 16.6.2 to find the number of data points in this study.

1. Who is the sample?
2. Who do might be the population?
3. What two different (quantitative) variables were measured?

Solution

1. The sample is the 13 chicken meals from one fast food restaurant.
2. The population might any chicken meal from that fast food restaurant, or even any chicken meal from any fast food restaurant.
3. The two variables measured were amount of fat and amount of carbohydrates in each meal.

Step 3: Analyze Data

Okay, here's where the tedious part starts! Use the data in Table 16.6.1 to start your calculations!

Table 16.6.1: Raw Scores in Empty Sum of Products Table

Total Fat	Difference: Fat - Mean	Fat Difference Squared	Total Carbs	Difference: Carbs - Mean	Carbs Difference Squared	Fat Diff * Carb Diff
11			19			
12			30			
13			31			
17			28			
24			53			
24			26			
26			52			
28			47			
31			66			
31			56			
35			53			
44			81			
46			53			
$\sum = 342.00$	$\sum = ?$	$\sum = ?$	$\sum = 595.00$	$\sum = ?$	$\sum = ?$	$\sum = ?$

To start filling in the Table 16.6.1, we'll need to find the means again for Total Fat and Total Carbs. Since we did that in the last chapter, let's just put them in a Table 16.6.2.

Table 16.6.2- Descriptive Statistics of Nutrition Information

Measurement	Mean	Standard Deviation	N
Total Fat	26.31	11.32	13
Total Carb	45.77	17.89	13

Now that we have the means, we can fill in the complete Sum of Products table.

Exercise 16.6.1

Fill in Table 16.6.1 by finding the differences of each score from that variable's mean, squaring the differences, multiplying them, then finding the sums of each of these.

Answer available at the end of this section.

We can do a computation check to see if our difference scores each score minus the mean for that variable) were calculated correctly by seeing if they each sum to nearly zero, which they do; yay! We could use the sum of those squares to calculate the standard deviation for each variable, but that was already provided in Table 16.6.2.

So, we have what we need to complete this activity. But what is this activity? Ultimately, our goal is to predict Total Carbs using Total Fat with a regression line equation ($\hat{y} = a + b\text{x}$). So, we need to find the slope (b) and the intercept (a).

Here's the formula for slope:

$$\text{b} = \frac{(Diff_x \times Diff_y)}{Diff_x^2} = \frac{(Diff_F \times Diff_C)}{Diff_F^2}$$

What this equation is telling us to do is pretty simple since we have the Sum of Products table filled out.

Example 16.6.4

Calculate the slope for this scenario.

Solution

$$\text{b} = \frac{1997.92}{1539.77} = 1.30$$

The result means that as Total Fat (X) changes by 1 unit, Total Carb (Y) will change by 1.30. This is a positive relation.

Next, we use this slope (b = 1.30), along with the means of each variable, to compute the intercept:

$$a = \overline{X_y} - (b \times \overline{X_x}) = \overline{X_C} - (b \times \overline{X_F})$$

Example 16.6.5

Using the means and the slope (b) that we just calculated, calculate the intercept (a).

Solution

$$a = 45.77 - (1.30 * 26.31)$$

$$a = 45.77 - (34.20)$$

$$a = 11.57$$

Now that we have all of our parameters estimated, we can give the full equation for our line of best fit: $\hat{y} = a + b\text{x}$

Example 16.6.6

Construct the regression line equation for to predict Total Carbs (\(\widehat{y}).

Solution

$$\hat{y} = 11.57 + 1.30x$$

Let's look at that regression line on our scatterplot generated using statistical software.

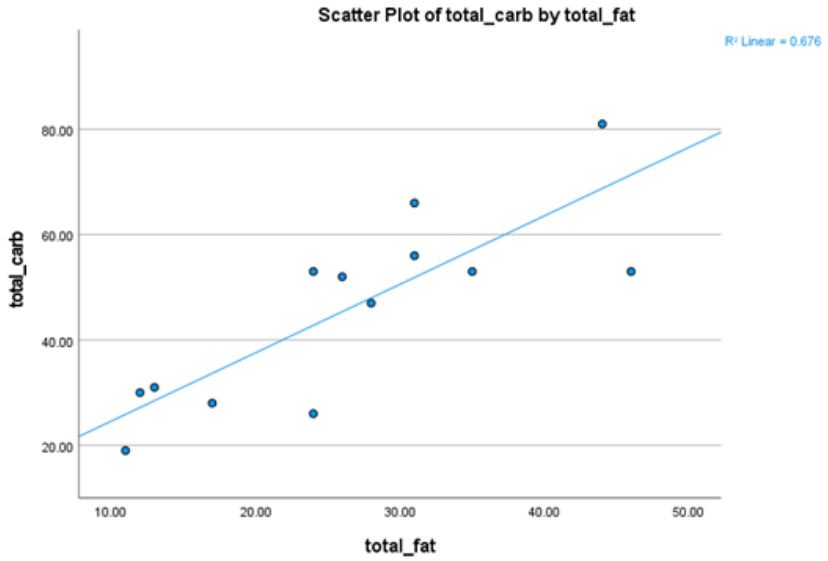


Figure 16.6.1: Scatterplot and Regression Line of Fast Food Nutrition Data (CC-BY-SA Michelle Oja via data from OpenIntro.org)

It looks like our regression line is a pretty good match for our data! But to make sure, let's make sure that we have a statistically significant model using this regression line. Table 16.6.4 shows an ANOVA Summary Table with the Sum of Squares already included from an analysis conducted using statistical software.

Table 16.6.4: ANOVA Summary Table with SS

Source	SS	df	MS	F
Model	1038.88			
Error	497.89			
Total				

Example 16.6.7

Fill in the rest of the ANOVA Summary Table from Table 16.6.4.

Solution

Table 16.6.5: ANOVA Summary Table

Source	SS	df	MS	F
Model	1038.88	1	$MS_M = \frac{SS_M}{df_M} = \frac{1038.88}{1} = 1038.88$	$F = \frac{MS_M}{MS_E} = \frac{1038.88}{45.26} = 22.95$

Source	SS	df	MS	F
Error	497.889	$N-2=13-2=11$	$MS_E=\frac{SS_E}{df_E}=\frac{497.89}{11}=45.26$	N/A
Total	173.99	$N-1=13-1=14$	N/A	N/A

We can do a computation check to make sure that our Degrees of Freedom are correct since the sum of the Model's df and the Error's df should equal the Total df. And since 1 + 11 = 12, we're on the right track.

This gives us an obtained F statistic of 22.95, which we will now use to test our hypothesis.

Step 4: Conclusion

We've done the hard part! Now we get to see if our research hypothesis is supported or not!

First, let's find the critical value so that we can compare it to our calculated value to see if the null hypothesis is retained or rejected. Our critical value for regression will come from the Table of Critical Values of F, which shows that the appropriate critical value at p=0.05 is $F_{Critical}(1,11)=4.84$.

We now have everything we need to make our final hypothesis testing decision. Our calculated test statistic was $F_{Calc}=22.95$ and our critical value was $F_{Crit}=4.84$. Since our calculated test statistic is greater than our critical value, we can reject the null hypothesis because this is *still* true:

Note

(Critical $<$ Calculated) = Reject null = There is a linear relationship. $=p<.05$

(Critical $>$ Calculated) = Retain null = There is not a linear relationship. $=p>.05$

But rejecting the null hypothesis doesn't always automatically mean that we support a directional research hypothesis. We again hypothesized a positive slope, and our Sum of Products was positive, though, so we did support our research hypothesis.

Step 5: Write-Up

Let's use the components needed for reporting results to organize our conclusions so far.

Example 16.6.8

Describe the components for reporting results for this scenario in complete sentences.

Solution

1. Re-state the research hypothesis: The research hypothesis was that here will be a positive slope in the regression line for Total Fat and Total Carbs.
2. State if the research hypothesis was supported or not: The results are statistically significant, and the positive Sum or Products shows a positive slope so the research hypothesis is supported.
3. Write out the statistical sentence F(1,11)=22.95, p<.05
4. If relevant, name the means for each variable: For our sample of 13 chicken dishes from one fast food restaurant, the average Total Fat was 26.31 and the average Total Carb was 45.77.

Using Regression

But really, what does that even mean? How is any of that useful? Well, it turns out that because our regression equation was found to be statistically significant, we can use it to make predictions.

Example 16.6.9

Imagine that you went to the fast food restaurant that this data is from, and they had a new chicken meal with even lower fat than anything currently on their menu (10 grams of Total Fat). Use the regression line equation to estimate how many Total Carbs would be in this new meal. Don't forget to end with a complete sentence.

Solution

With this regression line equation:

$$\hat{y} = 11.57 + 1.30\text{x}$$

We would replace the "x" placeholder with our new value of 10 grams of Total Fat to find:

$$\hat{y} = 11.57 + (1.30 * 10)$$

$$\hat{y} = 11.57 + (13.00)$$

$$\hat{y} = 24.57$$

The estimated Total Carbs of this new chicken meal with 10 Total Fat is 24.57 grams of carbohydrates.

That's not what we hypothesized, but this is the whole point of regression, to use a variable (or variables) that we have to predict variables that we don't have. Knowing this, let's try a final conclusion.

> In our sample of 13 chicken dishes from one fast food restaurant, we found the average Total Fat to be 26.31 and the average Total Carbs to be 45.77. This data was used to create a statistically significant regression equation (F(1, 11)=22.95, p<.05), which supports the research hypothesis that there would be a positive slope. If the fast food restaurant came out with a new low-fat chicken dish (10 grams of fat), we can predict that the Total Carbs of that dish would be 24.57 grams.

Accuracy in Prediction

Wanna know how accurate that prediction is? Square root the Mean Square of the Error term from the ANOVA Summary Table. As this formula shows, this is similar to the standard deviation formula because it follows the same logic:

$$S_{(Y-\widehat{Y})} = \sqrt{\frac{\sum (Y-\widehat{Y})^2}{N-2}}$$

But since we have the MS of the Error in Table 16.6.5, we can just square root 42.26:

$$s_{(Y-\widehat{Y})} = \sqrt{MS_E} = \sqrt{45.26} = 6.73$$

So on average, our predictions will be almost 7 (6.73) grams away from the actual values. There are no specific cutoffs or guidelines for how big our standard error of the estimate can or should be. What do you think? Is a variation around the prediction of 7 grams of carbohydrates precise enough? There's no right or wrong answer, just your thoughts and opinion!

Hopefully you saw how regression equations can be useful for making predictions about values that we don't have yet based on samples of two variables that we do already have data for. In the next section, you'll learn how we can combine multiple variables to predict one variable.

Contributors and Attributions

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

Answers

Exercise 16.6.1

Completed Sum of Products table:

Table 16.6.3: Completed Sum of Products Table

Total Fat	Difference: Fat - Mean	Fat Difference Squared	Total Carbs	Difference: Carbs - Mean	Carbs Difference Squared	Fat Diff * Carb Diff
11	-15.31	234.40	19	-26.77	716.63	409.85
12	-14.31	204.78	30	-15.77	248.69	225.67
13	-13.31	177.16	31	-14.77	218.15	196.59
17	-9.31	86.68	28	-17.77	315.77	165.44
24	-2.31	5.34	53	7.23	52.27	-16.70
24	-2.31	5.34	26	-19.77	390.85	45.67
26	-0.31	0.10	52	6.23	38.81	-1.93
28	1.69	2.86	47	1.23	1.51	2.08
31	4.69	22.00	66	20.23	409.25	94.88
31	4.69	22.00	56	10.23	104.65	47.98
35	8.69	75.52	53	7.23	52.27	62.83
44	17.69	312.94	81	35.23	1241.15	623.22
46	19.69	387.70	53	7.23	52.27	142.36
$\sum = 342.00$	$\sum = 0.00$	$\sum = 1536.77$	$\sum = 595.00$	$\sum = -0.01$	$\sum = 3842.31$	$\sum = 1997.92$

This page titled 16.6: Practice Regression on Nutrition is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

16.7: Multiple Regression

Simple linear regression as presented here is only a stepping stone towards an entire field of research and application. Regression is an incredibly flexible and powerful tool, and the extensions and variations on it are far beyond the scope of this chapter (indeed, even entire books struggle to accommodate all possible applications of the simple principles laid out here).

Multiple Regression

The next step in regression is to study multiple regression, which uses multiple X variables as predictors for a single Y variable at the same time. In other words, we can use regression to "predict" one score/variable from many other scores/variables, as well as show which of the multiple variables contribute to the score on the target variable. This shows us, statistically, which variables are most related to the changes of the variable that we're trying to predict.

The general formula is pretty simple:

$$\widehat{Y} = a + (b \times X_1) + (b \times X_2) + (b \times X_\infty)$$

But the math of multiple regression is very complex but the logic is the same: we are trying to use variables that are statistically significantly related to our outcome to explain the variance we observe in that outcome.

We can keep adding as many predictor variables as we have data for! Imagine that you'd like to know what all contributes to getting an Associate degree in 2 years.

> **Note**
>
> If my target variable was graduation rates, what variables could affect it?

The first variables that come to mind are how many hours a students works, how many kids in the student's household, and how many units the student passes each semester. Did you think of others? Statistically, you throw all of the variables that you can think of into the regression to try to predict graduation. Then, you use the regression equation to figure out which variables actually contribute to graduation, and which don't. You can see how this "simple" statistical can be a really powerful tool to help college administrators make decisions. The tool can be used in any industry that has a modeling sample (a group of people who can provide all of the variables).

And More!

Other forms of regression include curvilinear models that can explain curves in the data rather than the straight lines used here, as well as moderation models that change the relation between two variables based on levels of a third. The possibilities are truly endless and offer a lifetime of discovery.

Before we wrap-up regression analyses, here are some practice exercises to see if you have the concepts down.

This page titled 16.7: Multiple Regression is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

- 13.6: Multiple Regression and Other Extensions by Foster et al. is licensed CC BY-NC-SA 4.0. Original source: https://irl.umsl.edu/oer/4.

17: CHI-SQUARE

17.1: Introduction to Chi-Square
17.2: Assumptions of the Test(s)
17.3: Introduction to Goodness-of-Fit
17.4: Interpretation of the Goodness-of-Fit Test
17.5: Goodness of Fit Formula
17.6: Practice Goodness of Fit on Pineapple on Pizza
17.7: Introduction to Test of Independence
17.8: Practice Test of Independence on College Sports
17.9: Practice Chi-Square on Fast Food Meals
17.10: Phi Correlation is Not a Correlation
17.11: Choosing the Correct Test- Chi-Square Edition

This page titled 17: Chi-Square is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

17.1: 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 17.1.1

What type is each of the following? Qualitative or Quantitative?

1. Hair color
2. Ounces of vodka
3. Ethnicity
4. MPG (miles per gallon)
5. Type of music

Answer available at the end of this section.

Exercise 17.1.2

Do you use means to find the average of qualitative or quantitative variables?

Answer available at the end of this section.

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 17.1.1: Pet Preferences

Contingency Table	Cat	Dog	Other	Total
Observed	14	17	5	36
Expected	*12*	*12*	*12*	*36*

Table 17.1.1 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.

Answers

Exercise 17.1.1

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. Ethnicity: Qualitative
4. MPG: Quantitative
5. Type of music: Qualitative

Exercise 17.1.2

Quantitative. Means are mathematical averages, so the variable has to be a number that measures something.

This page titled 17.1: Introduction to Chi-Square is shared under a CC BY-NC-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.

17.2: 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. Suppose Dr. Navarro is interested in proportion of babies born at a particular hospital that are assigned males at birth. She could walk around the maternity wards, and observe 20 infants assigned female at birth and only 10 infants assigned male at birth. Seems like a pretty convincing difference, right? But later on, it turns out that she'd actually walked into the same ward 10 times, and in fact had only seen 2 infants assigned as female and 1 infant assigned as male. Not as convincing, is it? The original 30 *observations* were massively non-independent because she really only had three observations. 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.

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. (1973). 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 17.2: 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.

17.3: Introduction to Goodness-of-Fit

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 17.3.1

What does "k" stand for?

Answer available at the end of this section.

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.

Answers

Exercise 17.3.1

The letter "k" usually stands for the number of groups. In Chi-Square, this would be the number of different categories.

This page titled 17.3: Introduction to Goodness-of-Fit is shared under a CC BY-NC-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.

17.4: Interpretation of the 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 17.4.1, called "Observed" for reasons that will become clear very soon:

Table 17.4.1- *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 17.4.1), 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

$$\text{Null Hypothesis for } P = (0.25, 0.25, 0.25, 0.25)$$

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:

$$\text{Research Hypothesis of } P \neq (0.25, 0.25, 0.25, 0.25)$$

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 17.4.2).

Table 17.4.2- Contingency Table of Cards Pretended to Select

Contingency Table	Clubs ♣	Diamonds ♢	Hearts ♡	Spades ♠
Observed Frequency (O)	35	51	64	50

Contingency Table	Clubs ♣	Diamonds ♢	Hearts ♡	Spades ♠
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 17.4.3.

Table 17.4.3-*Contingency Table with Differences Squared*

Contingency Table	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 17.4.4)!

Table 17.4.4- *Contingency Table with Observed, Expected, Difference, Difference Squared, and Divided by Expected*

Contingency Table	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 17.4.1 you can see that if we change the degrees of freedom, then the Chi-Square distribution changes shape quite substantially.

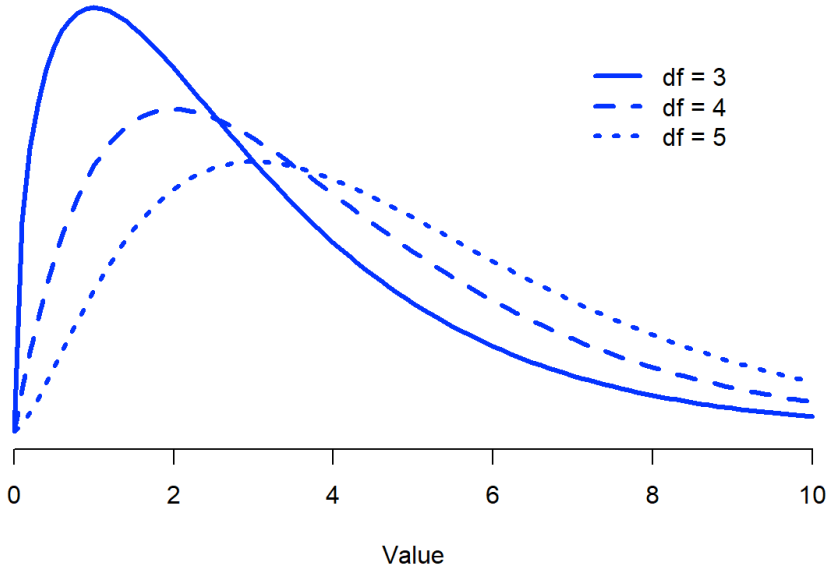


Figure 17.4.1- Chi-Square Distributions with Different df's(CC-BY-SA- Danielle Navarro from Learning Statistics with R).

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 17.4.2.

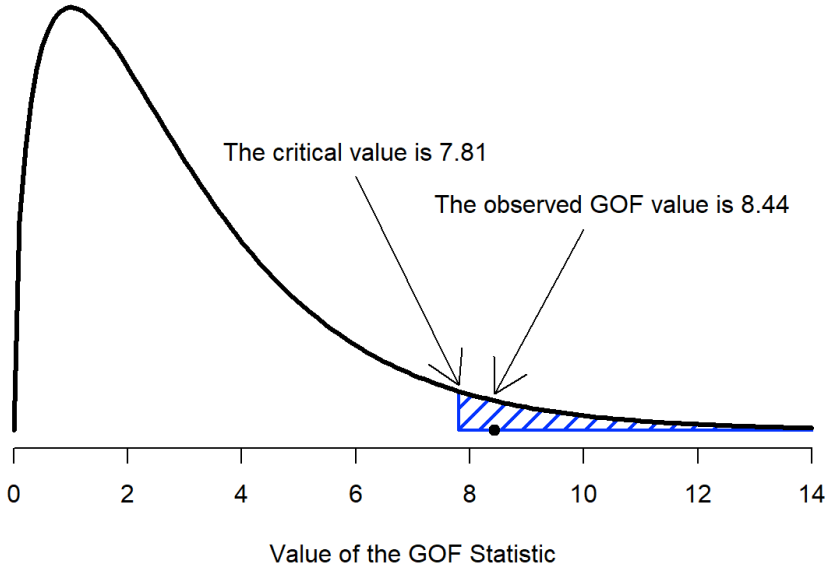


Figure 17.4.2- Illustration of how the hypothesis testing works for the chi-square goodness of fit test (CC-BY-SA- Danielle Navarro from Learning Statistics with R).

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 17.4.1

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 17.4: Interpretation of the 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/.

17.5: Goodness of Fit 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 17.5.1 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 17.5.1- Pet Preference Observations & Expectations

Contingency Table	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 17.5.1 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 17.5.1

Calculate the formula to complete Table 17.5.1.

Solution

Table 17.5.2- Pet Preference Observations & Expectations

Contingency Table	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 17.5: Goodness of Fit Formula is shared under a CC BY-NC-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.

17.6: Practice Goodness of Fit on 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 17.6.1

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 17.6.2

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 17.6.1

What is the critical value for this scenario?

Answer available at the end of this section.

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 17.6.1.

Table 17.6.1: Results of Data collection

Contingency Table	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 17.6.3

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 17.6.4

Complete the calculations labeled to fill in Table 17.6.2.

Table 17.6.2- Table to Complete Chi-Square Formula

Contingency Table	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 17.6.3- Table to Complete Chi-Square Formula

Contingency Table	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 17.6.5

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 17.6.6

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 17.6.2

What would our results look like in the statistical sentence?

Answer available at the end of this section.

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 17.6.7

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.

Answers

Exercise 17.6.1

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.

Exercise 17.6.2

$\chi^2(1)=1.08$, $p>.05$

This page titled 17.6: Practice Goodness of Fit on Pineapple on Pizza is shared under a CC BY-NC-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.

17.7: 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 17.7.1

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 17.7.1, 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 17.7.1: 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 17.7.1 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 17.7.1, 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 17.7.1, 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 17.7.1

Use the Observed frequencies from Table 17.7.1 to find the Expected frequencies for each combination of the two qualitative variables.

Solution

Table 17.7.2: 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 17.7: Introduction to Test of Independence is shared under a CC BY-NC-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.

17.8: Practice Test of Independence on 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 17.8.1: 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 17.8.2: 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 five step procedure that you've come to know.

Step 1: Hypotheses

Chi-Square tests 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 17.8.1

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; $2 \times 3 = 6$).

$$\frac{100}{6} = 16.67$$

So the probability that any random participant will fall into a specific cell is 0.167 for each cell.

Example 17.8.2

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 17.8.3

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: Collect Data

Okay, moving on from the "correlation doesn't equal causation" rant applied to Chi-Square! The data has already been collected, so let's identify what we have.

Example 17.8.4

Read the scenario to answer the following questions, although you might need Table 17.8.1 to find the number of students that the data was collected from in this study.

1. Who is the sample?
2. Who do might be the population?
3. What two different (qualitative) variables will we be testing?

Solution

1. The sample is the 168 students.
2. The population might be all students.
3. The two variables being analyzed are whether or not 168 college students watched college sports growing up and whether the students' final choice of which college to attend was influenced by the college's sports teams.

Step 3: Analyze Data

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 17.8.1 and Table 17.8.2 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 17.8.5

Use the previous two tables (Table 17.8.1 and Table 17.8.2) 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 17.8.3: 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 17.8.6

Square the difference scores in each cell of Table 17.8.3.

Solution

Table 17.8.4: 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 17.8.7

Using the squared differences in Table 17.8.4:, complete the formula by dividing each cell by it's own Expected frequency (found in Table 17.8.2). Then, add up all of the Total rows and columns to get the calculated χ^2.

Solution

Table 17.8.5: 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 17.8.8

Use the Chi-Square formula to calculate the χ^2 statistic

Solution

Using the information in Table 17.8.1 and Table 17.8.2, 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 about the research hypothesis.

Step 4: Conclusion

What is the decision: Should we support the research hypothesis or not?

You should be familiar with the first step to making this decision: compare the calculated value to the critical value. 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 17.8.1

What is the critical value for this scenario from the p = 0.05 column?

Answer available at the end of this section.

Comparing this critical value to our calculated value, should we retain or reject the null hypothesis?

Exercise 17.8.2

Should the null hypothesis be retained or rejected?

Answer available at the end of this section.

So what would the statistical sentence look like?

Exercise 17.8.3

What would our results look like in the statistical sentence?

Answer available at the end of this section.

When we reject the null hypothesis, we know that we can support a directional research hypothesis but that we might not (if the direction we found doesn't match the direction that we expected). However, the research hypothesis for Chi-Square is non-directional; we're saying that there will be a pattern of relationship between the variables but we aren't predicting what that pattern will be. This means that we can support the (non-directional) research hypothesis when we reject the null hypothesis! While non-directional research hypotheses slightly simplifies things, we are losing the specificity of predicting a direction and knowing if that prediction was supported or not.

Step 5: Write-Up

Let's use all that we've done to let people know what we found in...

Can you write this up with the 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 17.8.9

Report the results in a concluding write-up paragraph.

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 (χ^2(2)=20.31, p<.05). As can be seen in Table 17.8.1, 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 study 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<.05or the probability that nothing is happening in the population based on our sample is less than 5%) 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. However, 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.

Answers

Exercise 17.8.1

With our df = 2 ($df = (2-1)*(3-1) = 1 \times 2 = 2$), the critical value is 5.991.

Exercise 17.8.2

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.

Exercise 17.8.3

χ^2(2)=20.31, p<.05

This page titled 17.8: Practice Test of Independence on College Sports is shared under a CC BY-NC-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.

17.9: Practice Chi-Square on 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 17.9.1 shows a contingency table with the Observed frequencies:

Table 17.9.1: 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 ate sandwiches, but we have the minimum that is required.

Step 1: Hypotheses

We'll start with the research hypothesis. Look at Table 17.9.1. 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 17.9.1

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; $2 \times 2 = 4$).

$$\frac{100}{4} = 25.00$$

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: $P_{EachCell} \neq 0.25$.

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 17.9.1

What is the null hypothesis in words and symbols for this scenario?

Answer available at the end of this section.

Step 2: Collect Data

We're using data from fast food restaurants from OpenIntro.org, so we again did not collect our own data. However, can you answer the questions in Exercise 17.9.2 to show that you understand information Table 17.9.1 is providing?

Exercise 17.9.2

Use the information from Table 17.9.1 to answer the following questions.

1. Who is the sample?
2. Who do might be the population?
3. What two different (qualitative) variables will we be testing?

Answer available at the end of this section.

Step 3: Analyze Data

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 17.9.2

Find the Expected frequencies for the Observed frequencies in Table 17.9.1.

Solution

Table 17.9.2: 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 17.9.1. 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 17.9.3

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 17.9.3: 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 17.9.4

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 17.9.4: 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 17.9.5

Use the Chi-Square formula to calculate the χ^2 statistic.

Solution

Using the information in Table 17.9.1 and Table 17.9.2, 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: Conclusion

To determine if the research hypothesis is supported or not, we first do what? Yes, compare the Chi-Square that we just calculated to a critical value. Let's find the critical value, the easiest step in our process!

Exercise 17.9.3

Using the Critical Value of Chi-Square Table page (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 available at the end of this section.

What do you think when this is still true:

$$\textbf{(Critical} < \textbf{Calculated)} = \textbf{Reject null} = \textbf{There is a pattern of relationship.} \ = p < .05$$

$$\textbf{(Critical} > \textbf{Calculated)} = \textbf{Retain null} = \textbf{There is no pattern of relationship.} \ = p > .05$$

Exercise 17.9.4

Should the null hypothesis be retained or rejected? What should the statistical sentence look like?

Answer available at the end of this section.

We have all we need to write-up a conclusion now.

Step 5: Write-Up

Example 17.9.6

Report the results in a concluding paragraph with the four requirements for reporting results but refer to Table 17.9.1: 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 17.9.1, 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?

Contributors and Attributions

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

Answers

Exercise 17.9.1

- 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\)

Exercise 17.9.2

1. The sample is the 183 meals from a specific fast food company.
2. The population might be all fast food meals from the specific company.
3. The two variables being analyzed are type of meal (adult or kids) and type of sandwich (hamburger or sandwich).

Exercise 17.9.3

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.

Exercise 17.9.4

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.

This page titled 17.9: Practice Chi-Square on Fast Food Meals is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

17.10: Phi Correlation is Not a Correlation

As we we've seen throughout this textbook, if your data are purely quantitative (ratio or interval scales of measurement), then Pearson's is perfect. If your data happens to be rankings or ordinal scale of measurement, then Spearman's is the way to go. And if your data is purely qualitative, then Chi-Square is the way to go. But there's one more cool variation of data that we haven't talked about until now, and that's called *binary* or *dichotomous*.

> **Note**
>
> Look up "binary" or "dichotomous" to see what they mean.

The root of both words (bi- or di-) mean "two" but the Phi (sounds like "fee," rhymes with "reality") correlation actually uses two variables that only have two levels. You might be thinking, "That sounds like a 2x2 factorial design!" but the difference is that a 2x2 factorial design has two IVs, each with two levels, but also has a DV (the outcome variable that you measure and want to improve). The Phi correlation has one of the two variables as the DV and the DV only has two options, and one of the two variables in the IV and that IV only has two levels. Let's see some example variables:

- Pass or Fail
- Pregnant or Not pregnant
- Urban or rural
- Yes or No
- On or Off
- Conservative or Progressive

How does this work into IVs and DVs?

- Does tutor (tutored or not-tutored) affect passing (pass or fail)?
- Does caffeine (coffee or no-coffee) affect energy (sleepy or not-sleepy)?
- Does exercise (exercise or no-exercise) affect weight loss (lose 10+ pounds or do not lose 10+ pounds)?

We are not going to get into the formula for this one, but the formula is built off of the foundation of Chi-Square analyses. The Phi correlation is a unique statistical analysis that is designed for very specific variables. So when you have two binary variables, the Phi correlation is exactly what you should be using!

At this point, you've learned about all introductory statistical analyses, and some that are a little more advanced. The next section helps you to think about which analysis you would use with the different types of variables (and their relationships).

Contributors

- Dr. MO (Taft College)

This page titled 17.10: Phi Correlation is Not a Correlation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

17.11: 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 17.11.1

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 available at the end of this section.

Exercise 17.11.2

Which test would you use to decide whether two qualitative factors factors have a pattern of relationship?

Answer available at the end of this section.

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!

Answers

Exercise 17.11.1

Goodness of Fit test

Exercise 17.11.2

Test of Independence

This page titled 17.11: 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.

18: WRAP UP

18.1: Choosing the Test
18.2: Why did you take this class?

This page titled 18: Wrap Up is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

18.1: Choosing the Test

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.

Choose 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 18.1.1

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 18.1: Choosing the Test is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

18.2: 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 the first chapter? There was a whole section about why this course might be required for you. In this section, we'll review those ideas to see if you agree!

The first two reasons why this class might be required were related to your major:

> *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 18.2.1

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! The following examples are usually not based on specific studies, but some are based on general research findings; in sum, if you find one of these results interesting, use your library's databases to find any empirical research articles that test these variables.

Let's start interpreting!

Example 18.2.2

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 18.2.1

What is the t-test with two samples, other than an independent samples t-test?

Answer available at the end of this section.

Let's try interpreting the results of a dependent samples t-test.

Example 18.2.3

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 18.2.1

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 18.2.2

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 available at the end of this section.

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 18.2.5

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 18.2.3

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 available at the end of this section.

Improved Thinking

The second set of reasons why a math class might be required were:

> *Third reason that you are taking this class: Learning statistical analyses develops critical thinking.*
>
> *Fourth reason that you are taking this class: Understanding how numbers work will help you make predictions and spot inaccurate information.*

Now that you've practiced all of the equations, gone through all of the steps of Null Hypothesis Significance Testing SO MANY times, do you think that you're better able to follow a process and use logic to make decisions! Dr. MO sure hopes so! And after doing SO MANY practice problems, Dr. MO is confident that many of you will have a better sense of numbers. You can apply this in your daily life when looking at sales prices, tipping at restaurants, and when media provides you with data and graphs; do the numbers make sense? Your improved math literacy from the course will help you determine that!

Dr. MO also wrote: "A bonus benefit of learning how to *analyze*, **interpret**, and report data is that 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!

Answers

Exercise 18.2.1

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.

Exercise 18.2.2

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.

Exercise 18.2.3

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!

This page titled 18.2: 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.

Common Formulas

The following formulas are in the order in which you learn about them in this textook. Use the Table of Contents to look for a specific equation.

Descriptive Statistics

Mean

$$\overline{X} = \frac{\sum X}{N} \tag{1}$$

Standard Deviation

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

Which is also: $s = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}} = \sqrt{\frac{SS}{df}}$

Some instructors prefer this formula because it is easier to calculate (but more difficult to see what's happening):

$$\sqrt{\frac{\left(\sum (X^2) - \frac{(\sum X)^2}{N}\right)}{(N - 1)}} \tag{3}$$

z-score

To find the z-score when you have a raw score:

$$z = \frac{X - \overline{X}}{s} \tag{4}$$

To find a raw score when you have a z-score:

$$x = zs + \overline{X} \tag{5}$$

t-tests

One-Sample t-test

These are the same formulas, but formatted slightly differently.

$$t = \frac{(\overline{X} - \mu)}{\left(\frac{s}{\sqrt{n}}\right)} \tag{6}$$

Confidence Interval

$$\text{Margin of Error } = t \times \left(\frac{s}{\sqrt{N}}\right)$$

$$\text{Confidence Interval } = \overline{X} \pm (t \times \left(\frac{s}{\sqrt{N}}\right)) \tag{7}$$

Independent Sample t-test

Unequal N

You can always use this formula:

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{\left[\frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}\right] \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{8}$$

Equal N

You should only use this formula when your two independent groups are the same size (N), meaning the same number of people in each group.

$$\frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\left(\frac{s_1^2}{N_1}\right) + \left(\frac{s_2^2}{N_2}\right)}} \tag{9}$$

Dependent Sample t-test

Conceptual Formula (symbols)

$$\frac{\left(\overline{X}_D \times \sqrt{N}\right)}{s_D} \tag{10}$$

Full Formula

$$\frac{\left(\frac{\Sigma X_D}{N}\right) \times \left(\sqrt{N}\right)}{\sqrt{\left(\frac{\sum\left((X_D - \overline{X}_D)^2\right)}{(N-1)}\right)}}$$

ANOVA

Sums of Squares for Between Groups Designs

Between Groups

$$SS_B = \sum_{EachGroup} \left[\left(\overline{X}_{group} - \overline{X}_T\right)^2 \times (n_{group})\right] \tag{11}$$

Within Groups

$$SS_W = \sum_{EachGroup} \left[\sum \left(\left(X - \overline{X}_{group} \right)^2 \right) \right] \tag{12}$$

Total

$$SS_T = \sum \left[\left(X - \overline{X}_T \right)^2 \right] \tag{13}$$

Tukey's HSD for Pairwise Comparison

$$HSD = q \times \sqrt{\frac{MSw}{n_{group}}} \tag{14}$$

Sums of Squares for Repeated Measures Designs

Between Groups

Same as above.

Participants

$$SS_{Ps} = \left[\sum \left(\frac{(\sum X_{Ps})^2}{k} \right) \right] - \frac{((\sum X)^2)}{N} \tag{15}$$

Within Groups (Error)

$$SS_{WG} = SS_T - SS_{BG} - S_P$$

Total

Same as above.

Pearson's r (Correlation)

The following formulas are the same. Use the first one when you already have the standard deviation calculated.

These are paired data, so *N is the number of pairs*.

SD Already Calculated:

$$r = \frac{\left(\frac{\sum((x_{Each} - \overline{X_x}) \times (y_{Each} - \overline{X_y}))}{(N-1)} \right)}{(s_x \times s_y)} \tag{16}$$

SD Not Calculated:

$$r = \frac{\left(\frac{\sum((x-\overline{X_x})\times(y-\overline{X_y}))}{(N-1)}\right)}{\left(\sqrt{\frac{\sum\left((x-\overline{X_x})^2\right)}{N-1}}\right)\times\left(\sqrt{\frac{\sum\left((y-\overline{X_y})^2\right)}{N-1}}\right)} \tag{17}$$

Regression Line Equation

$$\widehat{\mathrm{Y}} = \mathrm{a} + (\mathrm{b}\times X) \tag{18}$$

a (intercept):

$$\mathrm{a} = \overline{X_y} - (\mathrm{b}\times\overline{X_x}) \tag{19}$$

b (slope):

$$\frac{\sum(Diff_x \times Diff_y)}{\sum(Diff_X{}^2)} \tag{20}$$

In which "Diff" means the differences between each score and that variable's mean.

Pearson's χ^2 (Chi-Square)

$$\chi^2 = \sum_{Each}\left(\frac{(E-O)^2}{E}\right) \tag{21}$$

Expected Frequencies

Goodness of Fit:

$$\frac{N}{k} \tag{22}$$

Test of Independence:

$$E_{EachCell} = \frac{RT\times CT}{N} \tag{23}$$

In which RT = Row Total and CT = Column Total

This page titled Common Formulas is shared under a Public Domain license and was authored, remixed, and/or curated by Michelle Oja.

COMMON CRITICAL VALUE TABLES

The following pages include critical value tables for the analyses covered in this textbook. When needed, how to calculate the degrees of freedom (df) is also included. Sometimes, additional information is provided for context.

Although you can find most of these tables online somewhere, each version of each table is slightly different, so it's best to use the tables of critical values provided to all students in this textbook or by your professor

1.1: Table of Critical Values of z

1.2: Table of Critical Values of t

1.3: Table of Critical Values of F

1.4: Table of Critical Values of r

1.5: Table of Critical Values of Chi-Square

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

1.1: Table of Critical Values of z

You can use this table from Wikimedia Commons (website address: https://commons.wikimedia.org/wiki/File:Tabla_z.png) to find the proportion (multiply the p by 100) of a normal distribution that is more extreme than the z-score. Use the table that is easiest for you! They all have the same numbers.

Table 1.1.1 shows shows **positive** z-scores, their probability (p-value), and percentage of scores that fall *below* that proportion of a standard normal curve. The key to using this table is to understand the the percentages are showing the portion of scores that should be *below* the z-score that you're using. This means that you'll have to do a little bit of math to The figures below the Table 1.1.1 can help to understand what showing that scores below the z-score might mean.

z-Score	p-value	% Below
0.00	0.5000	50.00%
0.01	0.5040	50.40%
0.02	0.5080	50.80%
0.03	0.5120	51.20%
0.04	0.5160	51.60%
0.05	0.5199	51.99%
0.06	0.5239	52.39%
0.07	0.5279	52.79%
0.08	0.5319	53.19%
0.09	0.5359	53.59%
0.10	0.5398	53.98%
0.11	0.5438	54.38%
0.12	0.5478	54.78%
0.13	0.5517	55.17%
0.14	0.5557	55.57%
0.15	0.5596	55.96%
0.16	0.5636	56.36%
0.17	0.5675	56.75%
0.18	0.5714	57.14%
0.19	0.5754	57.54%
0.20	0.5793	57.93%
0.21	0.5832	58.32%
0.22	0.5871	58.71%
0.23	0.5910	59.10%
0.24	0.5948	59.48%
0.25	0.5987	59.87%
0.26	0.6026	60.26%
0.27	0.6064	60.64%
0.28	0.6103	61.03%
0.29	0.6141	61.41%
0.30	0.6179	61.79%
0.31	0.6217	62.17%
0.32	0.6255	62.55%
0.33	0.6293	62.93%

Table 1.1.1*: Positive z-Scores*

z-Score	p-value	% Below
0.34	0.6331	63.31%
0.35	0.6368	63.68%
0.36	0.6406	64.06%
0.37	0.6443	64.43%
0.38	0.6480	64.80%
0.39	0.6517	65.17%
0.40	0.6554	65.54%
0.41	0.6591	65.91%
0.42	0.6628	66.28%
0.43	0.6664	66.64%
0.44	0.6700	67.00%
0.45	0.6736	67.36%
0.46	0.6772	67.72%
0.47	0.6808	68.08%
0.48	0.6844	68.44%
0.49	0.6879	68.79%
0.50	0.6915	69.15%
0.51	0.6950	69.50%
0.52	0.6985	69.85%
0.53	0.7019	70.19%
0.54	0.7054	70.54%
0.55	0.7088	70.88%
0.56	0.7123	71.23%
0.67	0.7157	71.57%
0.58	0.7190	71.90%
0.59	0.7224	72.24%
0.60	0.7258	72.58%
0.61	0.7291	72.91%
0.62	0.7324	73.24%
0.63	0.7357	73.57%
0.64	0.7389	73.89%
0.65	0.7422	74.22%
0.66	0.7454	74.54%
0.67	0.7486	74.86%
0.68	0.7518	75.18%

z-Score	p-value	% Below
0.69	0.7549	75.49%
0.70	0.7580	75.80%
0.71	0.7612	76.12%
0.72	0.7642	76.42%
0.73	0.7673	76.73%
0.74	0.7704	77.04%
0.75	0.7734	77.34%
0.76	0.7764	77.64%
0.77	0.7794	77.94%
0.78	0.7823	78.23%
0.79	0.7852	78.52%
0.80	0.7881	78.81%
0.81	0.7910	79.10%
0.82	0.7939	79.39%
0.83	0.7967	79.67%
0.84	0.7996	79.96%
0.85	0.8023	80.23%
0.86	0.8051	80.51%
0.87	0.8079	80.79%
0.88	0.8106	81.06%
0.89	0.8133	81.33%
0.90	0.8159	81.59%
0.91	0.8186	81.86%
0.92	0.8212	82.12%
0.93	0.8238	82.38%
0.94	0.8264	82.64%
0.95	0.8289	82.89%
0.96	0.8315	83.15%
0.97	0.8340	83.40%
0.98	0.8365	83.65%
0.99	0.8389	83.89%
1.00	0.8413	84.13%
1.01	0.8438	84.38%
1.02	0.8461	84.61%
1.03	0.8485	84.85%

z-Score	p-value	% Below
1.04	0.8508	85.08%
1.05	0.8531	85.31%
1.06	0.8554	85.54%
1.07	0.8577	85.77%
1.08	0.8599	85.99%
1.09	0.8621	86.21%
1.10	0.8643	86.43%
1.11	0.8665	86.65%
1.12	0.8686	86.86%
1.13	0.8708	87.08%
1.14	0.8729	87.29%
1.15	0.8749	87.49%
1.16	0.8770	87.70%
1.17	0.8790	87.90%
1.18	0.8810	88.10%
1.19	0.8830	88.30%
1.20	0.8849	88.49%
1.21	0.8869	88.69%
1.22	0.8838	88.38%
1.23	0.8907	89.07%
1.24	0.8925	89.25%
1.25	0.8944	89.44%
1.26	0.8562	85.62%
1.27	0.8980	89.80%
1.28	0.8997	89.97%
1.29	0.9015	90.15%
1.30	0.9032	90.32%
1.31	0.9049	90.49%
1.32	0.9066	90.66%
1.33	0.9082	90.82%
1.34	0.9099	90.99%
1.35	0.9115	91.15%
1.36	0.9131	91.31%
1.37	0.9147	91.47%
1.38	0.9162	91.62%

z-Score	p-value	% Below
1.39	0.9177	91.77%
1.40	0.9192	91.92%
1.41	0.9207	92.07%
1.42	0.9222	92.22%
1.43	0.9236	92.36%
1.44	0.9251	92.51%
1.45	0.9265	92.65%
1.46	0.9279	92.79%
1.47	0.9292	92.92%
1.48	0.9306	93.06%
1.49	0.9319	93.19%
1.50	0.9332	93.32%
1.51	0.9345	93.45%
1.52	0.9357	93.57%
1.53	0.9370	93.70%
1.54	0.9382	93.82%
1.55	0.9394	93.94%
1.56	0.9406	94.06%
1.57	0.9418	94.18%
1.58	0.9430	94.30%
1.59	0.9441	94.41%
1.60	0.9452	94.52%
1.61	0.9463	94.63%
1.62	0.9474	94.74%
1.63	0.9485	94.85%
1.64	0.9495	94.95%
1.65	0.9505	95.05%
1.66	0.9515	95.15%
1.67	0.9525	95.25%
1.68	0.9535	95.35%
1.69	0.9545	95.45%
1.70	0.9554	95.54%
1.71	0.9564	95.64%
1.72	0.9573	95.73%
1.73	0.9582	95.82%

z-Score	p-value	% Below
1.74	0.9591	95.91%
1.75	0.9599	95.99%
1.76	0.9608	96.08%
1.77	0.9616	96.16%
1.78	0.9625	96.25%
1.79	0.9633	96.33%
1.80	0.9641	96.41%
1.81	0.9649	96.49%
1.82	0.9656	96.56%
1.83	0.9664	96.64%
1.84	0.9671	96.71%
1.85	0.9678	96.78%
1.86	0.9686	96.86%
1.87	0.9693	96.93%
1.88	0.9700	97.00%
1.89	0.9706	97.06%
1.90	0.9713	97.13%
1.91	0.9719	97.19%
1.92	0.9726	97.26%
1.93	0.9732	97.32%
1.94	0.9738	97.38%
1.95	0.9744	97.44%
1.96	0.9750	97.50%
1.97	0.9756	97.56%
1.98	0.9762	97.62%
1.99	0.9767	97.67%
2.00	0.9773	97.73%
2.01	0.9778	97.78%
2.02	0.9783	97.83%
2.03	0.9788	97.88%
2.04	0.9793	97.93%
2.05	0.9798	97.98%
2.06	0.9803	98.03%
2.07	0.9808	98.08%
2.08	0.9812	98.12%

z-Score	p-value	% Below
2.09	0.9817	98.17%
2.10	0.9821	98.21%
2.11	0.9826	98.26%
2.12	0.9830	98.30%
2.13	0.9834	98.34%
2.14	0.9838	98.38%
2.15	0.9842	98.42%
2.16	0.9846	98.46%
2.17	0.9850	98.50%
2.18	0.9854	98.54%
2.19	0.9857	98.57%
2.20	0.9861	98.61%
2.21	0.9865	98.65%
2.22	0.9868	98.68%
2.23	0.9871	98.71%
2.24	0.9875	98.75%
2.25	0.9878	98.78%
2.26	0.9881	98.81%
2.27	0.9884	98.84%
2.28	0.9887	98.87%
2.29	0.9890	98.90%
2.30	0.9893	98.93%
2.31	0.9896	98.96%
2.32	0.9898	98.98%
2.33	0.9901	99.01%
2.34	0.9904	99.04%
2.35	0.9906	99.06%
2.36	0.9909	99.09%
2.37	0.9911	99.11%
2.38	0.9913	99.13%
2.39	0.9916	99.16%
2.40	0.9918	99.18%
2.41	0.9920	99.20%
2.42	0.9922	99.22%
2.43	0.9925	99.25%

z-Score	p-value	% Below
2.44	0.9927	99.27%
2.45	0.9929	99.29%
2.46	0.9931	99.31%
2.47	0.9932	99.32%
2.48	0.9934	99.34%
2.49	0.9936	99.36%
2.50	0.9938	99.38%
2.51	0.9940	99.40%
2.52	0.9941	99.41%
2.53	0.9943	99.43%
2.54	0.9945	99.45%
2.55	0.9946	99.46%
2.56	0.9948	99.48%
2.57	0.9949	99.49%
2.58	0.9951	99.51%
2.59	0.9952	99.52%
2.60	0.9953	99.53%
2.61	0.9955	99.55%
2.62	0.9956	99.56%
2.63	0.9957	99.57%
2.64	0.9959	99.59%
2.65	0.9960	99.60%
2.66	0.9961	99.61%
2.67	0.9962	99.62%
2.68	0.9963	99.63%
2.69	0.9964	99.64%
2.70	0.9965	99.65%
2.71	0.9968	99.68%
2.72	0.9967	99.67%
2.73	0.9968	99.68%
2.74	0.9969	99.69%
2.75	0.9970	99.70%
2.76	0.9971	99.71%
2.77	0.9972	99.72%
2.78	0.9973	99.73%

z-Score	p-value	% Below
2.79	0.9974	99.74%
2.80	0.9974	99.74%
2.81	0.9975	99.75%
2.82	0.9976	99.76%
2.83	0.9977	99.77%
2.84	0.9977	99.77%
2.85	0.9978	99.78%
2.86	0.9979	99.79%
2.87	0.9980	99.80%
2.88	0.9980	99.80%
2.89	0.9981	99.81%
2.90	0.9981	99.81%
2.91	0.9982	99.82%
2.92	0.9983	99.83%
2.93	0.9983	99.83%
2.94	0.9984	99.84%
2.95	0.9984	99.84%
2.96	0.9985	99.85%
2.97	0.9985	99.85%
2.98	0.9986	99.86%
2.99	0.9986	99.86%
3.00	0.9987	99.87%

Table 1.1.1: Positive z-Scores. (CC-BY-SA; modified by Michelle Oja from Jsmura via Wikimedia Commons)

If Table 1.1.1 is too unwieldy, here is a PDF of a z-score table with three sets of each columns (z-score, percent).

Again, this probabilities (p-values) and percentage below show the z-score that you calculated. Figure 1.1.1 and Figure 1.1.2 show what this might look like in a standard normal curve.



Figure 1.1.1: Example of z=1.00 (CC-BY-SA, Ashley Baumann via Wikipedia)

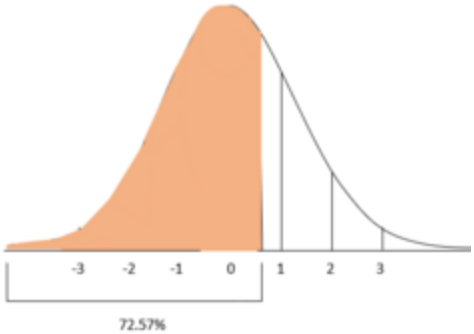


Figure 1.1.2: Example of z=0.60 (CC-BY-SA, Ashley Baumann via Wikipedia)

In Figure 1.1.1, the z-score was calculated to be 1.00. A z-score of 1.00 is one standard deviation above the mean because the z-score is positive. Using the z-table, this shows us that the percentage below that z-score of 1.00 is 84.13%. This percentage is shaded in Figure 1.1.1. Similarly, the shaded area in Figure 1.1.2 for a z-score of 0.60 is almost exactly the "% Below" shown in Table 1.1.1, which is 72.58%. More details on this example can be found on the Standard Score page of Wikipedia.

Negative Scores

Figure 1.1.1 shows a positive z-score of 1.00, but what if the mean was bigger than the target score and we end up with a negative z-score? You can see a z-score for -1 in Figure 1.1.1, but how can you find the percentage of scores that should be below a negative z-score using Table 1.1.1? Well, one way is to use a different z-table that includes negative z-scores (such as this PDF of very long columns, but includes both positive and negative z-scores). Another option is to think about the Standard Normal Distribution, and realize that 100% of all scores will fall somewhere within it. And if we know that 84.13% of scores should be below a z-score of 1.00, then we could figure out that 100% minus 84.13% is 15.87%. This is the percentage of scores that should fall below z=-1.00.

Example 1.1.1

What portion of scores should fall below a z-score of -0.60 (z=-0.60)?

Solution

Looking at Table 1.1.1, we see that the percentage of scores that should score below z=0.60 (a positive z-score) was 72.58%. If we subract that from 100%, then we get the percentage of scores that should fall below a z-score of -0.60.

$$100 - 72.58 = 27.42$$

The portion of scores that should score below z=-0.60 is 27.42%.

Attributions & Contributors

- Jsmura via Wikimedia Commons)
- Dr. MO (Taft College)

This page titled 1.1: Table of Critical Values of z is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

1.2: Table of Critical Values of t

Student's t Distribution

Table 1.2.1, which follows Figure 1.2.1, shows the critical t-score. If the absolute value of your calculated t-score is bigger (more extreme) than the critical t-score, then you reject the null hypothesis. In Figure 1.2.1, the critical t-score is represented by the line; if the absolute value of the calculated t-score is to the left (where the alpha sign is, α), then the null hypothesis should be rejected.

Figure 1.2.1- Upper critical values of Student's *t* Distribution (CC-BY by Barbara Illowsky & Susan Dean (De Anza College) from OpenStax)

Note

Remember

Critical < |Calculated| = Reject null = means are different = p<.05

Critical > |Calculated| = Retain null = means are similar = p>.05

The information in the Note above reminds you how to format the statistical sentence, with underlining meaning to replace the letters with the answers for that specific scenario: t(df)=calculated t-score, p<.05 or t(df)=calculated t-score, p>.05 (depending on whether your calculated t-scores is more extreme than the critical t-score that you find in Table 1.2.1).

Table of Critical Values for Student's t

Table 1.2.1 shows the critical t-scores for different probabilities (p-values) that represent how likely it would be to get a calculated t-scores this big if the sample was really from the population, by the Degrees of Freedom (df, to represent the size of the sample). More information Degrees of Freedom is below the table. If we think that the sample is not from the population, we would expect a larger t-score and want a small p-value.

Table 1.2.1- *Table of Critical t-values*

Degrees of Freedom (df)	p = 0.10	p = 0.05	p = 0.025	p = 0.01
1	3.078	*6.314*	12.706	31.821
2	1.886	*2.920*	4.303	6.965
3	1.638	*2.353*	3.182	4.541
4	1.533	*2.132*	2.776	3.747
5	1.476	*2.015*	2.571	3.365
6	1.440	*1.943*	2.447	3.143
7	1.415	*1.895*	2.365	2.998
8	1.397	*1.860*	2.306	2.896
9	1.383	*1.833*	2.262	2.821
10	1.372	*1.812*	2.228	2.764
11	1.363	*1.796*	2.201	2.718
12	1.356	*1.782*	2.179	2.681

Degrees of Freedom (df)	p = 0.10	p = 0.05	p = 0.025	p = 0.01
13	1.350	*1.771*	2.160	2.650
14	1.345	*1.761*	2.145	2.624
15	1.341	*1.753*	2.131	2.602
16	1.337	*1.746*	2.120	2.583
17	1.333	*1.740*	2.110	2.567
18	1.330	*1.734*	2.101	2.552
19	1.328	*1.729*	2.093	2.539
20	1.325	*1.725*	2.086	2.528
21	1.323	*1.721*	2.080	2.518
22	1.321	*1.717*	2.074	2.508
23	1.319	*1.714*	2.069	2.500
24	1.318	*1.711*	2.064	2.492
25	1.316	*1.708*	2.060	2.485
26	1.315	*1.706**	2.056	2.479
27	1.314	*1.703*	2.052	2.473
28	1.313	*1.701*	2.048	2.467
29	1.311	*1.699*	2.045	2.462
30	1.310	*1.697*	2.042	2.457
40	1.303	*1.684*	2.021	2.423
60	1.296	*1.671*	2.000	2.390
100	1.290	*1.660*	1.984	2.364
∞	1.282	*1.645*	1.960	2.326

Degrees of Freedom

- One-Sample t-test: N-1
- Independent Sample t-test: $N_1 + N_2 - 2$
- Dependent Sample t-test: N-1 (in which N is the *number of pairs*)

Because tables are limited by size, not all critical t-scores are listed. 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 example, if your df = 49. you would use the df row for 40, (so a p=0.05 would have a critical value of 1.684). If your df=55, you would use the df row for 60. It's sorta silly, but, mathematically, any score is closer to df=100 than infinity (∞), so if your sample is more than 100 scores, use df=100.
- Another option is to always we round down. For our example of df=49, we would still use the df row for 40. If your df=55, you would still use the df row for 40. And if your sample is more than 100 scores, use df=100. This option avoids inflating Type I Error (false positives).

Ask your professor which option you should use!

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 1.2: Table of Critical Values of t is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

1.3: Table of Critical Values of F

F Distribution

The tables following Figure 1.3.1 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 1.3.1- 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. $= p > .05$

The information in the Note above reminds you how to format the statistical sentence, with underlining meaning to replace the letters with the answers for that specific scenario: t(df$_{BG}$, df$_{WG}$)=calculated F-score, p<.05 or t(df$_{BG}$, df$_{WG}$)=calculated F-score, p>.05 (depending on whether your calculated t-scores is more extreme than the critical F-score that you find in the following tables).

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. There are three sets of tables, one for each of three probabilities that the null hypothesis is true (1%, 5%, and 10%). This textbook will always use the 5% alpha ($p = 0.05$, but your professor or your circumstances might indicate another alpha level is preferred. All of the 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 5. If you have 6 or more groups (or combinations), use the column for 5. But honestly, if you have more than 5 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!

To use each table, find the degrees of freedom for your denominator (sometimes Within Groups and sometimes called Error) on the first column, then the degrees of freedom of the numerator on the first row; use those to identify what the critical value is. If you're not sure how to calculate the degrees of freedom, the formulas for calculating the degrees of freedom for each (numerator and denominator) for each type of ANOVA are shown below the tables.

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.

Critical Values of F for p=1%

The first table (Table \(\PageIndex{1}\)) shows the critical values when p=0.01. Use the degrees of freedom for the numerator and the denominator to find the critical value; if you're not sure how to calculate the degrees of freedom, the formulas for calculating the degrees of freedom for each (numerator and denominator) for each type of ANOVA are shown below the tables.

Table 1.3.1- *Critical Values of F for* p=0.01

Degrees of Freedom	**Degrees of Freedom for Numerator**				
Degrees of Freedom for Denominator (Error)	1	2	3	4	5
5	16.26	13.27	12.06	11.39	10.97
6	13.75	10.92	9.78	9.15	8.75
7	12.25	9.55	8.45	7.85	7.46
8	11.26	8.65	7.59	7.01	6.63
9	10.56	8.02	6.99	6.42	6.06
10	10.04	7.56	6.55	5.99	5.64
11	9.65	7.21	6.22	5.67	5.32
12	9.33	6.93	5.95	5.41	5.06
13	9.07	6.70	5.74	5.21	4.86
14	8.86	6.51	5.56	5.04	4.69
15	8.68	6.36	5.42	4.89	4.56
16	8.53	6.23	5.29	4.77	4.44
17	8.40	6.11	5.19	4.67	4.34
18	8.29	6.01	5.09	4.58	4.25
19	10.56	8.02	6.99	6.42	6.06
20	8.10	5.85	4.94	4.43	4.10
21	8.02	5.78	4.87	4.37	4.04
22	7.95	5.72	4.82	4.31	3.99
23	7.88	5.66	4.76	4.26	3.94
24	7.82	5.61	4.72	4.22	3.90
25	7.77	5.57	4.68	4.18	3.85
26	7.72	5.53	4.64	4.14	3.82
27	7.68	5.49	4.60	4.11	3.78
28	7.64	5.45	4.57	4.07	3.75
29	7.60	5.42	4.54	4.04	3.73
30	7.56	5.39	4.51	4.02	3.70

Degrees of Freedom	Degrees of Freedom for Numerator				
40	7.31	5.18	4.31	3.83	3.51
50	7.17	5.06	4.20	3.72	3.41
60	7.08	4.98	4.13	3.65	3.34
100	6.90	4.82	3.98	3.51	3.21
200	6.76	4.71	3.88	3.41	3.11
1000	6.66	4.63	3.80	3.34	3.04

Critical Values of F for p=5%

The second table (Table \(\PageIndex{2}\)) shows the critical values when p=0.05. Use the degrees of freedom for the numerator and the denominator to find the critical value; if you're not sure how to calculate the degrees of freedom, the formulas for calculating the degrees of freedom for each (numerator and denominator) for each type of ANOVA are shown below the tables.

Table 1.3.2- Critical Values of F for p=0.05

Degrees of Freedom	Degrees of Freedom for Numerator				
Degrees of Freedom for Denominator (Error)	1	2	3	4	5
5	6.61	5.79	5.41	5.19	5.05
6	5.99	5.14	4.76	4.53	4.39
7	5.59	4.74	4.35	4.12	3.97
8	5.32	4.46	4.07	3.84	3.69
9	5.12	4.26	3.86	3.63	3.48
10	4.96	4.10	3.71	3.48	3.33
11	4.84	3.98	3.59	3.36	3.20
12	4.75	3.89	3.49	3.26	3.11
13	4.67	3.81	3.41	3.18	3.03
14	4.60	3.74	3.34	3.11	2.96
15	4.54	3.68	3.29	3.06	2.90
16	4.49	3.63	3.24	3.01	2.85
17	4.45	3.59	3.20	2.96	2.81
18	4.41	3.55	3.16	2.93	2.77
19	5.12	4.26	3.86	3.63	3.48
20	4.35	3.49	3.10	2.87	2.71
21	4.32	3.47	3.07	2.84	2.68
22	4.30	3.44	3.05	2.82	2.66
23	4.28	3.42	3.03	2.80	2.64
24	4.26	3.40	3.01	2.78	2.62
25	4.24	3.39	2.99	2.76	2.60
26	4.23	3.37	2.98	2.74	2.59
27	4.21	3.35	2.96	2.73	2.57

Degrees of Freedom	Degrees of Freedom for Numerator				
28	4.20	3.34	2.95	2.71	2.56
29	4.18	3.33	2.93	2.70	2.55
30	4.17	3.32	2.92	2.69	2.53
40	4.08	3.23	2.84	2.61	2.45
50	4.03	3.18	2.79	2.56	2.40
60	4.00	3.15	2.76	2.53	2.37
100	3.94	3.09	2.70	2.46	2.31
200	3.89	3.04	2.65	2.42	2.26
1000	3.85	3.00	2.61	2.38	2.22

Critical Values of F for p=10%

The third table (Table \(\PageIndex{3}\)) shows the critical values when p=0.10. Use the degrees of freedom for the numerator and the denominator to find the critical value; if you're not sure how to calculate the degrees of freedom, the formulas for calculating the degrees of freedom for each (numerator and denominator) for each type of ANOVA are shown below the tables.

Table 1.3.2- Critical Values of F for p=0.05

Degrees of Freedom	Degrees of Freedom for Numerator				
Degrees of Freedom for Denominator (Error)	1	2	3	4	5
5	4.06	3.78	3.62	3.52	3.45
6	3.78	3.46	3.29	3.18	3.11
7	3.59	3.26	3.07	2.96	2.88
8	3.46	3.11	2.92	2.81	2.73
9	3.36	3.01	2.81	2.69	2.61
10	3.29	2.92	2.73	2.61	2.52
11	3.23	2.86	2.66	2.54	2.45
12	3.18	2.81	2.61	2.48	2.39
13	3.14	2.76	2.56	2.43	2.35
14	3.10	2.73	2.52	2.39	2.31
15	3.07	2.70	2.49	2.36	2.27
16	3.05	2.67	2.46	2.33	2.24
17	3.03	2.64	2.44	2.31	2.22
18	3.01	2.62	2.42	2.29	2.20
19	3.36	3.01	2.81	2.69	2.61
20	2.97	2.59	2.38	2.25	2.16
21	2.96	2.57	2.36	2.23	2.14
22	2.95	2.56	2.35	2.22	2.13
23	2.94	2.55	2.34	2.21	2.11
24	2.93	2.54	2.33	2.19	2.10

Degrees of Freedom	Degrees of Freedom for Numerator				
25	2.92	2.53	2.32	2.18	2.09
26	2.91	2.52	2.31	2.17	2.08
27	2.90	2.51	2.30	2.17	2.07
28	2.89	2.50	2.29	2.16	2.06
29	2.89	2.50	2.28	2.15	2.06
30	2.88	2.49	2.28	2.14	2.05
40	2.84	2.44	2.23	2.09	2.00
50	2.81	2.41	2.20	2.06	1.97
60	2.79	2.39	2.18	2.04	1.95
100	2.76	2.36	2.14	2.00	1.91
200	2.73	2.33	2.11	1.97	1.88
1000	2.71	2.31	2.09	1.95	1.85

Degrees of Freedom

Between Groups ANOVA (independent groups)

- Numerator (effect of IV): k-1
- Denominator (error): N-k
- Total: N-1

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
- Total: N-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

Table of q-values

The table of q-values needed to complete Tukey's HSD pairwise comparison by hand can be found at this Real-Statistics.com webpage. There is not a separate, dedicated page in this textbook.

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 1.3: Table of Critical Values of F is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by OpenStax.

1.4: Table of Critical Values of r

When using this table, we're following the general pattern of rejecting the null hypothesis when the calculated value is larger than the critical value. The only difference from what we've been doing is the null hypothesis is not about whether group means are similar or not.

Note

(Critical $<$ Calculated) $=$ Reject null $=$ There is a linear relationship. $= p < .05$

(Critical $>$ Calculated) $=$ Retain null $=$ There is not a linear relationship. $= p > .05$

Degrees of Freedom

For Pearson's r, the **Degrees of Freedom are N-2** in which N is the number of pairs of data. This is usually the case because you get two different pieces of information from each participant. For example, if you had 10 people who all provided their height and weight (two different quantitative variables), then N=10, so df=8. If you were comparing weight before and after an exercise program, then you would not use a Pearson's correlation since you have one quantitative variable (weight) in two groups.

The information in the Note above reminds you how to format the statistical sentence, with underlining meaning to replace the letters with the answers for that specific scenario:r(df)=calculated r, p<.05 or r(df)=calculated r, p>.05 (depending on whether your calculated correlation is more extreme than the critical score that you find in Table 1.4.1).

Table of Critical Values of r

Table 1.4.1 is a simplified and accessible version of the table in Real Statistics Using Excel by Dr. Charles Zaiontz. Table 1.4.1 shows the critical scores of Pearson's r for different probabilities (p-values) that represent how likely it would be to get a calculated correlation this extreme if the two variables were unrelated in the population, by the Degrees of Freedom (df) to represent the size of the sample.

Table 1.4.1- *Critical Values for Pearson's r*

Degrees of Freedom (df)	p = 0.1	p = 0.05	p = 0.01
1	0.988	0.997	1.000
2	0.900	0.950	0.990
3	0.805	0.878	0.959
4	0.729	0.811	0.917
5	0.669	0.754	0.875
6	0.621	0.707	0.834
7	0.582	0.666	0.798
8	0.549	0.632	0.765
9	0.521	0.602	0.735
10	0.497	0.576	0.708
11	0.476	0.553	0.684
12	0.458	0.532	0.661
13	0.441	0.514	0.641

Degrees of Freedom (df)	p = 0.1	p = 0.05	p = 0.01
14	0.426	0.497	0.623
15	0.412	0.482	0.606
16	0.400	0.468	0.590
17	0.389	0.456	0.575
18	0.378	0.444	0.561
19	0.369	0.433	0.549
20	0.360	0.423	0.537
21	0.352	0.413	0.526
22	0.344	0.404	0.515
23	0.337	0.396	0.505
24	0.330	0.388	0.496
25	0.323	0.381	0.487
26	0.317	0.374	0.479
27	0.311	0.367	0.471
28	0.306	0.361	0.463
29	0.301	0.355	0.456
30	0.296	0.349	0.449
35	0.275	0.325	0.418
40	0.257	0.304	0.393
45	0.243	0.288	0.372
50	0.231	0.273	0.354
60	0.211	0.250	0.325
70	0.195	0.232	0.302
80	0.183	0.217	0.283
90	0.173	0.205	0.267
100	0.164	0.195	0.254
150	0.134	0.159	0.208
200	0.116	0.138	0.181
300	0.095	0.113	0.148
400	0.082	0.098	0.128
500	0.073	0.088	0.115
700	0.062	0.074	0.097
1000	0.052	0.062	0.081
5000	0.023	0.028	0.036

Because tables are limited by size, not all critical values are listed. For example, if you had 100 participants, your Degrees of Freedom would be 98 (df=N-2=100-2=98=100). However, the table provides df=90 or df=100. 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 r (98), that would mean that we would use the Degrees of Freedom of 100 because 98 is closer to 100 than to 90. That would mean that the critical r-value for r(98) would be 0.194604 for a p-value of 0.05.
- Another option is to always we round down. For our example of N=100, we use the Degrees of Freedom of 90 because it is the next lowest df listed. That would mean that the critical r-value for r(98) would be 0.204968 for a p-value of 0.05. 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 , regardless of which number is listed in the table; the table is used to decide if the null hypothesis should be rejected or retained.

Contributors and Attributions

Dr. MO (Taft College)

& Real Statistics Using Excel by Dr. Charles Zaiontz

This page titled 1.4: Table of Critical Values of r is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.

1.5: Table of Critical Values of Chi-Square

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 $1.5.1$ shows several χ-square distributions for different degrees of freedom.

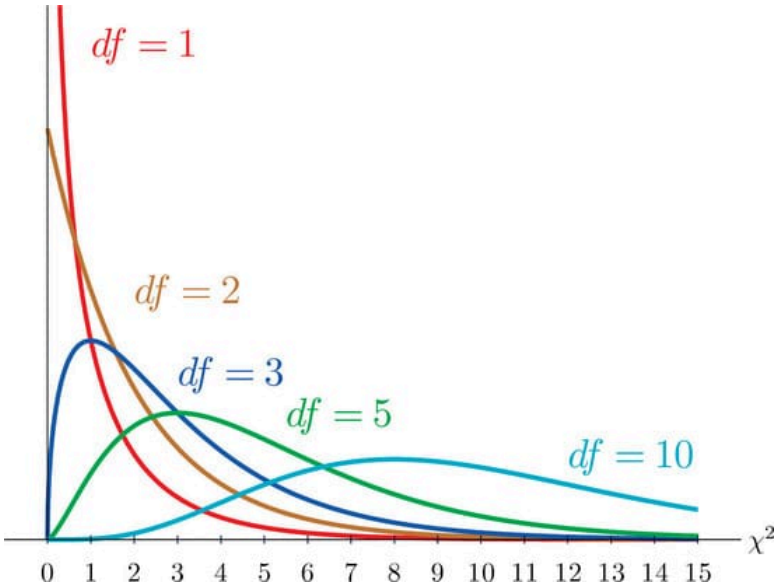


Figure $1.5.1$: Many χ Distributions (CC-BY-NC-SA Shafer & Zhang)

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. $= p > .05$

The information in the Note above reminds you how to format the statistical sentence, with underlining meaning to replace the letters with the answers for that specific scenario:χ^2(df)=calculated χ^2, p<.05 or χ^2(df)=calculated χ^2, p>.05 (depending on whether your calculated Chi-Square is more extreme than the critical score that you find in Table $1.5.1$).

Illustrated in Figure $1.5.2$, 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 $1.5.2$).

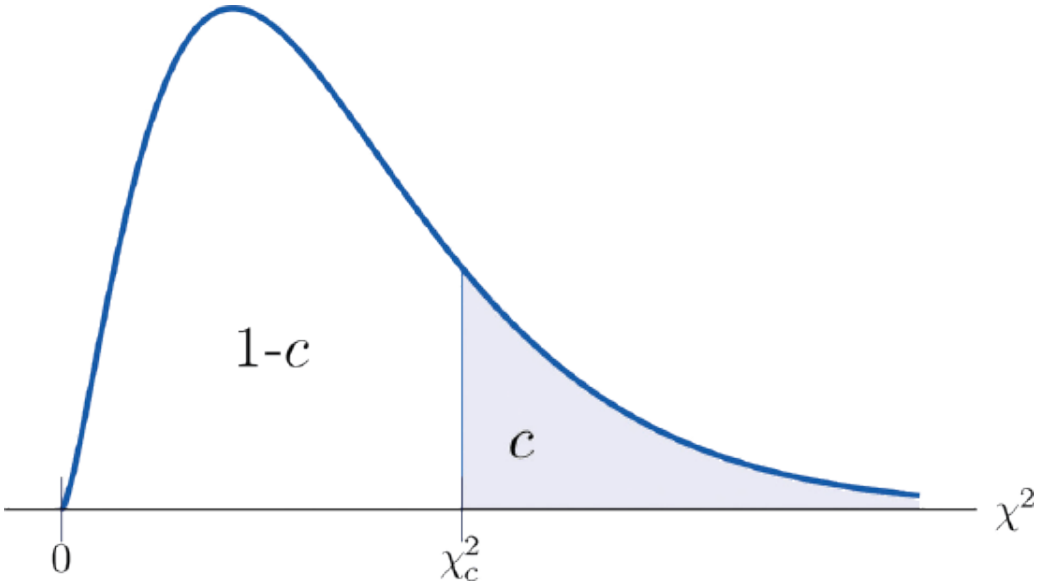


Figure 1.5.2: χ^2_c Illustration of Critical Values of Chi-Square. (CC-BY-NC-SA Shafer & Zhang)

Table of Critical Values for χ^2

Table 1.5.1 below gives values of χ^2 for various p-values and under several chi-square distributions with various degrees of freedom.

Table 1.5.1- 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

df	p = 0.10	p = 0.05	p = 0.01
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 1.5: Table of Critical Values of Chi-Square is shared under a CC BY-NC-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.

INDEX

B

Bonferroni Test
12.7: Pairwise Comparison Post Hoc Tests

box plots
2.11: Graphing Quantitative Data- Boxplots

C

comorbid
15.8: Practice Correlation on Anxiety and Depression

continuous data
1.4.2: Types of Variables

Covariance
15.7: Correlation Formula- Covariance Divided by Variability

D

dependent variable
1.4.1: IV and DV

descriptive statistics
2.1: Introduction to Looking at Data

discrete data
1.4.2: Types of Variables

E

Equal variance
15.6: Hypotheses

F

Frequency Polygons
2.8: Graphing Quantitative Data- Line Graphs

I

independent variable
1.4.1: IV and DV

K

Kruskal–Wallis Test
12.10: Non-Parametric Analysis Between Multiple Groups

kurtosis
2.7: Skew and Kurtosis

L

leptokurtic
2.7: Skew and Kurtosis

line graph
2.8: Graphing Quantitative Data- Line Graphs

line of best fit
16.1: Introduction- Line of Best Fit

linear correlation coefficient
15.6: Hypotheses

linear equations
16.3: Using Linear Equations

M

mean
3.1: Introduction to Measures of Central Tendency
3.3: Mode
3.4: Median
3.5: Mean
3.6: Interpreting All Three

mode
3.1: Introduction to Measures of Central Tendency
3.3: Mode
3.4: Median
3.5: Mean
3.6: Interpreting All Three

N

normal distribution
5.5: Three Important Distributions Together

P

pearson's correlation
15.3: Introduction to Pearson's r

PEMDAS
3.2: Math Refresher

pie chart
2.5: Graphing Qualitative Variables- Pie Charts
6.4: Introduction to the z table

pie charts
2.5: Graphing Qualitative Variables- Pie Charts

Population Standard Deviation
4.3: Practice SD on Wait Times

post hoc test
12.7: Pairwise Comparison Post Hoc Tests

Q

qualitative data
1.4.2: Types of Variables

quantitative data
1.4.2: Types of Variables

R

rounding
3.2.1: Decimals

S

sample standard deviation
4.3: Practice SD on Wait Times

skew
2.7: Skew and Kurtosis

skewness
2.7: Skew and Kurtosis

slope
16.3: Using Linear Equations

standard deviation
4.1: Introduction to Measures of Variability
4.2: Introduction to Standard Deviation Calculations
4.3: Practice SD on Wait Times
4.4: Interpreting Standard Deviations

T

Tukey's Honest Significant Difference (HSD)
12.7: Pairwise Comparison Post Hoc Tests

V

variability
4.1: Introduction to Measures of Variability
4.2: Introduction to Standard Deviation Calculations
4.4: Interpreting Standard Deviations

variance
4.1: Introduction to Measures of Variability
4.2: Introduction to Standard Deviation Calculations
4.4: Interpreting Standard Deviations

This page titled Index is shared under a license and was authored, remixed, and/or curated by LibreTexts.

GLOSSARY

A

Absolute value | Any number converted to a positive value [CC-BY-SA Mattew J. C. Crump]

B

Binary variable | A binary variable is a variable that only has two options (yes or no). Binary variables can be considered quantitative *or* qualitative. [CC-BY Michelle Oja]

C

Confound | A variable that systematically varies along with the independent variable. [CC-BY-NC-SA 4.0 by Rajiv S. Jhangiani, I-Chant A. Chiang, Carrie Cuttler, & Dana C. Leighton]

Continuous variable | A continuous variable is a variable whose value is obtained by measuring, i.e., one which can take on an uncountable set of values. [CC-BY-SA- Wikipedia]

D

Dependent Variable | The variable that you think is the effect (the thing that the IV changes). The DV is the outcome variable, the thing that you want to improve. [CC-BY Michelle Oja]

Descriptive Statistics | Used to describe or summarize the data from the sample. [CC-BY Michelle Oja]

Dichotomous variable | A dichotomous variable is a variable that only has two options (yes or no). Binary variables can be considered quantitative *or* qualitative. [CC-BY Michelle Oja]

Discrete variable | For a discrete variable, the number of permitted values is either finite or countably infinite. [CC-BY-SA- Wikipedia]

Distribution | "Distribution" is often short-hand for "frequency distribution," which is all of the data that you are working with. A distribution is often desplayed in a line graph (frequency polygon) or a bar chart, but could be displayed in a frequency table or as a list of numbers. There are other types of distributions, too. [CC-BY Michelle Oja]

E

Eugenics | A discredited set of beliefs and practices that aim to improve the genetic quality of a human population. Eugenicists often enact white supremacist beliefs by inhibiting the birth of those considered inferior (often through forced sterilization) or promoting the birth of those considered superior. [CC-BY-SA- Wikipedia]

F

Frequency Distribution | A distribution of data showing a count of frequency (how many) for each score or data point. [CC-BY Michelle Oja]

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. [CC-BY Michelle Oja]

I

Independent Variable | The variable that the researcher thinks is the cause of the effect (the DV). The IV is sometimes also called a "predictor" or "predicting variable". [CC-BY Michelle Oja]

Inferential Statistics | Used to make generalizations from the sample data to the population of interest. [CC-BY Michelle Oja]

Interaction | How the levels of two or more IVs *jointly* affect a DV; when one IV *interacts* with the other IV to affect the DV. [CC-BY Michelle Oja]

Interval | Created numerical scale of measurement; type of variable that is measured, and the intervals between each measurement are equal (but zero does not mean the absence of the measured item). [CC-BY Michelle Oja]

K

Kurtosis | A measure of the “tailedness” of the distribution of data (how wide or broad the distribution is) [CC-BY Michelle Oja]

L

Leptokurtic | A tall and narrow distribution of data. [CC-BY Michelle Oja]

M

Main Effect | Any statistically significant differences between the levels of one independent variable in a factorial design. [CC-BY-SA Mattew J. C. Crump]

Math literacy | A combination of statistical literacy and numeracy [CC-BY Michelle Oja]

Mesokurtic | A medium, bell-sharped distribution of data. [CC-BY Michelle Oja]

N

Negative Correlation | When two quantitative variables vary in opposite directions; when one variable increases, the other variable decreases. [CC-BY Michelle Oja]

Negative skew | The scores are bunched to the right, and the thin tail is pointing to the left. [CC-BY Michelle Oja]

Nominal | Scale of measurement that names the variable's levels; type of variable that has a quality or name, but not a number that means something. [CC-BY Michelle Oja]

Non-Parametric Analysis | Statistical analyses using ranked data; used when data sets are not normally distributed or with ranked data [CC-BY Michelle Oja]

Null Hypothesis | A prediction that nothing is going on. The null hypothesis is *always*: 1. There is no difference between the groups' means OR 2. There is no relationship between the variables. [CC-BY Michelle Oja]

Numeracy | The ability to understand, reason with, and apply simple numerical concepts [CC-BY-SA- Wikipedia]

O

Omnibus F | An overall statistical test to find significant mean differences of at least one pair of means when there are more than one pair of means involved. [CC-BY Michelle Oja]

Ordinal | Scale of measurement in which levels have an order; a type of variable that can be put in numerical order. The variables are in ranks (first, second, third, etc.). [CC-BY Michelle Oja]

Outlier | An extreme score, a score that seems much higher or much lower than most of the other scores (There is a technical way to calculate whether a score is an outlier or not, but you don't need to know it.) [CC-BY Michelle Oja]

P

Parameter | Statistic describing characteristics of the population (usually mean and standard deviation of the population) [CC-BY Michelle Oja]

Percentage | Parts per hundred [CC-BY Green]

Platykurtic | A wide and flat distribution of data. [CC-BY Michelle Oja]

Population | The biggest group that your sample can represent. [CC-BY Michelle Oja]

Positive Correlation | When two quantitative variables vary together in the same direction; when one increases, the other one also increases (and when one decreases, the other also decreases) [CC-BY Michelle Oja]

Positive skew | The scores are bunched to the left, and the thin tail is pointing to the right. [CC-BY Michelle Oja]

Q

Qualitative variable | Type of variable that has different values to represent different *categories* or kinds This is the same as the nominal scale of measurement. [CC-BY Michelle Oja]

Quantitative variable | Type of variable that is measured with some sort of scale that uses numbers that measure something. [CC-BY Michelle Oja]

R

Random sample | Sampling where each member of the population has an equal probability of being selected [CC-BY Jhangiani et al. (2019)]

Range | The difference between the highest score and the lowest score in a distribution of quantitative data. [CC-BY Michelle Oja]

Range restriction | When measuring a variable, the full expression of the variable is not assessed. Instead, the top (ceiling) or bottom (floor) are cut off or excluded [CC-BY Michelle Oja]

Ratio | True numerical scale of measurement; type of variable that is measured, and zero means that there is none of it. [CC-BY Michelle Oja]

Research Hypothesis | A prediction of how groups are related. When comparing means, a complete research hypothesis includes:

1. The name of the groups being compared. This is sometimes considered the IV.

2. What was measured. This is the DV.
3. Which group are we predicting will have the higher mean.

[CC-BY Michelle Oja]

Robust | A term used by statisticians to mean resilient or resistant to [CC-BY-NC-SA Foster et al.]

S

Sample | People who participate in a study; the smaller group that the data is gathered from. [CC-BY Michelle Oja]

Sample size | See "sample". Samples are often designated by the letter "n". Lower case "n" is used for groups, and upper case "N" is used for the population or full sample. [CC-BY Michelle Oja]

Sampling | How a researcher recruits participants into their study [CC-BY Michelle Oja]

Sampling bias | In statistics, sampling bias is a bias in which a sample is collected in such a way that some members of the intended population have a lower or higher sampling probability than others. [CC-BY-SA- Wikipedia]

Skew | A distribution in which many scores are bunched up to one side, and there are only a few scores on the other side. [CC-BY Michelle Oja]

Statistic | The results of statistical analyses [CC-BY Michelle Oja]

Statistical analysis | Procedures to organize and interpret numerical information [CC-BY Michelle Oja]

Statistical literacy | The ability to understand and reason with statistics and data [CC-BY-SA- Wikipedia]

This page titled Glossary is shared under a license and was authored, remixed, and/or curated by LibreTexts.

DETAILED LICENSING

Overview

Title: Behavioral Statistics 2e

Webpages: 194

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-SA 4.0: 68.6% (133 pages)
- CC BY-NC-SA 4.0: 20.6% (40 pages)
- Undeclared: 6.2% (12 pages)
- CC BY 4.0: 2.6% (5 pages)
- Unknown License: 1% (2 pages)
- Other: 0.5% (1 page)
- Public Domain: 0.5% (1 page)
- Unknown License: 0% (0 page)

By Page

- Behavioral Statistics 2e - *Unknown License*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Other*
    - Table of Contents - *Undeclared*
    - Acknowledgements - *CC BY-SA 4.0*
    - Licensing - *Undeclared*
    - Licensing - *Undeclared*
    - For Faculty - *CC BY 4.0*
      - 1.1: Changes in Second Edition - *Undeclared*
  - 1: Introduction to Behavioral Statistics - *Undeclared*
    - 1.1: Why are you taking this course? - *CC BY-SA 4.0*
    - 1.2: Making Meaning of Numbers - *CC BY-SA 4.0*
    - 1.3: How Do You Know? - *CC BY-NC-SA 4.0*
      - 1.3.1: The Scientific Method - *CC BY-SA 4.0*
    - 1.4: Types of Data - *CC BY-NC-SA 4.0*
      - 1.4.1: IV and DV - *CC BY-SA 4.0*
      - 1.4.2: Types of Variables - *CC BY 4.0*
      - 1.4.3: Scales of Measurement - *CC BY-SA 4.0*
    - 1.5: Populations and Samples - *CC BY-SA 4.0*
    - 1.6: Collecting Data- More Practice with Populations and Samples - *CC BY-NC-SA 4.0*
    - 1.7: "Research shows that..." - *CC BY-SA 4.0*
    - 1.8: Learning (Statistics) - *CC BY-SA 4.0*
  - 2: What Do Data Look Like? (Graphs) - *CC BY-SA 4.0*
    - 2.1: Introduction to Looking at Data - *CC BY-SA 4.0*
    - 2.2: Frequency Tables - *CC BY-SA 4.0*
    - 2.3: APA Style Tables - *CC BY-SA 4.0*
    - 2.4: Graphing Qualitative Variables- Bar Graphs - *CC BY-SA 4.0*
    - 2.5: Graphing Qualitative Variables- Pie Charts - *CC BY-SA 4.0*
    - 2.6: Graphing Quantitative Variables - *CC BY-SA 4.0*
    - 2.7: Skew and Kurtosis - *CC BY-SA 4.0*
    - 2.8: Graphing Quantitative Data- Line Graphs - *CC BY-SA 4.0*
    - 2.9: Graphing Quantitative Data- Histograms - *CC BY-SA 4.0*
    - 2.10: Graphing Quantitative Data- Means Plot - *CC BY-SA 4.0*
    - 2.11: Graphing Quantitative Data- Boxplots - *CC BY-SA 4.0*
    - 2.12: Graphing Quantitative Data- Scatterplots - *CC BY-SA 4.0*
    - 2.13: Summary and Some Honesty - *CC BY-SA 4.0*
    - 2.14: APA Style Charts - *CC BY-SA 4.0*
  - 3: Measures of Central Tendency - *Undeclared*
    - 3.1: Introduction to Measures of Central Tendency - *CC BY-SA 4.0*
    - 3.2: Math Refresher - *CC BY-NC-SA 4.0*
      - 3.2.1: Decimals - *CC BY 4.0*
    - 3.3: Mode - *CC BY-SA 4.0*
    - 3.4: Median - *CC BY-SA 4.0*
    - 3.5: Mean - *CC BY-SA 4.0*
    - 3.6: Interpreting All Three - *CC BY-SA 4.0*
    - 3.7: Summary of Measures of Central Tendency - *CC BY-NC-SA 4.0*

- 4: Measures of Variability - *CC BY-SA 4.0*
  - 4.1: Introduction to Measures of Variability - *CC BY-SA 4.0*
  - 4.2: Introduction to Standard Deviation Calculations - *CC BY-SA 4.0*
  - 4.3: Practice SD on Wait Times - *CC BY-SA 4.0*
  - 4.4: Interpreting Standard Deviations - *CC BY-SA 4.0*
  - 4.5: Fractions, Decimals, and Percents, Oh My! - *CC BY 4.0*
  - 4.6: Putting All of the Measures Together - *CC BY 4.0*
- 5: Distributions - *CC BY-SA 4.0*
  - 5.1: Introduction to Distributions - *CC BY-SA 4.0*
  - 5.2: Introduction to Probability - *CC BY-NC-SA 4.0*
  - 5.3: The Binomial Distribution - *CC BY-SA 4.0*
  - 5.4: The Law of Large Numbers - *CC BY-SA 4.0*
  - 5.5: Three Important Distributions Together - *CC BY-SA 4.0*
  - 5.6: Sampling Distributions and the Central Limit Theorem - *CC BY-SA 4.0*
  - 5.7: Putting it All Together - *CC BY-SA 4.0*
- 6: Using z - *CC BY-SA 4.0*
  - 6.1: Introduction to z-scores - *CC BY-NC-SA 4.0*
  - 6.2: Calculating z-scores - *CC BY-SA 4.0*
  - 6.3: Practice z-score Calculation on IQ Scores - *CC BY-SA 4.0*
  - 6.4: Introduction to the z table - *CC BY-SA 4.0*
  - 6.5: Practice Using the z Table - *CC BY-SA 4.0*
  - 6.6: Predicting Amounts - *CC BY-SA 4.0*
  - 6.7: Summary of z Scores - *CC BY-SA 4.0*
- 7: APA Style - *CC BY-NC-SA 4.0*
  - 7.1: APA and APA Style - *CC BY-NC-SA 4.0*
  - 7.2: APA Style Resources - *CC BY-NC-SA 4.0*
  - 7.3: General Paper Format - *CC BY-NC-SA 4.0*
  - 7.4: Formatting by Section - *CC BY-NC-SA 4.0*
  - 7.5: Tables and Figures - *CC BY-NC-SA 4.0*
- 8: Inferential Statistics and Hypothesis Testing - *CC BY-SA 4.0*
  - 8.1: Growth Mindset - *CC BY-SA 4.0*
  - 8.2: Samples and Populations Refresher - *CC BY-SA 4.0*
  - 8.3: Can Samples Predict Populations? - *CC BY-SA 4.0*
  - 8.4: Descriptive versus Inferential Statistics - *CC BY-NC-SA 4.0*
  - 8.5: The Research Hypothesis and the Null Hypothesis - *CC BY-NC-SA 4.0*
  - 8.6: Null Hypothesis Significance Testing - *CC BY-SA 4.0*
  - 8.7: Critical Values - *CC BY-SA 4.0*
    - 8.7.1: Rejection Region Visualization - *CC BY-SA 4.0*
  - 8.8: Practice with p-values and NHST - *CC BY-SA 4.0*
  - 8.9: Steps of the Hypothesis Testing Process - *CC BY-SA 4.0*
  - 8.10: The Two Errors in Null Hypothesis Significance Testing - *CC BY-SA 4.0*
  - 8.11: Power and Sample Size - *CC BY-SA 4.0*
  - 8.12: Views of NHST - *CC BY-SA 4.0*
- 9: One Sample t-test - *CC BY-SA 4.0*
  - 9.1: Introduction to One-Sample t-tests - *CC BY-SA 4.0*
  - 9.2: One-Sample t-test Calculations - *CC BY-SA 4.0*
  - 9.3: Reporting Results - *CC BY-SA 4.0*
  - 9.4: Practice One-Sample t-tests on Mindset Scores - *CC BY-SA 4.0*
  - 9.5: Confidence Intervals - *CC BY-SA 4.0*
- 10: Independent Samples t-test - *CC BY-SA 4.0*
  - 10.1: Introduction to Independent Sample t-tests - *CC BY-SA 4.0*
  - 10.2: Independent Samples t-test Equation - *CC BY-SA 4.0*
  - 10.3: Hypotheses with Two Samples - *CC BY-SA 4.0*
  - 10.4: Practice Independent t-test on Movies and Mood - *CC BY-SA 4.0*
  - 10.5: Practice Independent t-test on Growth Mindset - *CC BY-SA 4.0*
  - 10.6: When to NOT use the Independent Samples t-test - *CC BY-SA 4.0*
  - 10.7: Non-Parametric Independent Sample t-Test - *CC BY-SA 4.0*
- 11: Dependent Samples t-test - *CC BY-SA 4.0*
  - 11.1: Introduction to Dependent Samples - *CC BY-SA 4.0*
  - 11.2: Dependent Sample t-test Calculations - *CC BY-SA 4.0*
  - 11.3: Practice Dependent t-test on Job Satisfaction - *CC BY-SA 4.0*
  - 11.4: Practice Dependent t-test on Changes in Mindset - *CC BY-SA 4.0*
  - 11.5: Non-Parametric Analysis of Dependent Samples - *CC BY-SA 4.0*
  - 11.6: Choosing Which Statistic- t-test Edition - *CC BY-SA 4.0*
- 12: BG ANOVA - *CC BY-SA 4.0*
  - 12.1: Why ANOVA? - *CC BY-NC-SA 4.0*
    - 12.1.1: Observing and Interpreting Variability - *CC BY-NC-SA 4.0*
    - 12.1.2: Ratio of Variability - *CC BY-NC-SA 4.0*
  - 12.2: Introduction to ANOVA's Sum of Squares - *CC BY-SA 4.0*
  - 12.3: Summary of ANOVA Summary Table - *CC BY-NC-SA 4.0*
  - 12.4: Hypotheses in ANOVA - *CC BY-NC-SA 4.0*
  - 12.5: Practice BG ANOVA on Job Applicants - *CC BY-NC-SA 4.0*
  - 12.6: Introduction to Pairwise Comparisons - *CC BY-SA 4.0*
  - 12.7: Pairwise Comparison Post Hoc Tests - *CC*

BY-NC-SA 4.0
- 12.8: Practice BG ANOVA and Pairwise Comparisons on Mindset Data - *CC BY-SA 4.0*
- 12.9: ANOVA versus t-test - *CC BY-SA 4.0*
- 12.10: Non-Parametric Analysis Between Multiple Groups - *CC BY-SA 4.0*

◦ 13: RM ANOVA - *CC BY-SA 4.0*
- 13.1: Introduction to Repeated Measures ANOVA - *CC BY-SA 4.0*
- 13.2: RM ANOVA Summary Table - *CC BY-SA 4.0*
- 13.3: Repeated Measures ANOVA Sum of Squares Formulas - *CC BY-SA 4.0*
- 13.4: Practice RM ANOVA Summary Table on Typing Speed - *CC BY-SA 4.0*
- 13.5: Practice RM ANOVA on Mindset Scores - *CC BY-SA 4.0*
- 13.6: Non-Parametric RM ANOVA - *CC BY-SA 4.0*

◦ 14: Factorial ANOVA (Two-Way) - *CC BY-SA 4.0*
- 14.1: Introduction to Factorial Designs - *CC BY-SA 4.0*
- 14.2: Factorial Notations and Square Tables - *CC BY-SA 4.0*
- 14.3: Introduction to Main Effects and Interactions - *CC BY-SA 4.0*
- 14.4: Example with Main Effects and Interactions - *CC BY-SA 4.0*
- 14.5: Graphing Main Effects and Interactions - *CC BY-SA 4.0*
- 14.6: Interpreting Main Effects and Interactions in Graphs - *CC BY-SA 4.0*
- 14.7: Do Main Effects Matter? - *CC BY-SA 4.0*
- 14.8: Beyond 2x2 in Graphs - *CC BY-SA 4.0*
- 14.9: Two-Way ANOVA Summary Table - *CC BY-NC-SA 4.0*
  - 14.9.1: Calculating Sum of Squares for the Factorial ANOVA Summary Table - *CC BY-SA 4.0*
- 14.10: When Should You Conduct Post-Hoc Pairwise Comparisons? - *CC BY-SA 4.0*
- 14.11: Practice Factorial on Attention - *CC BY-SA 4.0*
- 14.12: Practice Factorial on Mindset Scores - *CC BY-SA 4.0*
- 14.13: Choosing the Correct Analysis- Mean Comparison Edition - *CC BY-SA 4.0*

◦ 15: Correlations - *CC BY-SA 4.0*
- 15.1: Refresh to Prepare - *CC BY-SA 4.0*
- 15.2: What do Two Quantitative Variables Look Like? - *CC BY-SA 4.0*
- 15.3: Introduction to Pearson's r - *CC BY-NC-SA 4.0*
- 15.4: Correlation versus Causation - *CC BY-NC-SA 4.0*
  - 15.4.1: Correlation versus Causation in Graphs - *CC BY-SA 4.0*
- 15.5: Strength, Direction, and Linearity - *CC BY-SA 4.0*
- 15.6: Hypotheses - *CC BY-SA 4.0*
- 15.7: Correlation Formula- Covariance Divided by Variability - *CC BY-NC-SA 4.0*
- 15.8: Practice Correlation on Anxiety and Depression - *CC BY-NC-SA 4.0*
- 15.9: Practice Correlation on Nutrition - *CC BY-SA 4.0*
- 15.10: Alternatives to Pearson's Correlation - *CC BY-SA 4.0*
- 15.11: Final Considerations - *CC BY-NC-SA 4.0*

◦ 16: Regression - *CC BY-SA 4.0*
- 16.1: Introduction- Line of Best Fit - *CC BY-NC-SA 4.0*
- 16.2: Regression Line Equation - *CC BY-NC-SA 4.0*
- 16.3: Using Linear Equations - *CC BY-SA 4.0*
- 16.4: Hypothesis Testing- Slope to ANOVAs - *CC BY-NC-SA 4.0*
- 16.5: Practice Regression on Health and Happiness - *CC BY-NC-SA 4.0*
- 16.6: Practice Regression on Nutrition - *CC BY-SA 4.0*
- 16.7: Multiple Regression - *CC BY-NC-SA 4.0*

◦ 17: Chi-Square - *CC BY-SA 4.0*
- 17.1: Introduction to Chi-Square - *CC BY-NC-SA 4.0*
- 17.2: Assumptions of the Test(s) - *CC BY-SA 4.0*
- 17.3: Introduction to Goodness-of-Fit - *CC BY-NC-SA 4.0*
- 17.4: Interpretation of the Goodness-of-Fit Test - *CC BY-SA 4.0*
- 17.5: Goodness of Fit Formula - *CC BY-NC-SA 4.0*
- 17.6: Practice Goodness of Fit on Pineapple on Pizza - *CC BY-NC-SA 4.0*
- 17.7: Introduction to Test of Independence - *CC BY-NC-SA 4.0*
- 17.8: Practice Test of Independence on College Sports - *CC BY-NC-SA 4.0*
- 17.9: Practice Chi-Square on Fast Food Meals - *CC BY-SA 4.0*
- 17.10: Phi Correlation is Not a Correlation - *CC BY-SA 4.0*
- 17.11: Choosing the Correct Test- Chi-Square Edition - *CC BY-SA 4.0*

◦ 18: Wrap Up - *CC BY-SA 4.0*
- 18.1: Choosing the Test - *CC BY-SA 4.0*
- 18.2: Why did you take this class? - *CC BY-SA 4.0*

◦ Back Matter - *Undeclared*
- Common Formulas - *Public Domain*
- Common Critical Value Tables - *CC BY-SA 4.0*
  - 1.1: Table of Critical Values of z - *CC BY-SA 4.0*
  - 1.2: Table of Critical Values of t - *CC BY-SA 4.0*
  - 1.3: Table of Critical Values of F - *CC BY-SA*

4.0
- 1.4: Table of Critical Values of r - *CC BY-SA 4.0*
- 1.5: Table of Critical Values of Chi-Square - *CC BY-NC-SA 4.0*
- Index - *Undeclared*
- Glossary - *Unknown License*
- Detailed Licensing - *Undeclared*
- Detailed Licensing - *Undeclared*

This page titled Behavioral Statistics 2e is shared under a license and was authored, remixed, and/or curated by Michelle Oja.

DETAILED LICENSING

Overview

Title: Behavioral Statistics 2e

Webpages: 194

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-SA 4.0: 68.6% (133 pages)
- CC BY-NC-SA 4.0: 20.6% (40 pages)
- Undeclared: 6.2% (12 pages)
- CC BY 4.0: 2.6% (5 pages)
- Unknown License: 1% (2 pages)
- Other: 0.5% (1 page)
- Public Domain: 0.5% (1 page)
- Unknown License: 0% (0 page)

By Page

- Behavioral Statistics 2e - *Unknown License*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Other*
    - Table of Contents - *Undeclared*
    - Acknowledgements - *CC BY-SA 4.0*
    - Licensing - *Undeclared*
    - Licensing - *Undeclared*
    - For Faculty - *CC BY 4.0*
      - 1.1: Changes in Second Edition - *Undeclared*
  - 1: Introduction to Behavioral Statistics - *Undeclared*
    - 1.1: Why are you taking this course? - *CC BY-SA 4.0*
    - 1.2: Making Meaning of Numbers - *CC BY-SA 4.0*
    - 1.3: How Do You Know? - *CC BY-NC-SA 4.0*
      - 1.3.1: The Scientific Method - *CC BY-SA 4.0*
    - 1.4: Types of Data - *CC BY-NC-SA 4.0*
      - 1.4.1: IV and DV - *CC BY-SA 4.0*
      - 1.4.2: Types of Variables - *CC BY 4.0*
      - 1.4.3: Scales of Measurement - *CC BY-SA 4.0*
    - 1.5: Populations and Samples - *CC BY-SA 4.0*
    - 1.6: Collecting Data- More Practice with Populations and Samples - *CC BY-NC-SA 4.0*
    - 1.7: "Research shows that..." - *CC BY-SA 4.0*
    - 1.8: Learning (Statistics) - *CC BY-SA 4.0*
  - 2: What Do Data Look Like? (Graphs) - *CC BY-SA 4.0*
    - 2.1: Introduction to Looking at Data - *CC BY-SA 4.0*
    - 2.2: Frequency Tables - *CC BY-SA 4.0*
    - 2.3: APA Style Tables - *CC BY-SA 4.0*
    - 2.4: Graphing Qualitative Variables- Bar Graphs - *CC BY-SA 4.0*
    - 2.5: Graphing Qualitative Variables- Pie Charts - *CC BY-SA 4.0*
    - 2.6: Graphing Quantitative Variables - *CC BY-SA 4.0*
    - 2.7: Skew and Kurtosis - *CC BY-SA 4.0*
    - 2.8: Graphing Quantitative Data- Line Graphs - *CC BY-SA 4.0*
    - 2.9: Graphing Quantitative Data- Histograms - *CC BY-SA 4.0*
    - 2.10: Graphing Quantitative Data- Means Plot - *CC BY-SA 4.0*
    - 2.11: Graphing Quantitative Data- Boxplots - *CC BY-SA 4.0*
    - 2.12: Graphing Quantitative Data- Scatterplots - *CC BY-SA 4.0*
    - 2.13: Summary and Some Honesty - *CC BY-SA 4.0*
    - 2.14: APA Style Charts - *CC BY-SA 4.0*
  - 3: Measures of Central Tendency - *Undeclared*
    - 3.1: Introduction to Measures of Central Tendency - *CC BY-SA 4.0*
    - 3.2: Math Refresher - *CC BY-NC-SA 4.0*
      - 3.2.1: Decimals - *CC BY 4.0*
    - 3.3: Mode - *CC BY-SA 4.0*
    - 3.4: Median - *CC BY-SA 4.0*
    - 3.5: Mean - *CC BY-SA 4.0*
    - 3.6: Interpreting All Three - *CC BY-SA 4.0*
    - 3.7: Summary of Measures of Central Tendency - *CC BY-NC-SA 4.0*

- 4: Measures of Variability - *CC BY-SA 4.0*
  - 4.1: Introduction to Measures of Variability - *CC BY-SA 4.0*
  - 4.2: Introduction to Standard Deviation Calculations - *CC BY-SA 4.0*
  - 4.3: Practice SD on Wait Times - *CC BY-SA 4.0*
  - 4.4: Interpreting Standard Deviations - *CC BY-SA 4.0*
  - 4.5: Fractions, Decimals, and Percents, Oh My! - *CC BY 4.0*
  - 4.6: Putting All of the Measures Together - *CC BY 4.0*
- 5: Distributions - *CC BY-SA 4.0*
  - 5.1: Introduction to Distributions - *CC BY-SA 4.0*
  - 5.2: Introduction to Probability - *CC BY-NC-SA 4.0*
  - 5.3: The Binomial Distribution - *CC BY-SA 4.0*
  - 5.4: The Law of Large Numbers - *CC BY-SA 4.0*
  - 5.5: Three Important Distributions Together - *CC BY-SA 4.0*
  - 5.6: Sampling Distributions and the Central Limit Theorem - *CC BY-SA 4.0*
  - 5.7: Putting it All Together - *CC BY-SA 4.0*
- 6: Using z - *CC BY-SA 4.0*
  - 6.1: Introduction to z-scores - *CC BY-NC-SA 4.0*
  - 6.2: Calculating z-scores - *CC BY-SA 4.0*
  - 6.3: Practice z-score Calculation on IQ Scores - *CC BY-SA 4.0*
  - 6.4: Introduction to the z table - *CC BY-SA 4.0*
  - 6.5: Practice Using the z Table - *CC BY-SA 4.0*
  - 6.6: Predicting Amounts - *CC BY-SA 4.0*
  - 6.7: Summary of z Scores - *CC BY-SA 4.0*
- 7: APA Style - *CC BY-NC-SA 4.0*
  - 7.1: APA and APA Style - *CC BY-NC-SA 4.0*
  - 7.2: APA Style Resources - *CC BY-NC-SA 4.0*
  - 7.3: General Paper Format - *CC BY-NC-SA 4.0*
  - 7.4: Formatting by Section - *CC BY-NC-SA 4.0*
  - 7.5: Tables and Figures - *CC BY-NC-SA 4.0*
- 8: Inferential Statistics and Hypothesis Testing - *CC BY-SA 4.0*
  - 8.1: Growth Mindset - *CC BY-SA 4.0*
  - 8.2: Samples and Populations Refresher - *CC BY-SA 4.0*
  - 8.3: Can Samples Predict Populations? - *CC BY-SA 4.0*
  - 8.4: Descriptive versus Inferential Statistics - *CC BY-NC-SA 4.0*
  - 8.5: The Research Hypothesis and the Null Hypothesis - *CC BY-NC-SA 4.0*
  - 8.6: Null Hypothesis Significance Testing - *CC BY-SA 4.0*
  - 8.7: Critical Values - *CC BY-SA 4.0*
    - 8.7.1: Rejection Region Visualization - *CC BY-SA 4.0*
  - 8.8: Practice with p-values and NHST - *CC BY-SA 4.0*
  - 8.9: Steps of the Hypothesis Testing Process - *CC BY-SA 4.0*
  - 8.10: The Two Errors in Null Hypothesis Significance Testing - *CC BY-SA 4.0*
  - 8.11: Power and Sample Size - *CC BY-SA 4.0*
  - 8.12: Views of NHST - *CC BY-SA 4.0*
- 9: One Sample t-test - *CC BY-SA 4.0*
  - 9.1: Introduction to One-Sample t-tests - *CC BY-SA 4.0*
  - 9.2: One-Sample t-test Calculations - *CC BY-SA 4.0*
  - 9.3: Reporting Results - *CC BY-SA 4.0*
  - 9.4: Practice One-Sample t-tests on Mindset Scores - *CC BY-SA 4.0*
  - 9.5: Confidence Intervals - *CC BY-SA 4.0*
- 10: Independent Samples t-test - *CC BY-SA 4.0*
  - 10.1: Introduction to Independent Sample t-tests - *CC BY-SA 4.0*
  - 10.2: Independent Samples t-test Equation - *CC BY-SA 4.0*
  - 10.3: Hypotheses with Two Samples - *CC BY-SA 4.0*
  - 10.4: Practice Independent t-test on Movies and Mood - *CC BY-SA 4.0*
  - 10.5: Practice Independent t-test on Growth Mindset - *CC BY-SA 4.0*
  - 10.6: When to NOT use the Independent Samples t-test - *CC BY-SA 4.0*
  - 10.7: Non-Parametric Independent Sample t-Test - *CC BY-SA 4.0*
- 11: Dependent Samples t-test - *CC BY-SA 4.0*
  - 11.1: Introduction to Dependent Samples - *CC BY-SA 4.0*
  - 11.2: Dependent Sample t-test Calculations - *CC BY-SA 4.0*
  - 11.3: Practice Dependent t-test on Job Satisfaction - *CC BY-SA 4.0*
  - 11.4: Practice Dependent t-test on Changes in Mindset - *CC BY-SA 4.0*
  - 11.5: Non-Parametric Analysis of Dependent Samples - *CC BY-SA 4.0*
  - 11.6: Choosing Which Statistic- t-test Edition - *CC BY-SA 4.0*
- 12: BG ANOVA - *CC BY-SA 4.0*
  - 12.1: Why ANOVA? - *CC BY-NC-SA 4.0*
    - 12.1.1: Observing and Interpreting Variability - *CC BY-NC-SA 4.0*
    - 12.1.2: Ratio of Variability - *CC BY-NC-SA 4.0*
  - 12.2: Introduction to ANOVA's Sum of Squares - *CC BY-SA 4.0*
  - 12.3: Summary of ANOVA Summary Table - *CC BY-NC-SA 4.0*
  - 12.4: Hypotheses in ANOVA - *CC BY-NC-SA 4.0*
  - 12.5: Practice BG ANOVA on Job Applicants - *CC BY-NC-SA 4.0*
  - 12.6: Introduction to Pairwise Comparisons - *CC BY-SA 4.0*
  - 12.7: Pairwise Comparison Post Hoc Tests - *CC*

BY-NC-SA 4.0
- 12.8: Practice BG ANOVA and Pairwise Comparisons on Mindset Data - *CC BY-SA 4.0*
- 12.9: ANOVA versus t-test - *CC BY-SA 4.0*
- 12.10: Non-Parametric Analysis Between Multiple Groups - *CC BY-SA 4.0*

◦ 13: RM ANOVA - *CC BY-SA 4.0*
- 13.1: Introduction to Repeated Measures ANOVA - *CC BY-SA 4.0*
- 13.2: RM ANOVA Summary Table - *CC BY-SA 4.0*
- 13.3: Repeated Measures ANOVA Sum of Squares Formulas - *CC BY-SA 4.0*
- 13.4: Practice RM ANOVA Summary Table on Typing Speed - *CC BY-SA 4.0*
- 13.5: Practice RM ANOVA on Mindset Scores - *CC BY-SA 4.0*
- 13.6: Non-Parametric RM ANOVA - *CC BY-SA 4.0*

◦ 14: Factorial ANOVA (Two-Way) - *CC BY-SA 4.0*
- 14.1: Introduction to Factorial Designs - *CC BY-SA 4.0*
- 14.2: Factorial Notations and Square Tables - *CC BY-SA 4.0*
- 14.3: Introduction to Main Effects and Interactions - *CC BY-SA 4.0*
- 14.4: Example with Main Effects and Interactions - *CC BY-SA 4.0*
- 14.5: Graphing Main Effects and Interactions - *CC BY-SA 4.0*
- 14.6: Interpreting Main Effects and Interactions in Graphs - *CC BY-SA 4.0*
- 14.7: Do Main Effects Matter? - *CC BY-SA 4.0*
- 14.8: Beyond 2x2 in Graphs - *CC BY-SA 4.0*
- 14.9: Two-Way ANOVA Summary Table - *CC BY-NC-SA 4.0*
  - 14.9.1: Calculating Sum of Squares for the Factorial ANOVA Summary Table - *CC BY-SA 4.0*
- 14.10: When Should You Conduct Post-Hoc Pairwise Comparisons? - *CC BY-SA 4.0*
- 14.11: Practice Factorial on Attention - *CC BY-SA 4.0*
- 14.12: Practice Factorial on Mindset Scores - *CC BY-SA 4.0*
- 14.13: Choosing the Correct Analysis- Mean Comparison Edition - *CC BY-SA 4.0*

◦ 15: Correlations - *CC BY-SA 4.0*
- 15.1: Refresh to Prepare - *CC BY-SA 4.0*
- 15.2: What do Two Quantitative Variables Look Like? - *CC BY-SA 4.0*
- 15.3: Introduction to Pearson's r - *CC BY-NC-SA 4.0*
- 15.4: Correlation versus Causation - *CC BY-NC-SA 4.0*
  - 15.4.1: Correlation versus Causation in Graphs - *CC BY-SA 4.0*
- 15.5: Strength, Direction, and Linearity - *CC BY-SA 4.0*
- 15.6: Hypotheses - *CC BY-SA 4.0*
- 15.7: Correlation Formula- Covariance Divided by Variability - *CC BY-NC-SA 4.0*
- 15.8: Practice Correlation on Anxiety and Depression - *CC BY-NC-SA 4.0*
- 15.9: Practice Correlation on Nutrition - *CC BY-SA 4.0*
- 15.10: Alternatives to Pearson's Correlation - *CC BY-SA 4.0*
- 15.11: Final Considerations - *CC BY-NC-SA 4.0*

◦ 16: Regression - *CC BY-SA 4.0*
- 16.1: Introduction- Line of Best Fit - *CC BY-NC-SA 4.0*
- 16.2: Regression Line Equation - *CC BY-NC-SA 4.0*
- 16.3: Using Linear Equations - *CC BY-SA 4.0*
- 16.4: Hypothesis Testing- Slope to ANOVAs - *CC BY-NC-SA 4.0*
- 16.5: Practice Regression on Health and Happiness - *CC BY-NC-SA 4.0*
- 16.6: Practice Regression on Nutrition - *CC BY-SA 4.0*
- 16.7: Multiple Regression - *CC BY-NC-SA 4.0*

◦ 17: Chi-Square - *CC BY-SA 4.0*
- 17.1: Introduction to Chi-Square - *CC BY-NC-SA 4.0*
- 17.2: Assumptions of the Test(s) - *CC BY-SA 4.0*
- 17.3: Introduction to Goodness-of-Fit - *CC BY-NC-SA 4.0*
- 17.4: Interpretation of the Goodness-of-Fit Test - *CC BY-SA 4.0*
- 17.5: Goodness of Fit Formula - *CC BY-NC-SA 4.0*
- 17.6: Practice Goodness of Fit on Pineapple on Pizza - *CC BY-NC-SA 4.0*
- 17.7: Introduction to Test of Independence - *CC BY-NC-SA 4.0*
- 17.8: Practice Test of Independence on College Sports - *CC BY-NC-SA 4.0*
- 17.9: Practice Chi-Square on Fast Food Meals - *CC BY-SA 4.0*
- 17.10: Phi Correlation is Not a Correlation - *CC BY-SA 4.0*
- 17.11: Choosing the Correct Test- Chi-Square Edition - *CC BY-SA 4.0*

◦ 18: Wrap Up - *CC BY-SA 4.0*
- 18.1: Choosing the Test - *CC BY-SA 4.0*
- 18.2: Why did you take this class? - *CC BY-SA 4.0*

◦ Back Matter - *Undeclared*
- Common Formulas - *Public Domain*
- Common Critical Value Tables - *CC BY-SA 4.0*
  - 1.1: Table of Critical Values of z - *CC BY-SA 4.0*
  - 1.2: Table of Critical Values of t - *CC BY-SA 4.0*
  - 1.3: Table of Critical Values of F - *CC BY-SA*

4.0
- 1.4: Table of Critical Values of r - *CC BY-SA 4.0*
- 1.5: Table of Critical Values of Chi-Square - *CC BY-NC-SA 4.0*
- Index - *Undeclared*
- Glossary - *Unknown License*
- Detailed Licensing - *Undeclared*
- Detailed Licensing - *Undeclared*

This page titled Behavioral Statistics 2e is shared under a license and was authored, remixed, and/or curated by Michelle Oja.