DATA AND THE MODERN WORLD



Alan Polansky
Northern Illinois University

LibreTexts™

Data and the Modern World

InfoPage

Your page has been created!

Remove this content and add your own.

Edit page

Click the Edit page button in your user bar. You will see a suggested structure for your content. Add your content and hit Save.

Tips:

Drag and drop

Drag one or more image files from your computer and drop them onto your browser window to add them to your page.

Classifications

Tags are used to link pages to one another along common themes. Tags are also used as markers for the dynamic organization of content in the CXone Expert framework.

Working with templates

CXone Expert templates help guide and organize your documentation, making it flow easier and more uniformly. Edit existing templates or create your own.

Visit for all help topics.

TABLE OF CONTENTS

Licensing

Preface

1: Data and the Modern World

- 1.1: What to Expect
- 1.2: From Data to Decisions
- 1.3: The Use of Data in the Modern World
  - 1.3.1: The Pharmaceutical Industry
  - 1.3.2: The Retail Industry
  - 1.3.3: Sports
  - 1.3.4: The Online World
  - 1.3.5: Education
  - 1.3.6: Meteorology
  - 1.3.7: Insurance
  - 1.3.8: Energy
- 1.4: Some Applications
  - 1.4.1: Voter Engagement
  - 1.4.2: Economic Uncertainty and Corruption
  - 1.4.3: Severity of School Discipline
  - 1.4.4: Product Placement in Movies
  - 1.4.5: The Effect of the COVID Crisis
  - 1.4.6: Attitudes About Marital Rape
  - 1.4.7: Sexuality, Race, and Suicide
  - 1.4.8: Race and Gender in Sentencing Disparity
  - 1.4.9: The Major League Baseball Pitch Clock
  - 1.4.10: Artificial Intelligence in Music Education
- 1.5: About the Quotation
- 1.6: Practice and Exploration
- 1.7: References

2: Empiricism

- 2.1: What to Expect
- 2.2: Empiricism and Knowledge
- 2.3: The Modern Scientific Method
- 2.4: Feminist Empiricism
- 2.5: Critical Analysis
  - 2.5.1: Labor Market Discrimination
  - 2.5.2: Same-Sex Marriage Laws and Hate Crimes
  - 2.5.3: Performance After Surgery in Baseball Players
  - 2.5.4: The Supreme Court and Personal Attitudes
  - 2.5.5: Success Factors for Mobile Work in Health Care
  - 2.5.6: Improving Web Searches
  - 2.5.7: Customer Avoidance of Advertising on Social Media

- 2.6: Summary
- 2.7: About the Quote
- 2.8: Practice and Exploration
- 2.9: Collaborative Activity
- 2.10: References

3: Experiments

- 3.1: What to Expect
- 3.2: Experiments
- 3.3: Designed Experiments
- 3.4: Observational Studies
- 3.5: Case Control Studies
- 3.6: Unethical Experiments
- 3.7: Critical Analysis
  - 3.7.1: Labor Market Discrimination
  - 3.7.2: Visual Exercise Environments and Treadmill Running
  - 3.7.3: Same-Sex Marriage Laws and Hate Crimes
  - 3.7.4: Effects of a New Special Formula Fertilizer on Potato Growth
  - 3.7.5: The Supreme Court and Personal Attitudes
  - 3.7.6: Sleep and College Students
  - 3.7.7: Gender and COVID-19 Outcomes
  - 3.7.8: Thought and the Growth of Fungi Cultures
  - 3.7.9: Risk Factors for Homelessness
- 3.8: Summary
- 3.9: About the Quote
- 3.10: Practice and Exploration
- 3.11: Collaborative Activity
- 3.12: References

4: The Nature of Populations

- 4.1: What to Expect
- 4.2: Defining a Population
- 4.3: Practicality
- 4.4: Established Data
- 4.5: Defining Populations in Litigation
- 4.6: Critical Analysis
  - 4.6.1: The Presidential Election of 1936
  - 4.6.2: Young Women in the Justice System
  - 4.6.3: Phone Use Policies in Schools in England
  - 4.6.4: Same-Sex Marriage and LGBT Discrimination
  - 4.6.5: Sampling Animals in Populations
  - 4.6.6: Land Cover and Racial Segregation
  - 4.6.7: Tobacco Retailer Density
- 4.7: Summary
- 4.8: About the Quote
- 4.9: Practice and Exploration
- 4.10: Collaborative Activity
- 4.11: References

5: The Nature of Data

- 5.1: What to Expect
- 5.2: Quantitative and Qualitative Data
- 5.3: The Structure of Data
- 5.4: Critical Analysis
  - 5.4.1: Inequities in Employment During COVID-19
  - 5.4.2: Attitudes About Marital Rape
  - 5.4.3: Satisfaction with Sport Participation
  - 5.4.4: The Supreme Court and Personal Attitudes
  - 5.4.5: Product Placement in Movies
  - 5.4.6: The Racial Wealth Gap
- 5.5: Summary
- 5.6: About the Quotation
- 5.7: Practice and Exploration
- 5.8: Collaborative Activity
- 5.9: References

6: Measuring Abstract Information

- 6.1: What to Expect
- 6.2: Abstract Concepts
- 6.3: Conceptualization and Operationalization
- 6.4: Validity and Reliability
- 6.5: Established Measurement Systems
- 6.6: Critical Analysis
  - 6.6.1: Attitudes About Marital Rape
  - 6.6.2: Gender and Race-Related Stress
  - 6.6.3: Measuring Satisfaction in Sports
- 6.7: Summary
- 6.8: About the Quotation
- 6.9: Practice and Exploration
- 6.10: Collaborative Activity
- 6.11: References

7: Probability and Randomness

- 7.1: What to Expect
- 7.2: Randomness
- 7.3: Defining Randomness
- 7.4: Probability
- 7.5: Odds
- 7.6: Critical Analysis
  - 7.6.1: The Birthday Problem
  - 7.6.2: Risk of Death from a Shark Attack
  - 7.6.3: Bias in the Application of the Death Penalty
  - 7.6.4: The Vietnam War Draft Lottery
  - 7.6.5: Ovid and Gambling
  - 7.6.6: Sports Activities and Lower Back Pain in Adolescence
- 7.7: Summary

- 7.8: About the Quotation
- 7.9: Practice and Exploration
- 7.10: Collaborative Activity
- 7.11: References

8: Observing Data from Populations

- 8.1: What to Expect
- 8.2: Motivation
- 8.3: The Concept of Sampling
- 8.4: Randomness and Observing Data
- 8.5: More Complicated Sampling Methods
- 8.6: Nonresponse
- 8.7: Critical Analysis
  - 8.7.1: The Achievement Gap
  - 8.7.2: Tobacco Retailer Density
  - 8.7.3: Gender and Race-Related Stress
  - 8.7.4: Sampling Animals in Populations
  - 8.7.5: Sports Activities and Lower Back Pain in Adolescence
  - 8.7.6: Sleep and College Students
  - 8.7.7: Visual Exercise Environments and Treadmill Running
- 8.8: Summary
- 8.9: About the Quotation
- 8.10: Practice and Exploration
- 8.11: Collaborative Activity
- 8.12: References

9: Numerical Methods for Summarizing Data

- 9.1: What to Expect
- 9.2: Summarizing Data
- 9.3: Location
- 9.4: Variation
- 9.5: Percentiles
- 9.6: Critical Analysis
  - 9.6.1: Race and COVID-19 Mortality
  - 9.6.2: Effects of a New Special Formula Fertilizer on Potato Growth
  - 9.6.3: Visual Exercise Environments and Treadmill Running
  - 9.6.4: Satisfaction with Sport Participation
- 9.7: Summary
- 9.8: About the Quotation
- 9.9: Practice and Exploration
- 9.10: Collaborative Activity
- 9.11: References

10: Frequency Distributions

- 10.1: What to Expect
- 10.2: Qualitative Data
- 10.3: Quantitative Data
- 10.4: Cross Classified Data

- 10.5: Simpson's Paradox
- 10.6: Critical Analysis
  - 10.6.1: African American Student Debt
  - 10.6.2: Race and School Discipline
- 10.7: Summary
- 10.8: About the Quotation
- 10.9: Practice and Exploration
- 10.10: Collaborative Activity
- 10.11: References

11: Visual Methods for Summarizing Data

- 11.1: What to Expect
- 11.2: Bar Graphs
- 11.3: Pie Charts
- 11.4: Line Graphs
- 11.5: Histograms
- 11.6: Scatterplots
- 11.7: Critical Analysis
  - 11.7.1: Sampling Animals in Populations
- 11.8: Summary
- 11.9: About the Quotation
- 11.10: Practice and Exploration
- 11.11: References

12: The Logical Structure of Formal Statistical Inference

- 12.1: What to Expect
- 12.2: Hypotheses
- 12.3: Estimation and Evidence
- 12.4: Rejection Regions and Risk

Index

Glossary

Examples Used in the Text

Detailed Licensing

Detailed Licensing

Licensing

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

Preface

About this Open Educational Resource (OER)

This book and its ancillary materials (if applicable) were created by Alan M. Polansky at Northern Illinois University for Basic Statstics and was made possible through an Illinois Secretary of State/Illinois State Library (ISL) OER Grant awarded to Northern Illinois University Libraries' project: *Huskies Unbound: Developing OER for Undergraduate Success in High-Impact Courses*. Funded by a one-time $3 million allocation to the Illinois Secretary of State's Budget in FY24, and facilitated by the Consortium of Academic and Research Libraries in Illinois (CARLI), the ISL OER Grant Program was created to support the development of textbooks and ancillaries as open educational resources to reduce students' costs, encourage student success, and allow faculty to adapt materials to meet local needs.

I wish to express my gratitude to Illinois Governor JB Pritzker, Illinois Secretary of State Alexander Giannoulias, and Greg McCormick, Director of the Illinois State Library, for their commitment to Open Education.

Acknowledgements

The creation of this OER was made possible through the collaborative efforts of educators, librarians, and staff from Northern Illinois University. In particular, I acknowledge grant project members Jaime Schumacher, Missy Clapp, Cynthia Paralejas, Tracy Miller, and Lindsay Vreeland, whose time and expertise were invaluable to its development. I also extend my gratitude to Larissa Garcia and Dee Anna Phares, as co-directors of the grant, for their instrumental contributions in making this project possible and their management throughout the process.

In addition, I offer my appreciation for the training and support provided by Michele Leigh, Elizabeth Clarage, and Nicole Swanson and other CARLI staff members.

Peer Review

The Peer Reviewer process was completed by the following reviewer(s):

- Full name, Rank, Department or Field, Institution

Customization/Terms of Use

This textbook is licensed under a Creative Commons Attribution 4.0 International (CC-BY) license unless otherwise indicated, which means that you are free to:

- SHARE – copy and redistribute the material in any medium or format
- ADAPT – remix, transform, and build upon the material for any purpose

Licensing Terms

The licensor cannot revoke these freedoms as long as you follow the license terms.

- Attribution: You must give appropriate credit, provide a link to the license, and indicate if any changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- No Additional Restrictions: You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.
- Notice: You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable United States copyright exception or limitation.
- No Warranties are Given: The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Attribution

Some of the content for this textbook may have been adapted from other open educational resources. For specific reference information about what was used and/or changed in this adaptation, please refer to the references at the end of each section of the book.

Suggested Attribution Statement for this Book

Content that is not adapted from other OER should include the following attribution statement:

Last name, First name (and if applicable co-author's First and Last name). (2026). *Book Title*. Illinois Secretary of State/Illinois State Library OER Grant: Northern Illinois University. Licensed under CC BY 4.0.

References

This Preface is a derivative of Labatte, Jessica and Larissa Garcia. (2024). Illinois SCOERs. *Fundamentals of Photography and Creative Practice* by Consortium of Academic and Research Libraries in Illinois licensed under CC BY 4.0.

CHAPTER OVERVIEW

1: Data and the Modern World

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."

H. G. Wells

- 1.1: What to Expect
- 1.2: From Data to Decisions
- 1.3: The Use of Data in the Modern World
  - 1.3.1: The Pharmaceutical Industry
  - 1.3.2: The Retail Industry
  - 1.3.3: Sports
  - 1.3.4: The Online World
  - 1.3.5: Education
  - 1.3.6: Meteorology
  - 1.3.7: Insurance
  - 1.3.8: Energy
- 1.4: Some Applications
  - 1.4.1: Voter Engagement
  - 1.4.2: Economic Uncertainty and Corruption
  - 1.4.3: Severity of School Discipline
  - 1.4.4: Product Placement in Movies
  - 1.4.5: The Effect of the COVID Crisis
  - 1.4.6: Attitudes About Marital Rape
  - 1.4.7: Sexuality, Race, and Suicide
  - 1.4.8: Race and Gender in Sentencing Disparity
  - 1.4.9: The Major League Baseball Pitch Clock
  - 1.4.10: Artificial Intelligence in Music Education
- 1.5: About the Quotation
- 1.6: Practice and Exploration
- 1.7: References

This page titled 1: Data and the Modern World is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.1: What to Expect

This chapter discusses the ubiquitous nature of data in the modern world, and how that data is used to make important decisions. Of course, data has always existed, and one does not collect data unless they would like to learn something about the world. It was not until the 17th century that scientists and mathematicians began to systematically study what should be done with all the data that was being collected. By the mid 20th century, the use of formal statistical inference was commonplace in scientific investigations, and the world of business and manufacturing soon found that statistical reasoning was an invaluable part of the business decision-making process. By the late 20th century, the use of statistical methods became embedded in nearly every part of society. The new century witnessed new and incredibly powerful computing technology, and with it the use of new and sophisticated statistical reasoning and advanced computational algorithms that brought the promise of intelligent machine-based decision systems, or artificial intelligence. This chapter provides an overview of these developments and gives examples of how statistical reasoning and computational power are used in the world today.

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

- Discuss the general history of the use of data in the modern world.
- Discuss why knowledge of statistics and data science is important to everyone.
- Discuss examples of the use of data in real world applications.

This page titled 1.1: What to Expect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.2: From Data to Decisions

Not so very long ago the most convenient method for purchasing new media, such as books or music, started with a visit to a local bookstore or music store. Upon entering the store, a customer would browse through the different sections of the store, observing different types of books or music in each. Scanning the shelves, customers would observe the inventory of options available to them and would make purchasing decisions based on what they saw. The in-store experience was enhanced by special product placements and attentive staff members who were available to assist the customer should they need help. With the help of the staff, inventory could be searched, and specific items could be ordered. However, the hallmark of the attentive staff member was that they could respond to questions, offer suggestions, and perhaps even offer alternative items that the customer may enjoy even more than what they were originally interested in. In the best situation, the customer would develop a relationship with the staff members at the store, and the staff would be able to anticipate the needs of the customer.

As the retail industry entered the online digital age, stores sought methods to try to duplicate the in-store experience online. Simple menus allowed customers to browse through books or music selections in a similar way that they would in a regular physical store. In fact, it could be easier to browse online as the customer could enter a few keywords and search the inventory, sorting their results according to price, popularity, and other factors. It was not long until online retailers found other ways to not only replicate many of the advantages of in-store shopping but to enhance the online shopping experience beyond what could be expected in a physical store. Customers soon found that they had access to online reviews from both professional reviewers and other customers. Perhaps the most useful development was the personalization of the shopping experience for each online customer.

Customers soon found that when they logged into their online shopping account, they would be greeted with recommendations, notices of sales, and updates on their orders. What was particularly useful to the customers was that the recommendations and the notice of sales were for items that they might be interested in. That is, these recommendations were different for each customer, and they were specifically targeted to the interest of that customer. Online retailing had entered the modern age.

At first it may seem quite remarkable that an online store would be able to tailor recommendations individually for each customer visiting the store. It is perhaps even more impressive when one considers the fact that some online retailers may have thousands or perhaps even millions of customers. Of course, for the most part these recommendations were not being hand-picked by human staff but were being automatically generated by computer algorithms based on what had been observed about the purchasing and browsing history of the customer. But how does one train a computer to look at customer habits and come up with recommendations about items that they might be interested in purchasing?

The answer lies in being able to gather data on customer purchasing and browsing habits, summarize that data into a usable form, draw conclusions about the customer based on the data summary, and assess potential errors and how often they may happen. The scientific field that addresses all the activities is the field of statistics. Statistics also holds the theoretical basis for many new emerging fields such as data analytics and data science, which combine the theoretical power of statistical analysis with the computational power of modern computing devices.

The study of statistics begins in the 17th century when many European governments found it useful to systematically collect data on the demographics and economic status of their populations. Contributions to the field of statistics in this time consisted mainly of descriptions of data. A typical example of this type of analysis is John Graunt's 1662 work on the Bills of Mortality, which contains statistics on mortality due the plague, a detailed description of the ratio of female to male births, and the development of an early life table that estimates the probability that a person of a given age will live another specified number of years (Fienberg 1992).

By the early 19th century mathematicians and scientists began using the relatively new field of probability theory to look at data in new ways. Carl Friedrich Gauss and Pierre Simon Laplace started to pioneer a mathematical approach to statistics that culminated in the early 20th century with the development of formal procedures for statistical inference. The important feature of statistical inference is the ability to make decisions from data and do so in a way that one can specify the probability of making an error. This is an important distinction from simply describing data. As will be shown throughout this book, the fact that the potential size of errors as well as the rate at which one makes errors when making decisions based on data is a crucial aspect of making data analysis useful.

For example, suppose you are sitting in your comfortable chair at home talking with a friend, looking out on a snowy day, and you would like to know the outside temperature. You friend states that they can estimate the outside temperature by counting the number of chairs in your house, and based on this number your friend estimates the outside temperature to be 17 degrees Fahrenheit. Alternatively, you can look at the trusty thermometer mounted outside your window, which shows the temperature as 27 degrees Fahrenheit. It is important to realize that both methods provide a legitimate estimate of the outside temperature. Though it may seem a little silly, estimating the outside temperature by counting chairs does estimate the outside temperature. The important difference between the two methods is the typical size of their error: the difference between the actual outside temperature and the estimated temperature. The question is how you compute these errors and decide which method to use.

Statisticians in the early 20th century used technical mathematical arguments to formalize the process of looking at errors when using data to make decisions, and by the 1950s these methods were being used in agriculture, medicine, psychology, sociology, and many other fields. The advent of the first accessible digital computers revolutionized the use of statistical procedures in analyzing data, and more sophisticated methods were developed and implemented. By this time statistical methodology was seen as an indispensable part of the scientific method and was applied across almost all fields of study.

By the beginning of the 21st century, the advent of highly advanced computing technology and the development of new algorithms took the theoretical development of statistics to new and exciting places in the field of data science, where new techniques with interesting names like artificial intelligence, machine learning, and deep learning were solving problems that just a few years prior were considered too complex and too large to address.

Statistics and data science are not only useful in the online retail industry. More and more of our lives are becoming directly related to data, and that data is used to decide everything: available medical treatments, the placement of highway and rail systems, insurance premiums, and online gaming and sports odds, just to name a few applications. It is difficult to find an area of society that has not been affected using statistics and data science.

The purpose of this book is to introduce you to the world of statistics and data science, with an emphasis on how data is collected and used to affect your everyday life. Reading this book will not make you a statistician or a data scientist, but it will help you become a good consumer of data. There is not a great amount of technical mathematical information about statistical methodology or computer algorithms in the coming chapters. The emphasis of this book is on a conceptual understanding of the framework of statistical inference. By the end of this book, you should be able to critically analyze the use of data from a conceptual standpoint.

The constant collection and analysis of data means that there is always someone, somewhere, who is trying to convince you of something based on that data. Therefore, as a member of the modern data-driven society, it is imperative that you have the knowledge to critically assess the use of data in your life. This book is intended to help you gain that knowledge. The quote from H.G. Wells at the start of this chapter looked forward to the future. But that future is now: statistical thinking is as necessary for efficient citizenship as the ability to read and write.

This page titled 1.2: From Data to Decisions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.3: The Use of Data in the Modern World

Statistical reasoning is used throughout the modern world. The sections below provide some examples of how statistical reasoning is used in several important industries

This page titled 1.3: The Use of Data in the Modern World is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.3.1: The Pharmaceutical Industry

The pharmaceutical industry has a long history of the use of statistics and data science, and the importance of making data-based decisions in drug development is becoming more important each year. From assessing the effectiveness of proposed treatments for diseases to the development of marketing plans for new treatments, nearly every step of the process that a drug or treatment takes from discovery to final deployment is based on decisions made by using data.

The development of new and innovative medical treatments is a time-consuming and costly process that requires very carefully designed medical experiments. Treatment development often begins with preliminary trials based on either human or nonhuman subjects, depending on the conditional being treated and the type of treatment being administered. The main challenge is how to identify what treatments have potential based on these preliminary trials. The stakes are high as moving a possible treatment from the preliminary stage to later stages is expensive and time-consuming. There is a great deal of risk involved with spending too much time and resources on a treatment that is ultimately not useful or has side effects that outweigh the potential benefits of the treatment.

As a potential treatment moves through the process, the stakes become higher, and the complexity and costs of the experiments increase. Moving a treatment to human trials is a serious proposition. At this point there must be some evidence that the treatment has some potential benefit and is not harmful to the participants. This is particularly true when treating a serious disease for which a standard treatment exists. The decision to give someone a new treatment that may not work as well as a standard treatment for a condition that is life-threatening has serious ethical implications. In the final phase, when a treatment is ready to be released to the public, studies of possible side effects from the treatment provide practitioners and patients with important information that they need to know when considering using the treatment for their condition.

Statisticians and scientists have been working on methodology for performing these experiments for years. Specialized methods are used to determine what treatments might be useful early on, and what conditions are required for a potential treatment to proceed to the next phase of testing. All this methodology is based on data observed from experiments, where the important issue is how much evidence must the data provide before a decision can be made on whether to continue studying a treatment. One of the most important parts of this methodology is the assessment of risk. Of course, there is potential risk in moving an ineffective treatment to the next phase in terms of costs both to human lives and company resources. There is also risk involved with not identifying an effective treatment in terms of lost revenue to the company, and the obvious loss to society of an effective medical treatment.

As the methodology of statistics is developed in this book, the concept of risk is an important issue. Statistical methodologies are not useful solely because they can be used to identify trends and structure within data—they are useful because they include an assessment of risk. In a statistical decision-making process, the practitioner can specify the amount of risk they are willing to tolerate when making certain decisions.

This page titled 1.3.1: The Pharmaceutical Industry is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.3.2: The Retail Industry

The retail industry, like most business-related ventures, has always relied heavily on data of all sorts. The problem of inventory management alone is a problem that is inherent with the retail industry. Store owners require careful planning to know how much inventory to buy and to have on hand at any one time. They must be able to anticipate demand while being aware of the overhead costs associated with keeping inventory in store. Keeping too much inventory can be costly due to the investment and storage of stock that is not selling. Not having enough inventory results in lost sales opportunities. Setting sales prices is another universal problem for the retail industry. Price is not only a function of the cost to the retailer, but also of demand for the product and competition from other stores. A retailer may decide to reduce profit for certain goods by having a sale, in the hopes that customers will buy other more profitable products when visiting the store. Advertising can be costly, and a retailer needs to be able to determine what the net effect may be to the profit margin of the store for each dollar invested in advertising. These problems are decision-making issues based on data. Statistical and mathematical methodologies, often used in conjunction with economic and business principles, have proven themselves useful when solving these problems. It is now not uncommon for an undergraduate degree in business to include at least one course in statistics and data management.

By the 1980s it was not uncommon to have computer-based point-of-sale systems in many retail operations. These systems were often linked, or networked, together so that the business would have access to large amounts of information about daily sales trends. As the computer systems became more complex, so did the data stored by these systems. By the new millennium one could track what individual products were purchased with what other individual products on a regular basis. In many cases, these purchase profiles could be linked to individual customer accounts. While data were inexpensive and easy to collect, it was not always obvious what should be done with the massive amounts of data that retail businesses were collecting.

The advent of advanced statistical and data science algorithms, along with the significant increase in computational power, brought the data to retail decision-makers in a way that was useful and timely. In the modern retail industry, targeted marketing can be specialized down to the level of a single customer. Complex interactive trends in consumer behavior are also being studied at a level never before possible. For example, it is now possible to model to an unprecedented degree how a change in the price of an item will affect store traffic, the sales of the item in question, and the sales of other items that are often bought along with the item that was put on sale.

As the introduction to this chapter demonstrates, the use of data and statistical algorithms are also crucial to the development and effective deployment of online retail operations. The use of advanced statistical algorithms applied to the large amounts of data available to these retailers is used to try to make the online shopping experience as immersive and helpful as shopping in person. In some ways the online experience can be even more helpful than a traditional store as targeted suggestions and sales can be tailored to individual shoppers.

This page titled 1.3.2: The Retail Industry is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.3.3: Sports

From simply keeping score to gathering data to compare players and teams, the use of data and statistical analysis in sports has a long history. The professional sport often most associated with the use of statistics is baseball. Statistics like batting average, number of home runs, number of strike outs, and number of runs batted in have been on the minds of baseball fans for over a century. Ask any baseball fan who their favorite player is and then ask them what makes them a great player. Chances are they will respond with statistics.

In the modern era the use of statistics and data science has produced the field of sports analytics. Sports analytics uses historical statistics to predict player and team performance in a way that can provide a competitive advantage to a team or individual. The results can be used to help players, coaches and staff make decisions both during and prior to sporting events.

Sports analytics was popularized following the release of the 2011 film *Moneyball*, in which a baseball team relies on the use of analytics to build a competitive team on a minimal budget. The essential idea behind this use of sports analytics is based on solving an optimization problem based on past performance data. Given the resources, that is, the allowable budget given to a team to pay player contracts, how should one find players to compose the team such that the winning potential of the team is the largest? The answer is not always obvious. For example, it may not be worth paying a large amount of money for a single high-impact player when the same money could be spent of several medium-impact players.

The use of statistics and data science goes well beyond the application of building a team. For example, in baseball, many analyses have considered the number of runs that are expected for a team under many different types of circumstances. This work traces its beginning back to 1963, where play-by-play data was used to compute an estimate of the average number of runs scored that could be expected in the remainder of the inning for each of the 24 possible states of an inning (Lindsey 1963). Later analyses became more complex and now provide insights into the general strategy of baseball. For example, it has been shown that a sacrifice bunt is rarely advantageous as most teams would be expected to score more runs with a runner on first and no outs than they would with a runner on second and one out (Tango et al.2007).

In American football one of the most prominent recent developments is an estimate of the probability that each team has of winning. This probability is updated after each play for the duration of the game. Many of these estimates are available during online streams. An early approach to this type of analysis used data based on the current score, the point spread, the number of timeouts remaining for each team, and statistics on the status of play to estimate the win probability (Lock and Nettleton 2014). More complicated methods have been developed that use sophisticated statistical modeling (Yurko et al. 2019; Carl and Baldwin 2022).

Of course, statistical methods and mathematical modeling are applied to every conceivable sport including basketball, soccer, ice hockey, rugby and even e-sports (Baumer et al. 2023). With the new prevalence of online wagering in many states, these analytical methods have increased importance not only for those making wagers but also more importantly,for those who set the odds for the games. As more data is available about each game and player, one can expect that the world of sports statistics and analytics will continue to develop and gain importance.

This page titled 1.3.3: Sports is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.3.4: The Online World

The use of network computer technology is ubiquitous in modern society. Nearly every time we use a computer, device, or phone, we are using network technology. Of course we are using digital computer technology at the same time. Computers are basically data crunching machines, and the web pages, games, and entertainment streams are simply data being processed using very sophisticated algorithms. Within these algorithms are methods that rely on statistical and data science methods.

A good example of the influence of statistical and data science methods on the internet can be observed by considering web-based search engines. You and a friend could enter the same terms into a search engine, you on your computer and them on their computer. Chances are that the results that you will see on your search will differ from those of your friend. This discrepancy is because most search engines have been trained using your web history and other online behavior to try to tailor your search results to be most useful for you. The complexity and the size of this problem is enormous. A search must somehow consider finding the pages relevant to your search terms—over the entire known internet—and sort the results so that the most important resources are listed first.

Creating these types of search engines requires methods for assessing the content of web pages and then measuring not only the amount of relevance of the page to the search parameters but also the relevance to the usual requirements of the user based on their search or web histories, as well as the general popularity of the pages based on others search histories. If someone enters the search term “apple,” one can think of two major search results that could result from this search term. The first, of course, is the fruit, while the second is the well-known computer and technology company. Within each of these options the search engine must decide how the relevant pages are ranked and sorted in the output window. All these comparisons are determined using sophisticated statistical and mathematical ideas and algorithms from computer science. Of course, there may be other factors as well. Companies might pay search engines to ensure that their results come up first. This is quite common with many large corporations.

This page titled 1.3.4: The Online World is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.3.5: Education

The field of education has long depended on student performance data to make informed decisions about education policy and pedagogical techniques. One of the most important questions when assessing a pedagogical approach to teaching a subject is determining whether the students learned and retained the material taught during a class. A related question concerns whether changes in pedagogical techniques can result in increased rates of learning and retention. As education researchers have known for quite some time, these studies are very difficult to design and interpret because the concept of learning is very abstract, and hence their research has also considered how to measure the effectiveness of classroom pedagogy on learning in the first place.

Answering these types of questions requires many types of methods from statistics and data science. For example, one might be tempted to conclude that any student who gets a good grade, say a B or better in a course, has demonstrated that they have learned and retained the material taught in the course. But such a conclusion makes many assumptions about how the course grade relates to the work done in the course. One might consider an extreme example where the course grade was only based on attendance. In this case the student has not demonstrated that they have any knowledge of the course material, only that they have attended the course. An exam in a course will only demonstrate a student's knowledge under certain conditions. For example, an instructor could review the actual exam the hour before the class and then give the exam. Or the instructor could do what is known as "teaching to the exam," where class time and student work is focus on passing the exam and not on learning the material. The context under which the exam is given is also important. Students must be given the proper time and atmosphere to demonstrate their knowledge. Therefore, the design of the course assessment materials, as well as the content of the course is very important to assessing whether a student has learned the material or not. The field of statistics has developed methodology that aids with these problems, and it helps researchers get an idea of the true effects that things like teaching materials and different pedagogical approaches can have on student learning.

Prediction is another important, and sometimes controversial, aspect of education research. Those who work in admissions for colleges and universities are tasked with the problem of admitting students who are likely to be successful in their college career. It does neither the college nor the student any favor for them to be admitted to a program in which they are unlikely to succeed. This is the reason that colleges require transcripts and other information about potential students when they apply for admission. The problem is one of prediction, trying to determine what will happen in the future based on what is known at the present. Statistical methods can be used to make such predictions, and these predictions can be quite accurate if sufficient information is known about the present. This is the challenge for colleges and universities, and there have been questions from students and parents as to what type of information is relevant and should be used in the admission process. Current trends in this area include the use of artificial intelligence and data science methods to help answer these questions.

This page titled 1.3.5: Education is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.3.6: Meteorology

The science of forecasting, or predicting, weather conditions is heavily dependent on statistical modeling and data science. Weather forecasting is essentially based on the wealth of historical regional weather data that has been collected for approximately the last one hundred years or so. For short term forecasts the process, from a very simplified viewpoint, is relatively straightforward. For example, one can look at the regional weather conditions today and look back through the historical data for days that had similar regional weather conditions. If we want to predict whether it will rain tomorrow, we can look at how many times there were the same conditions, and it rained the next day. This method may work to some extent, but modern weather forecasting is much more sophisticated.

The need for sophistication and advanced computer algorithms can be appreciated when one thinks about the five-day forecast that is quite common on weather reporting news services. As one seeks to predict conditions much later beyond the present time, the atmospheric conditions over a much wider area becomes important. One can imagine that if we want to predict what the weather will look like in an hour, then relatively local conditions should suffice for making that prediction. A storm system that is 500 miles away is unlikely to have a large immediate effect of the local weather. When one increases the prediction to five days, then even the potential formation of storms 500 miles away might be very important to a local weather forecast.

The potential for large errors will generally increase as the time frame becomes longer. It might be relatively easy to predict what the temperature might be in an hour, and in most circumstances even if there is some error, it might not be very great. This is simply because in most conditions the temperature does not make substantial changes in a brief amount of time. If it is 90 degrees Fahrenheit at 2:00 p.m., it is usually likely to be near that temperature at 2:30 p.m. Now consider the five-day forecast horizon. Predicting the temperature at 2:30 p.m. in five days would be very difficult unless you were able to take in the many possible changes in conditions that might occur in relatively distant regions. Once a prediction is made, any small changes in those remote regions can have a large effect on the outcome. For example, a meteorologist might see a cold front moving in a way that would cross the area in interest in four days. This front obviously has the potential to have a large effect on temperature predictions in local regions on the fifth day. In the meantime, it is entirely possible that the regional conditions may change sufficiently so that the front turns northward and misses the area of interest. Predictions made based on the front will crossing the region on the fourth day will be very far from the actual outcome.

Despite what some individuals might feel personally, the business of weather forecasting has made great technological strides over the past 100 years. The very thought of a five-day forecast one hundred years ago must have seemed impossible. The advent of satellite technology and the seeming ubiquity of weather stations across the globe has provided modern meteorologists with more data than they have ever had before. Sophisticated modeling algorithms based on mathematical and statistical models and advanced data processing algorithms have made predicting the weather more accurate than it has ever been. Of course, prediction is not always perfect, and meteorologists often make mistakes.

One important point that this book emphasizes is that errors based on data cannot be avoided. The key idea is to know how big the errors might be and how often they might occur. Armed with these two bits of information, one can still make informed decisions in the face of uncertainty.

This page titled 1.3.6: Meteorology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.3.7: Insurance

The insurance industry was one of the first large sectors that had a major dependence on at least the idea of uncertainty. The concept of insurance is quite old and is really a form of wagering. An individual might be afraid that an expensive adverse event might occur in the next year. A well-capitalized business associate recognizes the potential of this adverse event but is of the opinion that the event is very unlikely to occur. With this realization the two individuals enter into a business agreement. The individual will pay the business associate a set fee that is much smaller than the potential cost of the adverse event. The agreement is then that if the adverse event occurs, the business associate will cover the cost of the event, thus protecting the individual. However, if the event does not occur, the business associate will keep the fee. Hence it is a kind of a bet. The business associate is essentially betting that the adverse event will not occur, in which case they will get to keep the fee they charged the individual. The individual is essentially betting that the adverse effect will occur, in which case the cost associated with the adverse event is covered.

The most common modern application that most people are familiar with is automobile insurance. In the United States most automobile owners are required to have some minimal type of insurance to cover damage that that might be incurred due to accidents. In most cases the minimum type of required insurance is known as liability insurance, which will pay the cost of injury and damage incurred by the automobile owner, thus protecting the public. In this case the insurance company is betting that you will not have an accident over the coverage period, but in the case that you do, you are covered by the company.

Insurance has a history reaching back to roughly 3000 BCE, with early applications associated with marine insurance against losses of ships or their cargo in places like China and the Middle East and later in the Roman Empire (Dewan 2008; Vaughan 1997). Insurance became far more sophisticated and by the 15th century in Europe, specialized types of insurance were developed. The development of property insurance was precipitated by the Great Fire of London in 1666. In 1681, Nicholas Barbon and his associates established the first fire insurance company to homes (Dickson 1960). Around the same time Edward Lloyd’s coffee house became the meeting place for parties in the shipping industry and those willing to underwrite such ventures, which eventually led to the establishment of the now famous Lloyd's of London. By the 18th century, life insurance policies were being offered by the Amicable Society for a Perpetual Assurance Office, founded in London in 1706. The advent of accident insurance can be traced to the Railway Passengers Assurance Company, which was formed in 1848 in England to insure against fatalities on the railway systems.

Statistical methods allow insurance companies to assess the risk incurred by taking on a policyholder. This risk assessment is then used to determine how much the customer should pay as a premium. It is a well-known statistical property that if insurance companies can effectively spread their risk over many clients, the premiums paid by policyholders will be relatively small when compared to the possible losses by the policyholder. It is this property that makes insurance appealing to those who purchase it; for a relatively small premium they can guard themselves against large catastrophic losses.

The assessment of risk in insurance is based on very large sets of data gathered by the insurance company on the outcomes of previous policyholders and your own previous history. If you have been covered for a long time without any claims on your policy, then this is an empirical indication that the amount of risk associated with your policy is low. To set your rate the insurance company will consider that history along with the risk associated with other policyholders who have a similar profile to you. Other data is also considered. For example, a homeowner who lives in an area that is prone to catastrophic weather events has increased risk as the likelihood of a claim for home damage due to a storm is larger. Similarly, other aspects of your home may decrease risk, such as burglar and fire alarm systems that automatically contact emergency services in the event of an alarm.

Life insurance is based largely on what are known as life tables, which were a logical consequence of the “Bills of Mortality” complied in the 17th century in England. These publications listed and enumerated the number and types of deaths during a year in a particular area (David 1998). A life table is an estimate of the probability that an individual of a given age will live for another specified time. These tables are usually specific to gender and other factors. The estimates of the probabilities can, in principle, be based on looking at historical data. For a simplified example consider a woman who is 40 years old, and you wish to know the probability that that that individual is still alive at age 60. In principle, this probability can be estimated by looking at the proportion of women who were 40 years old 20 years ago and are still alive today. Modern methods use large sets of data on mortality and its causes along with statistical modeling techniques to produce these estimates. These estimates are the basis for setting life insurance rates as well as ensuring that commercial and public retirement plans remain solvent.

This page titled 1.3.7: Insurance is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.3.8: Energy

The modern world runs on energy. Nearly every modern activity from driving an automobile to enjoying the latest applications on phones or tablets, requires the consumption of energy. As a commodity, energy production and consumption have important global political and social consequences. As such, policymakers require a range of data-based information on energy production, consumption, imports, exports, inventories, prices, and environmental effects. Statisticians and data scientists then use these data to provide projections for future activities and analyses of the possible effects of changes in technology, regulations, and tax policies to promote sound policymaking, efficient markets, and public understanding of energy and its place in the economy and environment.

The amount of energy produced and consumed has serious implications for the price of energy as well as issues such as national security due to possible disruptions in energy supplies. Data on energy production and consumption is complicated and in application only statistical estimates of these quantities are often available in the short term. This is because it is impractical to measure energy consumption for millions of homes and businesses. Therefore, consumption must be estimated from measures such as decreases in inventories at fuel storage points.

Beyond the simple tracking of energy consumption and production, prediction is crucially important to developing energy distribution systems and studying the effect of energy efficiency programs. For example, it is important to know how much energy is needed to satisfy demand in the next 30 years. Of equal importance is the closely related question of what energy supplies will be available to meet this demand. These issues have profound implications on the expected costs of energy, the industries that will be profitable, and the modes of transportation that will be economical.

This page titled 1.3.8: Energy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.4: Some Applications

General information about how statistics and data are used in the real world provide an overview of the type of applications that occur in practice. The examples of the previous section give only a hint of what statistics and data science can achieve when applied across many fields. In this section some specific applications are considered in a little more detail. Each of the applications highlighted below is based on formal studies that have been published in professional and academic journals. Of course, other types of applications exist—ones that are used internally for businesses, for example. It is often difficult to obtain specific details about these applications and therefore this book will mainly rely on published studies for examples.

This page titled 1.4: Some Applications is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.4.1: Voter Engagement

Participation in the democratic process is often considered one of most important duties of citizenship in the United States. Abundant research considers how social position, race, gender, and education level are associated with voter turnout (Fraga 2018; Verba et al. 1995; Wolfinger and Rosenstone 1980; Whitaker et al. 2008; Cuevas-Molina et al. 2022; Stauffer et al. 2022). The classification of individuals based on gender, race, income, and education level may also have an interactive effect on voter turnout. What this means is that the difference in voter participation by income may vary by race or gender. For example, research has shown that African Americans are disproportionately poor and working class (Branch and Jackson 2020; Laurison et al. 2020; Granderson 2022). Hence, when one considers the question of voter participation, there may be differences in the effect of income due to the societal imbalance in the respective populations.

In 2021 the results of a research study that further broke down voter participation rates by cross-sectional sub-classes of these characteristics considered this question in detail (Laurison, Brown and Rastogi 2021). The data used for this study are based on a survey by the Cooperative Congressional Election Study, an internet-based survey that matches survey responses to voter records to determine voter demographics and participation. This survey interviews respondents between September and November each presidential election year, with about 50,000 respondents each year. For this study the data from 2008–2016 was used, yielding approximately 150,000 responses.

The data provide an interesting look at how voter participation rates vary between these characteristics, and more interestingly how the characteristics interact with one another. For example, the data show that people in higher income groups tend to participate in elections at a higher rate than those in lower income groups, but that the effect differs depending on race. For example, both African American and white voters had a participation rate slightly less than 55% when their income was lower than $30,000 annually. When the income is increased to over $100,000 the voter participation rate is nearly 70% for white voters, while for African American voters the participation rate increased only to about 60%. Therefore, voter participation rates increase as incomes grow for both populations, but at a much slower rate for African American voters. This discrepancy creates a widening gap in voter participation between the two race groups at the higher income levels. If the data are not broken down by race this trend is invisible not only because race is not included, but also because of the relative lack of African American voters in the largest income class.

Race also tends to interact with gender in voting participation. The results of the study show that both African American men and white men vote at about the same rate, slightly above 50%. While it has been known that women vote at a higher rate, there is an interaction with the race of the voter. The study reports that while more white women vote than men, around 55%, the voting rate for African American women exceeds 60%. Hence the effect of gender is more pronounced for African American women than for white women.

The researchers who published the study point out that these results are not only important for understanding the results of elections but are also vitally important when addressing participation rates across all races, incomes, and genders. Indeed, one of the conclusions considered in the research is whether additional personal resources necessarily and universally imply greater voter participation. The results of this study show that there does not appear to be any such universal relationship, as the effect was not observed for some groups.

This page titled 1.4.1: Voter Engagement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.4.2: Economic Uncertainty and Corruption

Government corruption, particularly within some areas of Africa, has been a source of conflict that can lead to increased rates of poverty and wider levels of income inequality (Durojaye et al. 2025). Economic research of this type often focuses on African countries because corruption there is largely systemic. Countries are often undergoing a combination of economic restructuring and have weak institutions, and the private sector tends to be relatively small and weak as compared to economies elsewhere. One study used data from African countries to investigate the effects of corruption on economic growth and the distribution of income (Gyimah-Brempong 2002). This study found that corruption decreases economic growth through decreased investment in physical capital. That is, the money that would have been invested in physical capital was instead being used to bribe officials within the government. Statistical methods used by the researchers demonstrated that a unit increase in corruption reduces the growth rates of gross domestic product and per capita income by between 0.75 and 0.90 percentage points and between 0.39 and 0.41 percentage points per year, respectively. What this means is that as corruption increases, the rate of growth of the economy in a country slows. Secondly, increased corruption is associated with higher rates of income inequality. These results suggest that corruption is more detrimental to the poor than the rich in the African countries considered in the study (Gyimah-Brempong 2002).

This page titled 1.4.2: Economic Uncertainty and Corruption is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.4.3: Severity of School Discipline

Discipline practices in schools affect the quality of the educational environment, and the ability of children to achieve academic and social success. A recent study documented patterns of discipline referrals in 364 elementary and middle schools during the 2005–2006 academic year (Skiba et al. 2011). Data for this study were reported by school personnel by uploading office discipline referrals to a database. An analysis of the data using statistical methods was used to show that students from African American families are roughly two to three times more likely to be referred to the office for problem behavior as their white peers. Additionally, students from African American and Latino families are more likely than their white peers to receive expulsion or out-of-school suspension as consequences for the same or similar problem behavior.

This type of study provides a school system with hard evidence of a systemic problem. It is certainly true that individual students or parents may feel that individual cases of discipline may have been motived by either an explicit or implicit bias. But individual cases are difficult to interpret because of the presence of complicated mitigating factors that may be involved with individual discipline decisions. An analysis based on the careful use of statistical methodology can remove the effect of these factors by providing a fair comparison between discipline rates. This yields strong conclusions about the presence of bias in school discipline that goes beyond what may be an individual bias. The research shows that there is a systemic problem throughout the entire school system. Such a study can also provide useful insights into potential solutions for the problems.

This page titled 1.4.3: Severity of School Discipline is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.4.4: Product Placement in Movies

It is not uncommon to be watching a movie and realize that your favorite actor is using a particular brand of product. In some cases, the product might be in the background, or may flash across the screen rather quickly, while in other cases the use of the product is obvious. Product placement in visual media has existed for a long time in television, movies, and online streaming media sometimes with some controversy. For example, Rod Serling, the creator of the popular television anthology series The Twilight Zone, which ran from 1959 to 1964, was an early critic of the practice (Bartlett 2023).

One study considered the reactions of a sample of 103 individuals who viewed 11 different movies with 18 different product placements (d'Astous and Chartier 2000). Each participant in the study viewed the product placements in a random order. After viewing each product placement, each participant rated their reaction to the placement using a list of 27 phrases that included terms like "aggressive," "amusing," "original," "related to film context." The participants also answered general questions about how often they view movies, their general view of product placement, and some demographic information. The participants were then interviewed a week later and asked what brands and products they recall from the movie placements.

Using statistical analysis techniques, the researchers were able to show that some characteristics of product placements have a positive effect while others have a negative effect. In fact, some characteristics can have either positive or negative effects based on other factors, according to the study. For example, they found that memory of a product placement is enhanced when the principal actor is present, and the placement is positively evaluated. They also found that memory of a product placement declined when the placement is judged as unacceptable and when it is well integrated into a scene.

Results like these can be very useful to companies who wish to place their products in movies for the purpose of advertising. Such companies do not wish to present their product in a way that is either ineffective or produces a negative response. Companies want to ensure that their advertising budget is used in a way the produces positive and effective results.

This page titled 1.4.4: Product Placement in Movies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.4.5: The Effect of the COVID Crisis

The COVID pandemic was the cause of not only a health crisis but also devastating economic effects. While economic opportunities have historically been different for those of different races, these dual crises had a disproportionate effect on African American families. The disparity of these effects can be traced back to pre-COVID conditions, including lower income, higher unemployment, and greater levels of food and housing insecurity for African American families (Hardy and Logan 2020; Bowen et al. 2021; O'Reilly et al. 2020). At the same time, African Americans suffer from health disparities such as a higher rate of preexisting health conditions (Benitez et al. 2020; Wiemers et al. 2020; Albert et al. 2021). The combined interaction of these effects has contributed to higher COVID-19 mortality rates for African Americans than for white Americans, but COVID will still be only the third leading cause of death for African Americans for that same tim period (APM Research Lab 2020).

From an economic viewpoint, historical discrimination and exclusion from education, housing, and economic opportunities for African Americans has caused an economic gap compared to white Americans that has been widening since the 1950s. African American workers are also more likely to have lower wages, more volatile work schedules, less accumulated wealth, and more volatile income (Gould and Wilson 2020; Hardy et al. 2018; Schneider and Harknett 2019; Ziliak et al. 2011; Darity and Mullen 2020; Emmons et al. 2018; Pfeffer and Killewald 2018; Kraus et al. 2019). These differences appear to be largely due to labor discrimination as African Americans tend to have lower returns from post-secondary education (Bertrand and Mullainathan 2004; Holzer et al. 2006; Bartik and Hershbein 2018). As a result of these differences, African American families were less likely to have the financial means to absorb the economic shock of the pandemic, which may have included adverse events such as layoffs, unemployment, and health crises.

With African American families already at an economic disadvantage from the onset of the pandemic, the type of employment is also a contributing factor for the health and welfare of this population, who are more likely to be employed in low-wage front-line and essential worker positions, including the health care support industry, than their white American counterparts. Therefore, these workers faced a higher likelihood of exposure to COVID. A similar effect applied to Hispanic workers as well (Gould and Wilson 2020).

The consequences for African American families have been devastating and could result in generational long-term losses of wealth and family stability (Alaimo et al. 2001; Brennan et al. 2011; Hardy and Marcotte 2020). In the short term, African American communities suffered from higher rates of COVID cases (Benitez et al. 2020; Jahromi and Hamidianjahromi 2020; Fouad et al. 2020; Yancy 2020; Reyes 2020; Wright and Merritt 2020; Kandil et al. 2020).

Families in these neighborhoods also faced higher rates of food and housing insecurity, including households with children. When schools closed to in-person instruction, these same households faced difficult choices that essentially weighed the cost of unemployment against the cost of possibly affordable but unsafe childcare options. College-age children were faced with the tough decision of suspending their higher education opportunities in favor of working or providing childcare to help support their families. Housing insecurity may be particularly troublesome for these communities as eviction moratoriums and loan forgiveness agreements expired in communities that were economically strained by the pandemic.

All of the conclusions discussed above are the result of care statistical data analyses on very complex and often very large sets of data from multiple sources. Without the aid of the statistical techniques and the data would be far too complex to be analyzed in any informal way. Similarly, without the associated computer algorithms the data would be impossible to collect and aggregate. Statistics and data science are crucial to the large health studies like those mentioned above.

This page titled 1.4.5: The Effect of the COVID Crisis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.4.6: Attitudes About Marital Rape

Rape is a substantial social problem that affects millions of people worldwide regardless of race, gender, class, and sexual identity (Ferro et al. 2008; Myhill et al. 2002). Rape is also one of the most underreported crimes in the United States, partially because many victims are intimately related to the offender (Tjaden and Thoennes 2006; Allen 2007). The issue of marital rape is often overlooked, though recent statistics suggest that 10–14% of married women are raped by their husbands in the United States (Bergen and Barnhill 2006). In fact, it was not until 1993 that marital rape became a crime in all 50 states. However, by 2005, only 20 states had eliminated marital rape exemptions from state laws while in the remaining 30 states, some exemptions given to husbands from rape prosecution remained (Bergen and Barnhill 2006). The signing of the Violence Against Women Act of 2005 into law led to systematic and widespread changes in state rape laws.

To study the current attitudes about marital rape, an internet survey was conducted on 85 undergraduate students and 44 alumni from a small liberal arts college. Of the undergraduate participants, 68% identified as female and 55% of the alumni identified as female. The undergraduate participants had varied ethnicities while all the alumni described themselves as white. The survey consisted of short accounts that described a rape scene that varied in only the victim-offender relationship. Some randomly selected participants had accounts that described the offender as a neighbor, while the remaining participants shared an account that described the offender as their spouse. These accounts were rated by the participants using the Illinois Rape Myth Acceptance Scale, which is designed to assess the participants' acceptance of rape myths. The purpose of the study was to determine if the relationship between the victim and the offender influenced how the participants rated the scenario described in the account.

Statistical techniques used on the observations from the study were used to show that participants who read the account that described the offender as a spouse were more likely to endorse false beliefs about rape than the participants who read the account that described the offender as a neighbor. Further, male participants were more likely than female participants to suggest that rape might be unintentional on the part of the perpetrator. Similarly, undergraduate participants were significantly more likely than alumni to suggest that rape might be unintentional on the part of the perpetrator (Ferro et al. 2008). Overall, while there were differences between the groups, the authors of the study found that the perceptions of rape of the participants are generally more realistic than were expected. The results of the study show that, overall, participants are not very supportive of false beliefs about rape and are generally sympathetic toward victims of rape. However, the participants were still likely to endorse certain types of rape myths such as the extent to which the offender should be held responsible for the rape, and where, when, and to whom rape can occur. Furthermore, the participants were more willing to rationalize sexual assault as accidental or uncontrollable on the part of the male perpetrator.

This page titled 1.4.6: Attitudes About Marital Rape is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.4.7: Sexuality, Race, and Suicide

In 2017 suicide was the leading cause of death in the United States with the rate increasing between 2005 and 2015 in nearly every geographic classification (Alvarez-Hernandez and Mowbray 2022; Hedegaard et al. 2018). In the Hispanic community between 1999 and 2017 the age-adjusted mortality rate due to suicide among Hispanic females increased about 40% and remained consistently high for Hispanic men at 11.2 per every 100,000 deaths (Curtin et al. 2016). Although mental health and substance use are factors associated with suicide, death by suicide is also a major concern for populations who experience discrimination. Hispanic individuals, as well as people who are in the LGBTQIA+ community, experience high levels of stress due to discrimination and violence from their social environment (Wang et al. 2021).

A study based on data from the National Violent Death Reporting System, a data gathering system managed by the Division of Violence Prevention at the National Center for Injury and Prevention Control, considered violent death data for Hispanic individuals between 2012 and 2016 for 32 states (Mercy et al. 2006; Steenkamp et al. 2006; Blair et al. 2016). The data showed that there is an association between manner of death and sexual orientation, where the percentage of LGBTQIA+ individuals was higher among those whose manner of death was suicide than those whose manner of death was not suicide.

Furthermore, the average age of individuals whose manner of death was suicide was higher when compared than those who died in other ways. Gender and manner of death were also associated in that women showed lower proportions among those who committed suicide compared to those who died in other ways. Finally, there was an association between education and manner of death where individuals with some college and individuals with a college degree or higher were over-represented among those whose manner of death was suicide. The study also found other links between mental health problems, alcohol related problems and rate of suicide. More advanced statistical modeling techniques were also used to show that there was an increasing trend in these rates over time (Alvarez-Hernandez and Mowbray 2022).

This page titled 1.4.7: Sexuality, Race, and Suicide is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.4.8: Race and Gender in Sentencing Disparity

The prison population in the United States is disproportionately represented in terms of race, gender, age, and socioeconomic status (Pettit 2012; Spohn 2013; Western 2006). Social justice researchers have long studied this problem, even as far back as the writings of W. E. B. Du Bois (Du Bois 2010). The relative ease of obtaining accurate prison population data over the past twenty years has increased the awareness of this issue as well as the number of social justice studies that share insights (Spohn 2000; Tonry 2019; Baumer 2013; Franklin 2018; Mitchell 2005; Ulmer 2012; Zatz 2000).

The clear finding in this literature is that African American and Hispanic offenders receive harsher sentences than their white counterparts. Other conclusions from recent studies have revealed that Asians receive similar or more lenient punishment outcomes than people who are white for federal offenses (Everett and Wojtkiewicz 2002; Franklin and Henry 2020; Franklin 2015; Johnson and Betsinger 2009; Kutateladze et al. 2014). Additionally, some research has shown that Native Americans experience harsher penalties than white offenders in the federal system, though these results might depend on jurisdictional and policy-related factors (Everett and Wojtkiewicz 2002; Franklin 2013; Wilmot and Delone 2010; Engen and Gainy 2000). These findings are so clear that some researchers have concluded that “it is overly simplistic to ask whether race and ethnicity matter at sentencing” (Spohn 2015, page 179). Hence, a newer focus of this research looks deeper into sentencing trends with respect to secondary characteristics.

With respect to gender, this research has traditionally found that male defendants are treated more punitively than females, though these differences decrease when the type of crime committed and prior record are considered in the analysis (Daly and Tonry 1997; Goulette et al. 2015; Griffin and Wooldredge 2006; Steffensmeier et al. 1993; Doerner and Demuth 2010). The differences between male and female defendants also appears to be diminishing with time (Bontrager et al. 2013). The effects of gender on sentencing additionally appears to be related to the types of responsibility the defendants have to their families, such as being the primary source of childcare or income (Daly 1987; Doerner and Demuth 2014; Koons-Witt 2002; Tasca et al. 2019).

Sentencing as related to gender appears to be also related to race. For example, African American and Hispanic defendants are penalized more if they are male. Further, some studies suggest that white females receive the most lenient criminal penalties (Brennan and Spohn 2009; Demuth and Steffensmeier 2004; Kramer and Ulmer 2002; Mustard 2001; Ulmer et al. 2016; Brennan 2006; Steffensmeier et al. 1993). When considered in conjunction with age, young African American and Hispanic males have consistently been identified as the groups that are most disadvantaged in criminal punishment (Doerner and Demuth 2010; Franklin 2015; Lehmann and Gomez 2021; Spohn and Holleran 2000; Steffensmeier et al. 1998; Warren et al. 2012).

The type of offense can also interact with race and gender in sentencing. Many studies have found that African American and Hispanics are particularly disadvantaged in the sentencing of lesser property and drug crimes (Demuth and Steffensmeier 2004; Hester and Hartman 2017; Steffensmeier and Demuth 2000; Warren et al. 2012). However, other studies of violent and drug offenses have mixed results (Kautt and Spohn 2002; Lehmann 2020; Spohn and Cederblom 1991; Steen et al. 2005). Research on the interaction between minority status and criminal history also show mixed results (Franklin and Henry 2020; Hester and Hartman 2017; Spohn and Cederblom 1991; Miethe and Moore 1986; Ulmer and Kramer 1996).

There are also studies that provide evidence suggesting that the relative size of African American or Hispanic populations is important in sentencing inequalities (Bontrager et al. 2005; Ulmer and Johnson 2004; Wang and Mears 2015). Other work considers African American or Hispanic communities that experienced recent population growth. Some of these studies showed associations with harsher sentencing while others with slightly different contexts were inconclusive in their conclusions (Feldmeyer et al. 2015; Ulmer and Parker 2020; Caravelis at al. 2011, 2013; Wang and Mears 2010).

Education level is also of interest. There is evidence that those who have more education are at a reduced risk of receiving harsh punishments (Freiburger 2011; Johnson and Betsinger 2009; Wooldredge 2010). Additionally, some research indicates that racial differences are less for defendants who are college graduates (Franklin 2017; Steffensmeier and Demuth 2000). Other research indicates that unemployed defendants are also at a disadvantage when being sentenced and offenders who rely on financial assistance were also more likely to receive longer sentences (Chiricos and Bales 1991; Spohn and Holleran 2000; Wooldredge 2010).

While overt discrimination in sentencing is expressly prohibited by law, recent research has studied why there is a persistent bias against certain groups in the justice system in the United States. Explanations include sociological theories based on implicit and

perhaps unintentional bias by judges. In essence judges may implicitly categorize offenders based on categories they perceive to be linked to issues such as the likelihood of future criminal activity (Lehmann and Gomez 2022).

As one can observe, these issues are very complex, and statistical techniques are required to carefully study these trends. A simple straightforward analysis, without the use of more sophisticated statistical methods, has the potential to miss important trends, particularly those trends where two or more conditions interact with one another. Potentially even more important, without the ability to separate the interactive trends from one another, simple trends may be observed that do not really exist but are the result of several other conditions interacting with one another. Hence, the role of statistical techniques are vitally important to ensuring that the conclusions from studies of complex situations provide a realistic view of what the important issues are in a situation.

This page titled 1.4.8: Race and Gender in Sentencing Disparity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.4.9: The Major League Baseball Pitch Clock

In the 2023 season Major League Baseball implemented a pitch clock to quicken the games. Thirty seconds is allowed between batters and fifteen seconds between pitches with the bases empty and 20 seconds with runners on. The timer was reduced to 18 seconds with runners on base starting in 2024. When introduced, it was widely hypothesized that would lead to increased injuries due to decreased recovery time. This hypothesis is based on theoretical physiological considerations that state that there is not enough time for muscles to recover from the fast movements required for pitching, which could lead to increased muscle fatigue and increased forces on the pitcher's ligaments and tendons (Puga et al. 2024).

A recent study of pitcher injuries before and after the rule change used statistical techniques on data from the 2021–2023 seasons. The injury timeframe was limited between opening day and the final day of the regular season. Injuries not related to the physical structure of the body (muscle, skeletal, nervous systems) were excluded from the study. Data were then categorized into major anatomical categories and subcategories.

The results of the study were somewhat surprising in that after one year after the implementation of the pitch clock, there was a decrease in the total number of injuries, and injuries in several major categories. Many other categories showed virtually no difference in injuries when comparing the 2023 to the 2021 or 2022 seasons. There was no difference in the injury rates of elbow injuries (Puga et al. 2024). The authors conclude by stating that long-term tracking of injuries is critical to see if the pitch clock will eventually affect injuries in pitchers.

This page titled 1.4.9: The Major League Baseball Pitch Clock is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.4.10: Artificial Intelligence in Music Education

Artificial intelligence, computer software that is designed to mimic human intelligence and interactions, has the potential to reshape music education, enabling students to learn musical instruments and compose music in a fun and easy way (Li and Wang 2023). A recent study examined the effectiveness of teaching music aided by artificial intelligence technology. This study assessed 98 piano students aged 14–17 who were enrolled in two 8-week learning programs in seven music schools in China. One of the classes was based on traditional training while the second class incorporated an artificial intelligence chatbot to enhance several skills such as sight reading and vocal skills.

The results of the study were analyzed from several perspectives. For the first analysis the authors considered the improvement of the student in a practice exercise that required them to play a scale. The ability for them to complete the exercise was scored and recorded by a computer application. The results showed that the students who used the chatbot had higher mean scores than the students who used traditional training. The results were further divided by the number of times that the students practiced the exercise according to three categories: fewer than 500 times, 500–1,000 times, and more than 1,000 times. The results of this analysis showed that use of the artificial intelligence chatbot increased scores over all three of these categories.

The authors of the study also considered the overall performance of the students after an entry-level course and four additional courses of increasing level. Throughout the levels the chatbot-assisted classes had higher performance scores that the classes that only used traditional methods. The authors concluded that the artificial intelligence chatbot was not only useful for helping students develop basic skills but can be useful for more advanced students as well.

A final analysis considered four types of skills that the students were learning: specialty and sight reading, music literature, solfeggio (vocal exercises and ear training) and vocals. In all these skills the chatbot-assisted classes had higher performance scores than the classes that only used traditional methods. The largest difference was in specialty and sight reading, with only modest increases in music literature and vocals. The authors once again concluded that the artificial intelligence chatbot was useful for helping students develop each of the skills.

This type of research not only shows the potential positive effect that artificial intelligence chatbots can have on music education but also points the researchers in directions where there could be additional improvement. For example, the results indicated that there were only slight gains in music literature and vocals, which could motivate further exploration into developing technologies that aid in these areas of music education.

This page titled 1.4.10: Artificial Intelligence in Music Education is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.5: About the Quotation

Herbert George Wells (1866–1946, Figure 1.1) was an English writer who wrote more than fifty novels and dozens of short stories. In addition to the science fiction novels that he is well known for, he also wrote nonfiction works on politics and history. He is often referred to as the "father of science fiction." He is also known as a forward-looking social critic who wrote several utopian works and foresaw much of the technology that we enjoy today.



This page titled 1.5: About the Quotation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.6: Practice and Exploration

1. One of the new jobs for the new millennium is that of an influencer, an individual who builds an online presence through engaging content like photos, videos, and updates. Influencers differ from traditional celebrities because they grow their following through social media rather than preexisting fame. Successful influencers are often paid by businesses for to promote products and services. For example, influencers have commonly worked in the fashion industries. Of course, for a business that is considering hiring an influencer, the relevant question is how much they should be paid. Discuss the type of data that the influence might provide to the business that would help the business decide how much the influencer would be worth in terms of their profits. What potential problems might exist with the data? What types of data could the business find on their own?
2. Suppose that you are interested in attempting to make a prediction about an upcoming soccer match. You would like to make the best prediction possible about the outcome, so you need to start thinking about the data you will use to make your prediction. What type of data might you consider when making your prediction? Where would you look for the data, and do you think that all the data you would be interested in is easily available? Given the best data that you could find, how accurate do you think your prediction might be? What could happen during the game that would make the data you collected less relevant to the outcome of the game that you might have assumed when you made your prediction?
3. Suppose you are considering opening a new business in your local area. For you to obtain financing from a bank, you must provide a business plan, and you must be able to demonstrate that there will be demand for the services that your business provides. What type of data might you consider gathering that will support your application for financing? Is this data readily available? Where do you think you might be able to obtain the data? What problems might you face in gathering the data that you need?
4. You are planning an outdoor event that will take place six months from now. The event will only run smoothly if the weather is reasonably warm and dry. If there is inclement weather, you will personally suffer a substantial financial loss; if the weather is nice, you should get a large financial gain. You are considering insuring the event. That is, for a fee an insurer will provide a payout to you if the weather is bad on the day of the outdoor event. What types of data do you think that the insurer would consider in deciding how big the fee should be?
5. Suppose you are considering purchasing a new automobile. You have narrowed it down to a single model that has three options: gasoline powered, hybrid powered (electric with gasoline backup when needed), and fully electric. What type of data should you consider when trying to decide which option to purchase? How would you gather this data? What types of problems might you have in finding the data you need? What type of things could happen after you make your purchase that would have changed your decision had you known they would occur?
6. Suppose you oversee your campus activities board, and you have been asked to book the hot new band called the “Probability Twins” for an on-campus concert. You will need to determine what price should be charged for the tickets. What type of data should you consider when trying to decide what the ticket price should be? How would you gather this data? What types of problems might you have in finding the data you need?
7. Suppose you are flying to Hawaii for a much-needed rest after a long semester. You need to be at the airport by 10:00 a.m. so that you have time to check-in, go through security, and reach your gate within the required time. Using your own home and local airport as a model, what type of data should you consider to decide when you should leave your home to get to the airport on time? How would you gather this data? What types of problems might you have in finding the data you need?

This page titled 1.6: Practice and Exploration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

1.7: References

Alaimo, Katherine, Christine M. Olson, Edward A. Frongillo, Jr., and Ronette R. Briefel. 2001. “Food Insufficiency , Family Income, and Health in US Preschool and School-Aged Children.” *American Journal of Public Health* 91, no. 5, 781. https://doi.org/10.2105/AJPH.91.5.781

Albert, Judith L., Claire M. Cohen, Thomas F. Brockmeyer, and Ana M. Malinow. 2021. “Racism, Chronic Disease, and Mental Health: Time to Change Our Racialized System of Second-Class Care.” In *Healthcare* 9, 1276. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/healthcare9101276

Allen, W. David. 2007. “The Reporting and Underreporting of Rape.” *Southern Economic Journal* 73, no. 3, 623-641. https://doi.org/10.1002/j.2325-8012.2007.tb00793.x

Alvarez-Hernandez, Luis R., and Orion Mowbray. 2022. “Trends in Suicide Among Lesbian, Gay, and Bisexual Hispanic Individuals: National Violent Death Reporting System Data from 2012-2016.” *Journal of Homosexuality* 70, no. 7, 1225–1239. https://doi.org/10.1080/00918369.2021.2018878

APM Research Lab. 2020. The Color of Coronavirus: COVID-19 Deaths by Race and Ethnicity in the US. https://www.apmresearchlab.org/covid/deaths-by-race

Bartik, Timothy J., and Brad Hershbein. 2018. "Degrees of Poverty: The Relationship Between Family Income Background and the Returns to Education." Upjohn Institute Working Paper 18–284. W.E. Upjohn Institute for Employment Research. https://doi.org/10.17848/wp18-284

Bartlett, C. 2023. “‘This I Rebel Against’: Television Advertising, Rod Serling’s The Twilight Zone, and a Changing Industry.” New Review of Film and Television Studies 21, no. 3, 505–526. https://doi.org/10.1080/17400309.2023.2216101

Baumer, Eric P. 2013. “Reassessing and Redirecting Research on Race and Sentencing.” Justice Quarterly 30, no. 2, 231–261. https://doi.org/10.1080/07418825.2012.682602

Baumer, B. S., G. J. Matthews and Q. Nguyen. 2023. “Big Ideas in Sports Analytics and Statistical Tools for Their Investigation.” Wiley Interdisciplinary Reviews: Computational Statistics 15, no. 6, e1612. https://doi.org/10.1002/wics.1612

Benitez, Joseph, Charles Courtemanche, and Aaron Yelowitz. 2020. “Racial and Ethnic Disparities in COVID-19: Evidence from Six Large Cities.” *Journal of Economics, Race, and Policy* 3, no. 4, 243–261. https://doi.org/10.1007/s41996-020-00068-9

Bergen, Raquel Kennedy, and Elizabeth Barnhill. 2006. *Marital Rape: New Research and Directions*. National Resource Center on Domestic Violence, Pennsylvania Coalition Against Domestic Violence. https://vawnet.org/material/marital-...and-directions

Bertrand, Marianne, and Sendhil Mullainathan. 2004. “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review* 94, no. 4, 991–1013. https://www.aeaweb.org/articles?id=1...02828042002561

Blair, Janet M., Katherine A. Fowler, Shane P. D. Jack, and Alexander E. Crosby. 2016. “The National Violent Death Reporting System: Overview and Future Directions.” *Injury Prevention* 22, Supplement 1, i6–i11. https://doi.org/10.1136/injuryprev-2015-041819

Bontrager, Stephanie, William Bales, and Ted Chiricos. 2005. “Race, Ethnicity, Threat and the Labeling of Convicted Felons.” *Criminology* 43, No. 3, 589–622. https://doi.org/10.1111/j.0011-1348.2005.00018.x

Bontrager, Stephanie, Kelle Barrick, and Elizabeth Stupi. 2013. “Gender and Sentencing: A Meta–Analysis of Contemporary Research.” *Journal of Gender, Race, and Justice* 16, 349–372. https://www.researchgate.net/publica...orary_research

Bowen, Sarah, Sinikka Elliott, and Annie Hardison-Moody. 2021. “The Structural Roots of Food Insecurity: How Racism Is a Fundamental Cause of Food Insecurity.” *Sociology* Compass 15, no. 7, e12846. https://doi.org/10.1111/soc4.12846

Branch, Enobong Hannah and Christina Jackson. 2020. *Black in America: The Paradox of the Color Line*. John Wiley & Sons. https://www.wiley.com/en-us/Black+in...-9781509531394

Brennan, Pauline K. 2006. “Sentencing Female Misdemeanants: An Examination of the Direct and Indirect Effects of Race/Ethnicity.” *Justice Quarterly* 23, no. 1, 60–95. https://doi.org/10.1080/07418820600552477

Brennan, Pauline K., and Cassia Spohn. 2009. “The Joint Effects of Offender Race/Ethnicity and Sex on Sentence Length Decisions in Federal Courts.” *Race and Social Problems* 1, no. 4, 200–217. https://doi.org/10.1007/s12552-009-9016-0

Brennan, Maya, Patrick Reed, and Lisa A. Sturtevant. 2011. “The Impacts of Affordable Housing on Education: A Research Summary.” Center for Housing Policy and National Housing Conference, Washington, D.C. https://nhc.org/wp-content/uploads/2...ducation-1.pdf

Caravelis, Cyndy, Ted Chiricos, and William Bales. 2011. “Static and Dynamic Indicators of Minority Threat in Sentencing Outcomes: A Multi-Level Analysis.” *Journal of Quantitative Criminology* 27, no. 4, 405–425. https://doi.org/10.1007/s10940-011-9130-1

Caravelis, Cyndy, Ted Chiricos, and William Bales. 2013. “Race, Ethnicity, Threat, and the Designation of Career Offenders.” *Justice Quarterly* 30, no. 5, 869–894. https://doi.org/10.1080/07418825.2011.635597

Carl, S., and B. Baldwin. 2022. *nflfastR: Functions to Efficiently Access NFL Play by Play Data*. https://CRAN.R-project.org/package=nflfastR

Chiricos, Theodore G., and William D. Bales. 1991. “Unemployment and Punishment: An Empirical Assessment.” *Criminology* 29, no. 4, 701-724. https://doi.org/10.1111/j.1745-9125.1991.tb01085.x

Cuevas-Molina, Ivelisse, Tatishe M. Nteta, Brian Schaffner, and Wouter van Erve. 2022. “The Conditional Effects of Latino Candidates and Partisanship on Latino Turnout.” *American Politics Research* 50, no. 5, 723-730. https://doi.org/10.1177/1532673X221090753

Curtin, Sally C., Margaret Warner, and Holly Hedegaard. 2016. *Suicide Rates for Females and Males by Race and Ethnicity: United States, 1999 and 2014*. NCHS Health E-Stat. https://www.cdc.gov/nchs/data/hestat..._1999_2014.pdf

D’Astous, Alain, and Francis Chartier. 2000. “A Study of Factors Affecting Consumer Evaluations and Memory of Product Placements in Movies.” *Journal of Current Issues and Research in Advertising* 22, no. 2, 31–40. https://doi.org/10.1080/10641734.2000.10505106

Daly, Kathleen. 1987. “Discrimination in the Criminal Courts: Family, Gender, and the Problem of Equal Treatment.” *Social Forces* 66, no. 1, 152–175. https://doi.org/10.1093/sf/66.1.152

Daly, Kathleen and Michael Tonry. 1997. “Gender, Race, and Sentencing.” *Crime and Justice* 22, 201–252. https://doi.org/10.1086/449263

Darity Jr., William A. and A. Kirsten Mullen. 2020. *From Here to Equality: Reparations for Black Americans in the Twenty-First Century*. UNC Press. https://uncpress.org/book/9781469671...econd-edition/

David, F. N. (1998). *Games, Gods and Gambling. A History of Probability and Statistical Ideas*. Dover. https://archive.org/details/gamesgod...ing-david-1962

Demuth, Stephen and Darrell Steffensmeier. 2004. “Ethnicity Effects on Sentence Outcomes in Large Urban Courts: Comparisons Among White, Black, and Hispanic Defendants.” *Social Science Quarterly* 85, no. 4, 994–1011. https://doi.org/10.1111/j.0038-4941.2004.00255.x

Dewan, N. 2008. *Indian Life and Health Insurance Industry: A Marketing Approach*. Gabler-Verlag. https://link.springer.com/book/10.10...-3-8349-9788-3

Dickson, P. G. M. 1960. *The Sun Insurance Office 1710–1960: The History of Two and a Half Centuries of British Insurance*. Oxford University Press. https://archive.org/details/suninsur...ge/n9/mode/2up

Doerner, Jill K. and Stephen Demuth. 2010. “The Independent and Joint Effects of Race/Ethnicity, Gender, and Age on Sentencing Outcomes in US Federal Courts.” *Justice Quarterly* 27, no. 1, 1–27. https://doi.org/10.1080/07418820902926197

Doerner, Jill K. and Stephen Demuth. 2014. “Gender and Sentencing in the Federal Courts: Are Women Treated More Leniently?” Criminal Justice Policy Review 25, no. 2, 242–269. https://doi.org/10.1177/0887403412466877

Du Bois, William Edward Burghardt. 2010. “The Philadelphia Negro.” In The Philadelphia Negro. University of Pennsylvania Press. https://www.pennpress.org/9781512824...delphia-negro/

Durojaye, E., G. Mirugi-Mukundi, J.-M. Iyi, and U. M. Assim, eds. 2025. *Conflict and Poverty in Africa: A Research Companion.* Routledge. https://www.routledge.com/Conflict-a...6i8DDei5UaKXUR

Engen, Rodney L., and Randy R. Gainey. 2000. “Modeling the Effects of Legally Relevant and Extralegal Factors Under Sentencing Guidelines: The Rules Have Changed.” *Criminology* 38, no. 4, 1207–1230. https://doi.org/10.1111/j.1745-9125.2000.tb01419.x

Emmons, William R., Ana H. Kent, and Lowell R. Ricketts. 2018. “A Lost Generation? Long-Lasting Wealth Impacts of the Great Recession on Young Families.” *The Demographics of Wealth*, May 2018. https://www.stlouisfed.org/dialogue-...ost-generation

Everett, Ronald S., and Roger A. Wojtkiewicz. 2002. “Difference, Disparity, and Race/Ethnic Bias in Federal Sentencing.” *Journal of Quantitative Criminology* 18, no. 2, 189–211. https://doi.org/10.1023/A:1015258732676

Feldmeyer, Ben, Patricia Y. Warren, Sonja E. Siennick, and Malisa Neptune. 2015. “Racial, Ethnic, and Immigrant Threat: Is There a New Criminal Threat on State Sentencing?” *Journal of Research in Crime and Delinquency* 52, no. 1, 62–92. https://doi.org/10.1177/0022427814548488

Ferro, Christine, Jill Cermele, and Ann Saltzman. 2008. “Current Perceptions of Marital Rape: Some Good and Not-So-Good News.” *Journal of Interpersonal Violence* 23, no. 6, 764–779. https://doi.org/10.1177/0886260507313947

Fienberg, Stephen E. 1992. “A Brief History of Statistics in Three and One-Half Chapters: A Review Essay.” *Statistical Science* 7, no. 2, 208–225. https://doi.org/10.1080/01615440.1991.9955300

Fouad, Mona N., John Ruffin, and Selwyn M. Vickers. 2020. “COVID-19 Is Disproportionately High in African Americans. This Will Come As No Surprise. . . .” *The American Journal of Medicine* 133, no. 10, e544–e545. https://doi.org/10.1016/j.amjmed.2020.04.008

Fraga, Bernard L. 2018. The Turnout Gap: Race, Ethnicity, and Political Inequality in a Diversifying America. Cambridge University Press. https://doi.org/10.1017/9781108566483

Franklin, Travis W. 2013. “Sentencing Native Americans in US Federal Courts: An Examination of Disparity.” *Justice Quarterly* 30, no. 2, 310–339. https://doi.org/10.1080/07418825.2011.605072

Franklin, Travis W. 2015. “Race and Ethnicity Effects in Federal Sentencing: A Propensity Score Analysis.” Justice Quarterly 32, no. 4, 653–679. https://doi.org/10.1080/07418825.2013.790990

Franklin, Travis W. 2017. “Sentencing Outcomes in US District Courts: Can Offenders’ Educational Attainment Guard Against Prevalent Criminal Stereotypes?” *Crime and Delinquency* 63, no. 2, 137–165. https://doi.org/10.1177/0011128715570627

Franklin, Travis W. 2018. “The State of Race and Punishment in America: Is Justice Really Blind?” *Journal of Criminal Justice* 59, 18–28, 2018. https://doi.org/10.1016/j.jcrimjus.2017.05.011

Franklin, Travis W., and Tri Keah S. Henry. 2020. “Racial Disparities in Federal Sentencing Outcomes: Clarifying the Role of Criminal History.” *Crime and Delinquency* 66, no. 1, 3–32. https://doi.org/10.1177/0011128719828353

Freiburger, Tina L. 2011. “The Impact of Gender, Offense Type, and Familial Role on the Decision to Incarcerate.” *Social Justice Research* 24, no. 2, 143–167. https://doi.org/10.1007/s11211-011-0133-8

Gould, E., and V. Wilson. 2020. “Black Workers Face Two of the Most Lethal Preexisting Conditions for Coronavirus—Racism and Economic Inequality.” Economic Policy Institute. Retrieved from https://coilink.org/20.500.12592/5j5nvq on 04 Feb 2025. COI: 20.500.12592/5j5nvq.

Goulette, Natalie, John Wooldredge, James Frank, and Lawrence Travis III. 2015. “From Initial Appearance to Sentencing: Do Female Defendants Experience Disparate Treatment?” *Journal of Criminal Justice* 43, no. 5, 406–417. https://doi.org/10.1016/j.jcrimjus.2015.07.003

Granderson, Kevin. 2022. “Structural Power, Systemic Poverty, and Entrenchment Among African American Males.” In *Addressing Issues of Systemic Racism During Turbulent Times*, 200–223. IGI Global Publishing. https://doi.org/10.4018/978-1-

7998-8532-0.ch013

Griffin, Timothy, and John Wooldredge. 2006. “Sex-Based Disparities in Felony Dispositions Before Versus After Sentencing Reform in Ohio.” *Criminology* 44, no. 4, 893–923. https://doi.org/10.1111/j.1745-9125.2006.00067.x

Gyimah-Brempong, K. 2002. “Corruption, Economic Growth, and Income Inequality in Africa.” *Economics of Governance* 3, 183–209. https://doi.org/10.1007/s101010200045

Hardy, Bradely, Timothy Smeeding, and James P Ziliak. 2018. “The Changing Safety Net for Low-Income Parents and Their Children: Structural or Cyclical Changes in Income Support Policy?” *Demography* 55, no. 1, 189–221. https://doi.org/10.1007/s13524-017-0642-7

Hardy, Bradley L., and Trevon D Logan. 2020. *Racial Economic Inequality Amid the COVID-19 Crisis*. The Hamilton Project Essay 2020-17. The Brookings Institution. https://www.brookings.edu/wp-content...an_LO_8.12.pdf

Hardy, Bradley L., and Dave E Marcotte. 2020. “Education and the Dynamics of Middle-Class Status.” Online Brookings Article. https://www.brookings.edu/articles/e...-class-status/

Hedegaard, Holly, Sally C. Curtin, and Margaret Warner. 2018. *Suicide Rates in the United States continue to Increase*. NCHS Datra Brief Volume 309. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/data/databriefs/db309.pdf

Hester, Rhys, and Todd K. Hartman. 2017. “Conditional Race Disparities in Criminal Sentencing: A Test of the Liberation Hypothesis from a Non-Guidelines State.” Journal of Quantitative Criminology 33, no. 1, 77–100. https://doi.org/10.1007/s10940-016-9283-z

Holzer, Harry J., Steven Raphael, and Michael A. Stoll. 2006. “Perceived Criminality, Criminal Background Checks, and the Racial Hiring Practices of Employers.” *The Journal of Law and Economics* 49, no. 2, 451–480. https://doi.org/10.1086/501089

Jahromi, Alireza Hamidian, and Anahid Hamidianjahromi. 2020. “Why African Americans Are a Potential Target for Covid-19 Infection in the United States.” *Journal of Medical Internet Research* 22, no. 6, e19934. https://www.jmir.org/2020/6/e19934/

Johnson, Brian D., and Sara Betsinger. 2009. “Punishing the ‘Model Minority’: Asian-American Criminal Sentencing Outcomes in Federal District Courts.” *Criminology* 47, no. 4, 1045–1090. https://doi.org/10.1111/j.1745-9125.2009.00169.x

Kandil, Emad, Abdallah S. Attia, Mohanad R. Youssef, Mohammad Hussein, Kareem Ibraheem, Mohamed Abdelgawad, Zaid Al-Qurayshi, and Juan Duchesne. 2020. “African Americans Struggle with the Current COVID-19.” *Annals of Surgery* 272, no. 3, e187. https://doi.org/10.1097/SLA.0000000000004185

Kautt, Paula, and Cassia Spohn. 2002. “Crack-ing Down on Black Drug Offenders? Testing for Interactions Among Offenders’ Race, Drug Type, and Sentencing Strategy in Federal Drug Sentences.” *Justice Quarterly* 19, no. 1, 1–35. https://doi.org/10.1080/07418820200095151

Koons-Witt, Barbara A. 2002. “The Effect of Gender on the Decision to Incarcerate Before and After the Introduction of Sentencing Guidelines.” *Criminology* 40, no. 2, 297–328. https://doi.org/10.1111/j.1745-9125.2002.tb00958.x

Kramer, John H., and Jeffery T. Ulmer. 2002. “Downward Departures for Serious Violent Offenders: Local Court ‘Corrections’ to Pennsylvania’s Sentencing Guidelines.” *Criminology* 40, no. 4, 897–932. https://doi.org/10.1111/j.1745-9125.2002.tb00977.x

Kraus, Michael W., Ivuoma N. Onyeador, Natalie M. Daumeyer, Julian M. Rucker, and Jennifer A. Richeson. 2019. “The Misperception of Racial Economic Inequality.” *Perspectives on Psychological Science* 14, no. 6, 899–921. https://doi.org/10.1177/1745691619863049

Kutateladze, Besiki L., Nancy R. Andiloro, Brian D. Johnson, and Cassia C. Spohn. 2014. “Cumulative Disadvantage: Examining Racial and Ethnic Disparity in Prosecution and Sentencing.” *Criminology* 52, no. 3, 514–551. https://doi.org/10.1111/1745-9125.12047

Laurison, Daniel, Dawn Dow, and Carolyn Chernoff. 2020. “Class Mobility and Reproduction for Black and White Adults in the United States: A Visualization.” Socius, 6, https://doi.org/10.1177/2378023120960959.

Laurison, Daniel, Hana Brown, and Ankit Rastogi. 2021. “Voting Intersections: Race, Class, and Participation in Presidential Elections in the United States 2008–2016.” *Sociological Perspectives* 65, no. 4, 768–789. https://doi.org/10.1177/07311214211059136

Lehmann, Peter S. 2020. "Race, Ethnicity, Crime Type, and the Sentencing of Violent Felony Offenders." *Crime and Delinquency* 66, no. 6–7, 770–805. https://doi.org/10.1177/0011128720902699

Lehmann, Peter S., and Anna I. Gomez. 2021. “Split Sentencing in Florida: Race/Ethnicity, Gender, Age, and the Mitigation of Prison Sentence Length.” *American Journal of Criminal Justice* 46, no. 2, 345–376. https://doi.org/10.1007/s12103-020-09550-4

Lehmann Peter S. and Anna I. Gomez. 2022. “Racial, Ethnic, Gender, and Economic Sentencing Disparity.” In *Handbook of Issues in Criminal Justice Reform in the United States*, 127–142. Elizabeth Jeglic and Cynthia Calkins, Editors. Springer. https://doi.org/10.1007/978-3-030-77565-0_8

Li, P. Ping, and B. Wang. 2023. “Artificial Intelligence in Music Education.” *International Journal of Human–Computer Interaction* 40, no. 16, 4183–4192. https://doi.org/10.1080/10447318.2023.2209984

Lindsey, G. R. 1961. “The Progress of the Score During a Baseball Game.” *Journal of the American Statistical Association* 56, no. 295, 703–728. https://doi.org/10.1080/01621459.1961.10480656

Lindsey, G. R. 1963. “An Investigation of Strategies in Baseball.” *Operations Research* 11, no. 4, 477–501. https://doi.org/10.1287/opre.11.4.477

Lock, D., and D. Nettleton 2014. “Using Random Forests to Estimate Win Probability Before Each Play of an NFL Game.” *Journal of Quantitative Analysis in Sports* 10, no. 2, 197–205. https://doi.org/10.1515/jqas-2013-0100

Mercy, James A., L. Barker, and Lorraine Frazier. 2006. “The Secrets of the National Violent Death Reporting System.” *Injury Prevention* 12, Supplement 2, ii1–ii2. https://doi.org/10.1136/ip.2006.012542

Miethe, Terance D., and Charles A. Moore. 1986. “Racial Differences in Criminal Processing: The Consequences of Model Selection on Conclusions About Differential Treatment.” *Sociological Quarterly* 27, no. 2, 217–237. https://doi.org/10.1111/j.1533-8525.1986.tb00258.x

Mitchell, Ojmarrh. 2005. “A Meta-Analysis of Race and Sentencing Research: Explaining the Inconsistencies.” *Journal of Quantitative Criminology* 21, no. 4, 439–466. https://doi.org/10.1007/s10940-005-7362-7

Mustard, David B. 2001. “Racial, Ethnic, and Gender Disparities in Sentencing: Evidence from the US Federal Courts.” *The Journal of Law and Economics* 44, no. 1, 285–314. https://doi.org/10.1086/320276

Myhill, Andy, and Jonathan Allen. 2002. *Rape and Sexual Assault of Women: The Extent and Nature of the Problem*. Home Office. https://citeseerx.ist.psu.edu/docume...ef6575fccbf6d0

O’Reilly, Nicole L., Erin R. Hager, Donna Harrington, and Maureen M. Black. 2020. “Assessment of Risk for Food Insecurity Among African American Urban Households: Utilizing Cumulative Risk Indices and Latent Class Analysis to Examine Accumulation of Risk Factors.” *Translational Behavioral Medicine* 10, no. 6, 1322–1329. https://doi.org/10.1093/tbm/ibaa027

Pettit, Becky. 2012. *Invisible Men: Mass Incarceration and the Myth of Black Progress*. Russell Sage Foundation. https://www.russellsage.org/publicat.../invisible-men

Pfeffer, Fabian, T., and Alexandra Killewald. 2018. “Generations of Advantage. Multigenerational Correlations in Family Wealth.” *Social Forces* 96, no. 4, 1411–1442. https://doi.org/10.1093/sf/sox086

Puga, Troy, Nicholas Scigliano, McKenna Box, John Riehl. 2024. “The Major League Baseball Pitch Clock: First Year Analysis of Pitcher Injuries.” *Journal of Orthopaedics* 57, 133–136. https://doi.org/10.1016/j.jor.2024.06.022

Reyes, Maritza Vasquez. 2020. “The Disproportional Impact of COVID-19 on African Americans.” *Health and Human Rights* 22, no. 2, 299. https://www.hhrjournal.org/wp-conten...quez-Reyes.pdf

Stauffer, Katelyn E., and Bernard L. Fraga. 2022. “Contextualizing the Gender Gap in Voter Turnout.” *Politics, Groups, and Identities* 10, no. 2, 334–341. https://doi.org/10.1080/21565503.2021.1893195

Schneider, Daniel, and Kristen Harknett. 2019. “Consequences of Routine Work-Schedule Instability for Worker Health and Well-Being.” *American Sociological Review* 84, No. 1, 82–114. https://doi.org/10.1177/0003122418823184

Skiba, Russell J., Robert H. Horner, Choong-Geun Chung, M. Karega Rausch, Seth L. May, and Tary Tobin. 2011. “Race is Not Neutral: A National Investigation of African American and Latino Disproportionality in School Discipline.” *School Psychology Review* 40, no. 1, 85–107. https://doi.org/10.1080/02796015.2011.12087730

Spohn, Cassia C. 2000. “Thirty Years of Sentencing Reform: The Quest for a Racially Neutral Sentencing.” *Criminal Justice* 3, no. 1, 427–501. https://www.researchgate.net/publica...encing_Process

Spohn, Cassia C. 2013. “Racial Disparities in Prosecution, Sentencing, and Punishment.” *The Oxford Handbook of Ethnicity, Crime, and Immigration*. 166–193. Oxford University Press. https://doi.org/10.1093/oxfordhb/978...859016.013.006

Spohn, Cassia. 2015. “Evolution of Sentencing Research.” *Criminology and Public Policy* 14, no. 2, 225–232. https://doi.org/10.1111/1745-9133.12125

Spohn, Cassia, and Jerry Cederblom. 1991. “Race and Disparities in Sentencing: A Test of the Liberation Hypothesis.” *Justice Quarterly* 8, no. 3, 305–327. https://doi.org/10.1080/07418829100091071

Spohn, Cassia, and David Holleran. 2000. “The Imprisonment Penalty Paid by Young, Unemployed Black and Hispanic Male Offenders.” *Criminology* 38, no. 1, 281–306. https://doi.org/10.1111/j.1745-9125.2000.tb00891.x

Steen, Sara, Rodney L. Engen, and Randy R. Gainey. 2005. “Images of Danger and Culpability: Racial Stereotyping, Case Processing, and Criminal Sentencing.” *Criminology* 43, no. 2, 435–468. https://doi.org/10.1111/j.0011-1348.2005.00013.x

Steenkamp, Malinda, Lorraine Frazier, N. Lipskiy, M. DeBerry, S. Thomas, L. Barker, and Debra Karch. 2006. “The National Violent Death Reporting System: An Exciting New Tool for Public Health Surveillance.” *Injury Prevention* 12, supplement 2, ii3–ii5. https://doi.org/10.1136/ip.2006.012518

Tango, T. M., M. G. Lichtman, and A. E. Dolphin. 2007. *The Book: Playing the Percentages in Baseball*. Potomac Books. https://www.google.com/books/edition...C?hl=en&gbpv=0

Steffensmeier, Darrell, John Kramer, and Cathy Streifel. 1993. “Gender and Imprisonment Decisions.” *Criminology* 31, no. 3, 411–446. https://doi.org/10.1111/j.1745-9125.1993.tb01136.x

Steffensmeier, Darrell, Jeffery Ulmer, and John Kramer. 1998. “The Interaction of Race, Gender, and Age in Criminal Sentencing: The Punishment Cost of Being Young, Black, and Male.” *Criminology* 36, no. 4, 763–798. https://doi.org/10.1111/j.1745-9125.1998.tb01265.x

Steffensmeier, Darrell and Stephen Demuth. 2000. “Ethnicity and Sentencing Outcomes in US Federal Courts: Who Is Punished More Harshly?” *American Sociological Review* 65, no. 5, 705–729. https://doi.org/10.1177/000312240006500505

Tasca, Melinda, Ahram Cho, Cassia Spohn, and Nancy Rodriguez. 2019. “The Role of Parental Status and Involvement in Sentence Length Decisions: A Comparison of Men and Women Sentenced to Prison.” *Crime and Delinquency* 65, no. 14, 1899–1924. https://doi.org/10.1177/0011128718811929

Tjaden, Patricia Godeke, and Nancy Thoennes. 2006. *Extent, Nature, and Consequences of Rape Victimization: Findings from the National Violence Against Women Survey.* National Institute of Justice, U. S. Department of Justice. https://nij.ojp.gov/library/publicat...ional-violence

Tonry, M. 2019. “Fifty Years of American Sentencing Reform: Nine Lessons.” *Crime and Justice* 48, no. 1, 1–34. https://doi.org/10.1086/701798

Ulmer, Jeffery T., and John H. Kramer. 1996. “Court Communities Under Sentencing Guidelines: Dilemmas of Formal Rationality and Sentencing Disparity.” *Criminology* 34, no. 3, 383–408. https://doi.org/10.1111/j.1745-9125.1996.tb01212.x

Ulmer, Jeffery T., and Brian Johnson. 2004. “Sentencing in Context: A Multilevel Analysis.” *Criminology* 42, no. 1, 137–178. https://doi.org/10.1111/j.1745-9125.2004.tb00516.x

Ulmer, Jeffery T. 2012. “Recent Developments and New Directions in Sentencing Research.” *Justice Quarterly* 29, no. 1, 1–40. https://doi.org/10.1080/07418825.2011.624115

Ulmer, Jeffery, Noah Painter-Davis, and Leigh Tinik. 2016. “Disproportional Imprisonment of Black and Hispanic Males: Sentencing Discretion, Processing Outcomes, and Policy Structures.” *Justice Quarterly* 33, no. 4, 642–681. https://doi.org/10.1080/07418825.2014.958186

Ulmer, Jeffery T., and Brandy R. Parker. 2020. “Federal Sentencing of Hispanic Defendants in Changing Immigrant Destinations.” *Justice Quarterly* 37, no. 3, 541–570. https://doi.org/10.1080/07418825.2019.1621362

Vaughan, E. J. 1997. *Risk Management*. John Wiley and Sons. https://www.google.com/books/edition...QgAACAAJ?hl=en

Verba, Sidney, Kay Lehman Schlozman, and Henry E. Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Harvard University Press. https://www.google.com/books/edition...C?hl=en&gbpv=0

Wang, Lei, Hsien-Chang Lin, and Y. Joel Wong. 2021. “Perceived Racial Discrimination on the Change of Suicide Risk Among Ethnic Minorities in the United States.” *Ethnicity and Health* 26, no. 5, 631–645. https://doi.org/10.1080/13557858.2018.1557117

Wang, Xia, and Daniel P. Mears. 2010. “Examining the Direct and Interactive Effects of Changes in Racial and Ethnic Threat on Sentencing Decisions.” *Journal of Research in Crime and Delinquency* 47, no. 4, 522–557. https://doi.org/10.1177/0022427810375576

Wang, Xia, and Daniel P. Mears. 2015. “Sentencing and State-Level Racial and Ethnic Contexts.” *Law and Society Review* 49, no. 4, 883–915. https://doi.org/10.1111/lasr.12164

Warren, Patricia, Ted Chiricos, and William Bales. 2012. “The Imprisonment Penalty for Young Black and Hispanic Males: A Crime-Specific Analysis.” *Journal of Research in Crime and Delinquency* 49, no. 1, 56–80. https://doi.org/10.1177/0022427810397945

Western, Bruce. 2006. *Punishment and Inequality in America*. Russell Sage Foundation. https://www.russellsage.org/publicat...lity-america-1

Whitaker, Lois Duke, Ed. 2008. *Voting the Gender Gap*. University of Illinois Press. https://www.press.uillinois.edu/books/?id=p075254

Wiemers, Emily E., Scott Abrahams, Marwa Al Fakhri, V. Joseph Hotz, Robert F. Schoeni, and Judith A. Seltzer. 2020. “Disparities in Vulnerability to Severe Complications from COVID-19 in the United States.” Technical Report. National Bureau of Economic Research. https://doi.org/10.1101/2020.05.28.20115899

Wilmot, Keith A., and Miriam A. Delone. 2010. “Sentencing of Native Americans: A Multistage Analysis Under the Minnesota Sentencing Guidelines.” *Journal of Ethnicity in Criminal Justice* 8, no. 3, 151–180. https://doi.org/10.1080/15377938.2010.502821

Wolfinger, Raymond E. and Steven J. Rosenstone. 1980. *Who Votes?* Yale University Press. https://yalebooks.yale.edu/book/9780...521/who-votes/

Wooldredge, John. 2010. “Judges’ Unequal Contributions to Extralegal Disparities in Imprisonment.” *Criminology* 48, no. 2, 539–567. https://doi.org/10.1111/j.1745-9125.2010.00195.x

Wright, James E., and Cullen C. Merritt. 2020. “Social Equity and COVID-19: The Case of African Americans.” *Public Administration Review* 80, no. 5, 820–826. https://doi.org/10.1111/puar.13251

Yancy, Clyde W. 2020. “COVID-19 and African Americans.” *Journal of the American Medical Association* 323, no. 19, 1891–1892. https://doi.org/10.1001/jama.2020.6548

Yurko, R., F. Matano, L. F. Richardson, N. Granered, T. Pospisil, K. Pelechrinis and S. L. Ventura. 2020. “Going deep: Models for Continuous-Time Within-Play Valuation of Game Outcomes in American Football with Tracking Data.” *Journal of Quantitative Analysis in Sports* 16, no. 2, 163–182. https://doi.org/10.1515/jqas-2019-0056

Zatz, Marjorie S. 2000. “The Convergence of Race, Ethnicity, Gender, and Class on Court Decision Making: Looking Toward the 21st century.” *Criminal Justice* 3, no. 1, 503–552. https://nij.ojp.gov/library/publicat...making-looking

Ziliak, James P., Bradley Hardy, and Christopher Bollinger. 2011. “Earnings Volatility in America: Evidence from Matched CPS.” *Labour Economics* 18, no. 6, 742–754.https://doi.org/10.1016/j.labeco.2011.06.015

This page titled 1.7: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

CHAPTER OVERVIEW

2: Empiricism

“Everything must be taken into account. If the fact will not fit the theory—let the theory go.”

Agatha Christie

2.1: What to Expect
2.2: Empiricism and Knowledge
2.3: The Modern Scientific Method
2.4: Feminist Empiricism
2.5: Critical Analysis
2.5.1: Labor Market Discrimination
2.5.2: Same-Sex Marriage Laws and Hate Crimes
2.5.3: Performance After Surgery in Baseball Players
2.5.4: The Supreme Court and Personal Attitudes
2.5.5: Success Factors for Mobile Work in Health Care
2.5.6: Improving Web Searches
2.5.7: Customer Avoidance of Advertising on Social Media
2.6: Summary
2.7: About the Quote
2.8: Practice and Exploration
2.9: Collaborative Activity
2.10: References

This page titled 2: Empiricism is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.1: What to Expect

This chapter discusses the philosophy and methodology of how we find things out about the world we live in. In most studies about our world, the mechanism for learning about reality is based on the scientific method and the philosophy of empiricism. With this method, a hypothesis is tested by observing some behavior in the world. The belief in the truth of the hypothesis is then determined by what is observed under controlled conditions, if possible. The hypothesis may then be refined and updated, and further observation is used to determine the truth of the refined hypothesis. It is through this cycle that we gain knowledge. This chapter will discuss the history and development of the scientific method. Several real studies are then considered with discussions on how observation was used to gain knowledge, and what possible problematic issues could arise when making conclusions based on the observations.

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

- Define epistemology.
- Define and discuss empiricism as a philosophy for gaining knowledge and discuss how it is connected to the scientific method.
- Define and discuss inductive and deductive inference.
- Define and discuss the concept of reasoning to the best explanation.
- Identify and discuss the steps of the scientific method.
- Define and discuss alternate philosophical theories related to feminist empiricism.
- Critically analyze the application of the scientific method in a research article.

This page titled 2.1: What to Expect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.2: Empiricism and Knowledge

One example from Chapter 1 considered a study that finds an association between a student's race and the severity of school discipline. In this study it is concluded that non-white students face, on average, more severe discipline for the same behavior than white students. Specifically, one of the findings indicates that students from African American and Latino families are more likely than their white peers to receive expulsion or out-of-school suspension as consequences for the same or similar problem behavior. The authors of the research based their conclusions on a review of the documented patterns of office discipline referrals in 364 elementary and middle schools during the 2005–2006 academic year as reported by school personnel (Skiba et al. 2011). The data were modeled using statistical techniques and the authors were able to demonstrate that the conclusions can be inferred from what is contained in the observed data.

For the moment we will not worry about the exact methods that were used to make the conclusions from the observed data. What is important at this time is the fact that the knowledge about this behavior was obtained by directly observing the patterns in society. That is, knowledge was gained from empirical observations. This is an example of empiricism, a philosophical methodology that is closely related to the modern scientific method.

The idea that we can observe the behavior of individuals in society and use those observations to obtain information or knowledge about society is a concept that is taken for granted by most scientists and researchers today. However, the idea that knowledge can be obtained from observation developed over a long time, going all the way back to the classical philosophers of Greece. This development begins with important questions about what is truly known to be true. That is, what is knowledge?

Epistemology refers to the study of knowledge—what separates belief and opinion from things that are known and what we really mean when we say that something is known. Hence, epistemology is concerned with questions that involve concepts such as knowledge, evidence, belief, justification, and likelihood (Fumerton 2009). It is important to note that the word knowledge as used in this context has a very specific meaning that separates it from opinion or belief. Knowledge is very specifically confined to those concepts that are taken to be facts, things that are known to be true. The process of obtaining knowledge is called science.

Definition: Science

Science is the systemic process of obtaining knowledge by deduction and careful observation in the form of testable hypotheses and predictions about the universe (New Websters Dictionary and Thesaurus 1992).

This is a contemporary definition of science that results from over two thousand years of philosophical development. The systematic process used to obtain knowledge is called the scientific method. Philosophers since the time of Aristotle have participated in very deep debates on how a known fact can be distinguished from a belief. The scientific method used today is a relatively recent methodological development and is still the subject of intense debate among philosophers and scientists.

The philosophy of science and the scientific method are generally considered to begin with Aristotle (see Figure **2.2.1**), though he considered philosophy and science to be indistinguishable (Waugh and Ariew 2008). Aristotle saw science as a stable and deductive structure based on first principles (Curd et al. 2014). This view of science is based on an axiomatic framework. That is, a scientific investigation begins with a set of axioms. The axioms are the starting point; they are first principles derived from experience and are known with certainty.

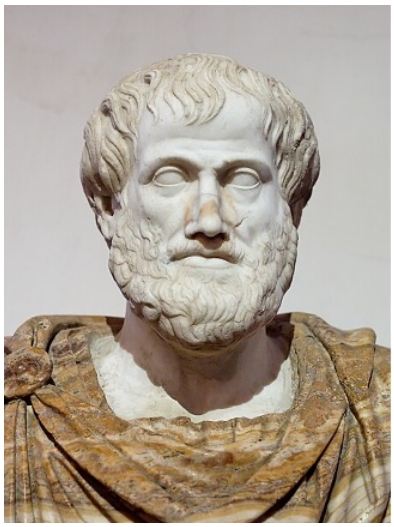


Figure **2.2.1**: Aristotle (384–322 BCE) was a Greek philosopher of the classical period in ancient Greece. His methods of inquiry correspond to the first development of the scientific method. He is portrayed here in a Roman marble bust copied from a Greek bronze original. (Public domain image)

Definition: Axiom

An *axiom* is a statement that is taken to be true and serves as a starting point for further reasoning.

Everything else that is taken to be knowledge is derived from these principles. The key to this method is that the first principles, or axioms, must be absolutely known to be true or the entire system fails.

The axiomatic approach to science depends on making conclusions based on the first principles, or axioms. The method that is used for expanding a theory from the axioms to other conclusions is deduction.

Definition: Deduction

A conclusion is based on *deduction* if it follows logically from a set of statements that are known to be true.

Essentially, deductive arguments reason that if a statement is true about a group, then the statement is true about any member of the group. For example, we might postulate as an axiom that “all cats are fuzzy.” We may observe that Felix (see Figure **2.2.2**), is quite obviously a cat. We can then deductively conclude that Felix is fuzzy. This is also objectively known to be true as evidenced by the figure. The theory would fail as soon as we observe a hairless cat, which are decidedly not fuzzy.



Figure **2.2.2**: Official portrait of Felix Ferdinand Polansky of Creston, IL (public domain image created by the author).

As another example, suppose we want to study rocks using this method. Based on our perhaps limited experience we might postulate an axiom that states “all rocks sink in water.” We might very well assume this to be a universally accepted fact that could not be debated. From this axiom we can then build up a larger scientific theory by deducing facts from this first principle. For example, we might deduce that because all rocks sink, we cannot build a raft from rock. Further development of our rock theory could proceed as usual and in our experience everything about the rock theory might seem to be scientific fact. Everything is fine, that is, until a traveler from a distant land brings us a sample of pumice! Pumice is a volcanic rock that is created when super-heated, highly pressurized rock is violently ejected from a volcano. Pumice has an unusual foamy configuration because of simultaneous rapid cooling and rapid depressurization. Pumice typically has a very low density, less than that of water, and hence a pumice stone will float in water. The problem now is that any part of your scientific theory that depended on the assumption that all rocks float is invalid. In fact, after the explosion of the Krakatoa volcano in 1883, rafts of pumice drifted through the Indian Ocean for up to 20 years (Devantier et al. 1992).

Euclid used this axiomatic method to develop geometry with a great deal of success, though modern research in geometry often challenges some of the axioms that Euclid relied on. Even in this case, standard Euclidean geometry is a valid geometric system for most applications. However, Aristotle's view of science has had some spectacular failures. A prominent example is in describing planetary motion, where the assumptions that postulated all earthly bodies move toward the center of the universe and that planets tend to move in circles around the center of the universe could not withstand observational scrutiny (Curd et al. 2014). Essentially, what was observed empirically did not match what should have been observed had the postulated model been true.

Nicolaus Copernicus formulated a model of the universe that placed the sun, rather than the Earth, at its center, at odds with earlier models that assumed that the Earth was the center of the universe. Though Aristarchus of Samos, an ancient Greek astronomer, had formulated such a model some eighteen centuries earlier, the publication of Copernicus's model in 1543 was a significant advancement in the history of science. Copernicus' ideas triggered the Copernican Revolution, a major reassessment of the nature of science and the universe that through the subsequent work of Galileo, Descartes, and finally Newton, showed that the planets in our solar system revolve in elliptical orbits around the sun.

Galileo Galilei (see Figure **2.2.3**) is widely regarded as the first modern physicist because he was the first to use mathematics to describe the dynamic behavior of moving objects. This was a major shift in scientific methodology since at the time it was thought that mathematics could not be applied to physical reality. Furthermore, Galileo emphasized testing hypotheses experimentally, which at the time was not considered a reliable means of obtaining knowledge. This is the beginning of the empirical approach to scientific investigation (Okasha 2002). The work of Galileo was refined by other scientists, namely René Descartes. Isaac Newton improved Descartes's theory—and invented calculus along the way—to develop what is known a Newtonian mechanics, a theory that remained essentially unchanged until the early twentieth century and which is the basis for most calculations involving the behavior of physical bodies in motion to this day.



Figure **2.2.3**: Galileo Galilei (1564–1642) was an Italian scientist who began the development of the modern scientific method, shown here in a 1636 portrait by Justus Sustermans (public domain image).

For our current discussion, the important aspects of the Copernican Revolution are the effects that its development had on the philosophy of science, and the rapid expansion of scientific thought that occurred roughly between the writings of Copernicus and the synthesis of Newton's model for the universe. It is in this rapid expansion of science that empiricism, the theory that knowledge is gained from experience, became an import idea in scientific thought. By the start of the Copernican Revolution, empiricism had already become an important component of science. For example, the early 14th-century philosopher William of Ockham started an intellectual movement toward empiricism (Hannam 2011).

Definition: Empiricism

Empiricism is the philosophical theory that knowledge is gained through experience.

British empiricism arose from the philosophical differences between Francis Bacon, an empiricist, and René Descartes, a rationalist whose philosophy was closer to that of Aristotle. Thomas Hobbes, George Berkeley, and David Hume were the primary supporters of British empiricism, and collectively they developed a sophisticated empirical framework as the basis of human knowledge.

An influential formulation of British empiricism is based on John Locke's "An Essay Concerning Human Understanding," which was published in 1689. In this essay he maintains that the only true knowledge that is accessible to the human mind is that which is based on experience. He believed that the human mind was created as a "blank tablet" upon which sensory impressions were recorded, and that knowledge was accumulated through a process of reflection. He wrote:

> Let us then suppose the mind to be, as we say, white paper void of all characters, without any ideas. How comes it to be furnished? Whence comes it by that vast store which the busy and boundless fancy of man has painted on it with the most endless variety? Whence has it all the materials of reason and knowledge? To this I answer in one word, from experience; in

> that all of our knowledge is founded, and from that it ultimately derives itself. Our observation, employed either about external sensible objects, or about the internal operations of our minds perceived and reflected on by ourselves, is that which supplies our understanding with all materials of thinking. These are the two fountains of knowledge, from whence all ideas we have, or can naturally have, do spring. (Locke 1974)

It is interesting to note that Locke anticipated the modern use of probability theory in the attainment of knowledge. In his theory of knowledge, it is suggested that there are two types of knowledge, the type that is known for certain and the type that is not certain but is wise to accept. Their reasoning is that very few things are known for certain and therefore we should act based on the probability that something is true (Russell 2013). That is, we cannot really have true knowledge about facts, only a probable belief. This type of reasoning is an important part of scientific study, particularly in fields like psychology, sociology, political science, biology, medicine, and education. We will discuss probable belief in detail in later chapters. It should be noted that the concept of a "blank slate" dates to Aristotle and was later developed by Stoic philosophers.

The empiricists changed the direction of scientific thought from one that is based solely on logical deduction to one that is based on a mixture of observation and logical deduction. The development of empiricism lead to the modern scientific method commonly used today.

This page titled 2.2: Empiricism and Knowledge is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.3: The Modern Scientific Method

The modern scientific method is an iterative process that depends on putting forth hypotheses, predicting outcomes based on these hypotheses, and observing whether the predictions hold through experimentation. The term scientific method dates from the nineteenth century when terminologies establishing clear boundaries between science and non-science were being developed (Harrison et al. 2011). The fundamental feature of science, and an inherent part of the scientific method, is that a scientific theory is one that can make definite predictions that can be tested against experience (Popper 1989). If experience does not match the predictions, then the corresponding theory is false.

Karl Popper used this argument to separate scientific theories such as Einstein's Theory of General Relativity, which was first empirically verified in 1919 when an expedition led by Eddington confirmed general relativity's prediction for the deflection of starlight by the sun during the total solar eclipse of 1919, from what they considered non-scientific theories such as Freud's psychoanalytic theory, which could not be reconciled with empirical findings. In practice, it should be noted that this viewpoint may be overly simplistic in practice.

Hence, scientific inquiry and the scientific method are an iterative process where theory and observation complement one another. For example, Newton's gravitational theory made specific predictions about the how the planets should behave as they orbit the sun. These predictions could be empirically verified except for the planet Uranus, because observations of Uranus varied from the theory in a consistent way. Two scientists, John Couch Adams working in England and Urbain Le Verrier working in France, independently suggested in 1846 that there was a yet undiscovered planet whose gravitational pull was responsible for the observed behavior of Uranus. That same year the planet Neptune was observed for the first time.

The key idea here is that Newton’s entire theory was not abandoned based on these observations, particularly because the theory apparently correctly predicted the behavior of the other known planets. Instead, the hypothesized theory and observation led to a new prediction that would explain why the empirical observations for Uranus did not match what the theory predicted. The new viewpoint was that there was another planet that was affecting the system in a way that is consistent with the hypothesized theory. Of course, the modified view was verified in short order in this case. If no such planet had been observed, then the theory would have been modified in a different way and tested again.

Another key element of the scientific method is the concept of induction. Earlier we stated that deduction is making a conclusion that logically follows from another. The argument in our example used the deduction to conclude that if all rocks sink in water, and we find a rock, then it will also sink in water. An inductive argument is one that moves from conclusions about cases we know about to a case we do not know about. In the example of the rocks, we might observe that six rocks have been observed to sink in water. An inductive argument, or inductive inference, would be used to conclude that a seventh rock will sink when placed in water. Hence, we take the knowledge about objects we have observed and apply it to try to make a conclusion about another object for which we have no observations.

Definition: Induction

A conclusion is based on *induction* if it is generalized from a set of observations.

Of course, inductive inference can be more dangerous than deductive inference. If the assumed premise is true, then a deductive conclusion always holds, whereas an inductive conclusion may not. In the example of the rocks the inductive inference obviously does not hold because what we observed in the six rocks is not true for all rocks.

The modern scientific method is heavily dependent on inductive inference. For example, the scientific method takes the notion that if we observe that the set of known planets all follow elliptical orbits around the sun, then we would conclude that if a new planet were observed, it too would follow this pattern. As an example from everyday life, the conclusion that a medical treatment is safe depends on what has been observed up to that point but may not always hold for someone who has not been observed to have the treatment. As we shall see, modern statistical theory attempts to mitigate this problem by assessing the risk of something that has not been observed will follow an established pattern from what has been observed. In fact, inductive reasoning is used any time one looks at data and draws general conclusions from it. In practice this is often referred to as empirical proof or experimental proof, but it is not proof in the sense of using a deductive argument.

While admitting that we use inductive inference almost constantly in our lives, David Hume insisted that inductive inference cannot be rationally justified, and that we are unconsciously assuming what he called the uniformity of nature (Okasha 2002). For example, if the first three eggs in a carton are good, then concluding that the rest in the carton will also be good makes an inherent assumption about the nature of the uniformity of egg goodness. Whenever an airplane takes off, we implicitly assume that the laws of gravity will continue to hold, and in fact will hold for that airplane on that flight; otherwise, we would probably not board the plane. While there is no logical reason for the universe to continue to work exactly as it has been, we implicitly believe this to be true. Because of these assumptions inductive inference apparently cannot be justified logically. This is known as Hume's Problem of Induction, a problem with which scientific philosophers struggle to this day.

Another type of reasoning used in science is in some ways closely related to inductive inference. This type of reasoning is known as inference to the best explanation. When using inference to the best explanation, a scientist considers all the information and produces the best plausible explanation for the observations. Suppose that you leave for class in the morning, and you leave a box of tasty donuts on the table in the kitchen. Your roommate is still sleeping as they have a later class, and they were out late the night before. You arrive home after your class and the box of donuts is empty. You conclude that your roommate ate the donuts in your absence. You do not know this for a fact, but you look at the information you have. The only other person who lives with you is your roommate, no one else usually visits on class days, and you were not gone that long. This conclusion is not a logical deduction from the information that is given, as there is a myriad of other possible explanations. Perhaps you accidentally left the door ajar, someone saw the donuts, came in and ate them, and left while your roommate slept. Perhaps a mouse got in and ate all the donuts. The conclusion that your roommate ate the donuts seems to be the most plausible, or best, explanation.

This type of reasoning may not seem to be inductive reasoning because the conclusion that your roommate ate your donuts does not generalize previous donut eating behavior. You did not observe your roommate eating the first few donuts and then conclude that they ate the rest of the donuts based on those observations. Some would argue, though, that your conclusion may be based on previous experience. You know by experience that it is not likely someone came into your apartment and just ate the donuts and left. You know by experience that your roommate is perfectly capable of eating your donuts in your absence. From this perspective some may then argue that this is a type of inductive reasoning. There is no correct answer to using this terminology, and it really does not matter if we are consistent. In our context we will use the term inductive inference for reasoning about the same situation. Donuts to doughnuts, so to speak.

In application, inference to the best explanation is one of the most used methods for scientific inquiry, but it can also be the most haphazard. The main problem stems from the fact that it can be very difficult to settle on which of a competing series of hypotheses about a question presents the most plausible explanation. The development of the Darwinian theory of evolution is based on observations of similarities between species and the conclusion that these species share a common ancestor as the most plausible explanation. This reasoning certainly did not prove that evolution was the mechanism of the natural development of species, and even in the face of modern advances in genetic theory, the theory is still debated today (Thagard and Findlay 2010; Hildering et al. 2013; Diogo 2020; Cavallo and McCall 2008; Crivellaro and Sperduti 2014; Cobern 1994; Gilson 2009).

Another example of this type of development occurred with the study of Brownian motion. In 1827 the botanist Robert Brown noted that when he observed pollen suspended in water through a microscope, the grains of pollen moved in an irregular fashion. Over the next roughly 80 years, several explanations for the phenomenon were proposed, including theories based on the electrical attraction between particles and convection currents in the water. The actual reason for the movement is because of the bombardment of the water molecules against the pollen grains, a theory proposed in the late 19th century. However, a theoretical mathematical treatment of the theory was elusive, and hence it was not accepted until the work of Albert Einstein in 1905. What is interesting about the theory is that it postulated the existence of molecules and atoms, and in a symbiotic logical relationship bolstered scientific belief in this theory as well (Okasha 2002; Haw 2002; Maiocchi 1990; Marchesoni and Hänggi 2005; Bigg 2008).

One of the key elements in accepting a theory as being plausible is based on the idea of parsimony or simplicity. It is usually much easier to believe that a simple explanation is the cause of a phenomenon rather than a complicated one. But this theory provokes a deep question about whether this is the preferred structure of the universe (Okasha 2002). Is the universe necessarily based on the simplest explanation? This concept is still widely debated (Huemer 2009; Stewart 1993; Friday 1982; Aarts 2007; Kluge 2001; Simon 2002; Schlesinger 1975).

So where does this leave us with the modern scientific method? Essentially the modern view of the scientific method is of an iterative process that relies on a cycle of hypothesis and testing, usually represented with a framework as shown in Figure **2.3.4**. There are many other representations with varying degrees of complexity, but the basic process is essentially the same. The scientific method begins with observation. Some type of behavior is observed in the universe, and we would like to use the scientific method to explain why it has occurred. The second step is then to form a hypothesis about why the behavior might occur. In the best cases this hypothesis is something that can be tested through experimentation;, otherwise the hypothesis is just a theory that has not been verified. The next step involves testing the hypothesis through experimentation. Generally, what happens is that the hypothesis is used to make predictions about what should occur under conditions that are controlled or observed by the researcher. The experiment is performed, and data is collected, and this brings us to the next step, which is to analyze the data to see if the data support or are contradictory to what was predicted by the hypothesis. At this point a formal conclusion is made. If the data are contradictory to what was predicted by the hypothesis, then we have observed something new, and a new hypothesis is needed. Then the process starts again. If the data support the predictions of the hypothesis, then in some sense we have confirmed our theory, though this is usually not the end.

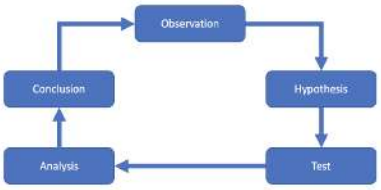


Figure **2.3.4**: A conceptual flowchart of the modern scientific method visualized as a cycle of observation, hypothesis, testing, and analysis (Created by the authors).

Richard Feynman summarized this situation in the following way:

> "The exception tests the rule." Or, put another way,
> "The exception proves that the rule is wrong."
> That is the principle of science.
> If there is an exception to any rule, and if it can be proved by observation, that rule is wrong.

The problem is that theories are never fully confirmed by the scientific method. What happens in the cyclical process is that evidence in support of a hypothesis is slowly built until there is a consensus in the scientific community that the hypothesis is accepted. However, we can never be certain that a hypothesis is correct.

An interesting example to consider is the development of the theory of the simple dynamics of physical bodies. Isaac Newton introduced three laws of motion in his book *Mathematical Principles of Natural Philosophy* published in 1687 (see Figure **2.3.5**). It is thought that Newton derived these laws based on empirical observation and mathematical logic. Similarly, Galileo considered empirical data on the motion of objects in his book *The Discourses and Mathematical Demonstrations Relating to Two New Sciences* published in 1638. In fact, it has been proven that the laws of motion could have been solely derived from this data (Katsikadelis 2015).

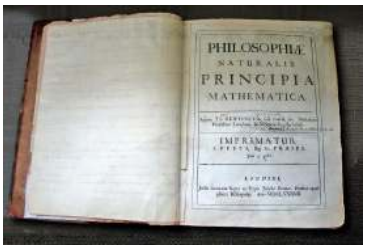


Figure **2.3.5**: The title page of Newton's *Mathematical Principles of Natural Philosophy* (public domain image).

These laws were accepted as fact up until the twentieth century, based partly on the supporting mathematics and partly on the ability of the laws to make accurate predictions of the everyday behavior of objects. However, the scientific theory of gravitational laws changed when Albert Einstein published his theory of general relativity in 1915. This theory shows that there are corrections due to the size of the gravitational effect and the velocity of the object that are not accounted for in Newton's laws. Indeed, there

has been no possible way to even observe these effects until recently, and in practice the Newtonian theory of gravity is a good approximation when gravitational effects are weak, and objects are moving slowly compared to the speed of light (Goldstein et al. 2002).

All of this does not mean that the scientific method is not useful, or that we cannot gain practical knowledge using the scientific method. What it means is that we must be diligent in applying the scientific method, and that we should consider the accumulated knowledge that scientific experimentation affords us. Scientific progress has cured many diseases, prevented many others, and has generally improved the quality of life for those in modern society. The development of the polio vaccine is an excellent example of the application of the scientific method in practice that has had a fundamental effect on our lives.

In the early part of the twentieth century the United States was suffering from a major polio epidemic that lasted forty years, claiming hundreds of thousands of victims, many of whom were children. Polio was a disease that tended to affect affluent communities, and this observation provided the basis for the hypothesis that a vaccine could be developed to prevent the disease. It was thought that children who lived in less affluent areas were exposed to mild cases of the disease at a very early age while they were still protected by antibodies from their mother. After becoming infected, these children would develop their own antibodies that would protect them from a more severe infection later. Hence, the hypothesis was that a vaccine could be used to train the immune systems of children to fight the disease. This led to the development of a vaccine that was used in test trials in the mid-1950s. By the end of the twentieth century polio was essentially eliminated in the United States due to the development of this vaccine.

This page titled 2.3: The Modern Scientific Method is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.4: Feminist Empiricism

Until relatively recently the dominant viewpoint was that scientific endeavors are best suited for the male viewpoint. In fact, science was generally considered to be an inherently masculine venture, and that women were not capable of being rational, neutral, and objective, the necessary attributes of being a good scientist (McNeal 1994). Unfortunately, gender discrimination in science is still a common problem (Casad et al. 2021; Schwerter and Ilg 2021; Stewart-Williams and Halsey 2021; Bird and Rhoton 2021).

The twentieth century did show some advancement with influential women in science, but the progress has been slow even in the best circumstances. Marie Curie, an early researcher into radiation, won the Nobel Prize in physics in 1903 and the Nobel Prize in chemistry in 1911 (see Figure **2.4.6**). She is the only person to win the Nobel prize twice in two different fields and was the first woman awarded the prize. The prizes were recognition for her work in isolating radium and polonium, two unknown elements at the time. At first the committee had intended to only honor her husband and another scientist, but after complaints from committee members and her husband, Marie's name was added to the nomination. Even with these achievements the French Academy of Sciences failed to elect her as a member in 1911. By 2012 only 43 women had been awarded Nobel prizes out of 862 awards (Van de Sande and Byvelds 2015).



Figure **2.4.6**: Marie Curie (1867–1934) was a physicist and chemist who conducted pioneering research on radioactivity (public domain image).

The discovery in 1953 of the double helix structure of DNA (deoxyribonucleic acid) by James Watson and Francis Crick has been heralded on one of the greatest discoveries in the history of science and gave rise to modern molecular biology. This discovery, and the medical advances that have followed, have fundamentally changed the modern world. In 1968 Watson published their account of the discovery in their book The Double Helix. The publication of the book was not without controversy, as Crick was unhappy with his portrayal in the book. This was not the only controversy, however. Another scientist, Rosalind Franklin, is given scant acknowledgement in the book, even though she provided crucial evidence on the structure (Klug 1968). This lack of acknowledgement is often attributed to attitudes about gender in science that were common at the time.

While scientific inquiry is intended to gain knowledge without bias, gender identity can have a fundamental effect on how the scientific method is applied in practice. Feminist empiricism shares many characteristics with empiricism, in that it is an epistemology that values knowledge based on experience. However, in the feminist approach the experience is based on a female perspective. This is a recognition of the interaction of evidence and objectivity with moral and political values. The researcher is not an isolated individual but is part of a political community. Feminist empiricists use what is known as a naturalist epistemology, an approach to empiricism that encourages disruption, multiple approaches to the evaluation of knowledge, and radicalism (Hundleby 2011).

Traditional epistemology relies on individualism and the idea that to achieve true knowledge one must free themselves from the influences of society. That is, knowledge can only be obtained by a separate individual who rejects the notion that social factors have a role in the production of knowledge. Feminist empiricism, however, is based on the belief that social factors are not only relevant to, but are also important to knowledge (Letherby 2003; Nelson 2010). Indeed, feminist epistemologists believe that

detachment from the subject matter to be a neutral observer is neither possible nor desirable. It is held in this empirical philosophy that emotions are important to and can enrich an investigation (Letherby 2003). Those using the traditional empirical method assume that they are neutral and objective. Hence, they cannot acknowledge the social factors that can influence their research (Van de Sande and Byvelds 2015).

To eliminate the bias associated with the sexism in science, it has been suggested that the context of discovery bears as much importance as the context of justification. The context of discovery corresponds to such things as the research hypothesis, theoretical framework, and the method of analysis used in the experiment. The context of justification refers to what evidence is used to assess a theory.

Helen Longino developed what is known as conceptualized feminist empiricism. Longino was frustrated by the fact that the traditional philosophy of science does not acknowledge the relationship between social values and scientific inquiry (Longino 1990). Acknowledging and reconciling social values with the objectivity of science, Longino suggests that background beliefs are biased and that better background beliefs, based on a community that includes feminist voices, will produce better scientific investigation. This transformation is based on a philosophy of science that includes communities of individuals working together on scientific inquiry and that in the end the community has the final say on scientific evidence. This is based on the idea that knowledge can be secured through social interaction (Longino 2018). This philosophy is ultimately empirical and parsimonious in nature in that Longino believes that knowledge is obtained through scientific investigation, and that scientific philosophy should also rely on criteria such as truth, accuracy, simplicity, and predictability.

Lynn Hankinson-Nelson developed a similar scientific philosophy known as naturalized feminist empiricism. Naturalized feminist empiricism starts with the basic principles of empiricism with the view that science must be understood within its political, social, and cultural context. However, this philosophy goes even further; there is a belief that the theoretical framework from which hypotheses are formulated is important and dependent on the social context. That is, the same observations can result in different conclusions depending on the social context of the underlying theory. Inherent in this idea is that observation is shaped by history and culture, and this should be accounted for in scientific investigations (Nelson 2010).

Sandra Harding developed a system of feminist philosophy that essentially proposes that science will only achieve its goals if gender plays an essential role. Indeed, her development goes beyond gender, advocating for a multicultural approach to science. This is based on the idea that those with different gender identities have differing relationships to the world around them and that science can become distorted by the linked meanings of objectivity and gender (Harding 1994; Harding 1998). The implementation of this philosophy is based on the development of strong objectivity, which removes the distortion and bias in science by ensuring a multicultural community of scientists.

The feminist perspective on empiricism is controversial and is not widely accepted outside the human and social sciences (Curd et al. 2014). Some modern philosophers argue that there is no evidence of bias in the standard approach to empiricism and that the necessity for a systematic change in the scientific method has not been substantially proven. Some of these arguments are circular and biased in themselves, sort of saying that it seems like science works so science must work. Of course, the bias is that we may not know if science can be better without acknowledging some of the issues raised in this section. It seems there can be little risk in inviting a diverse community of scientists to work together within whatever scientific philosophy is the basis for scientific inquiry. Whether there will be a long-term shift in scientific inquiry is still very much up to debate.

This page titled 2.4: Feminist Empiricism is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.5: Critical Analysis

When critically analyzing a study in terms of its use of empiricism, one should consider whether a study is based on an axiomatic structure with deductive reasoning or an empirical structure based on inductive reasoning through the framework of the scientific method. If a study is based on an axiomatic structure, then the axioms, the statements that initially are taken to be true, are crucial and should be carefully examined. Studies based on the scientific method should be examined in terms of the steps shown in Figure 2.5. A careful examination of what was first observed, and how these observations were used to develop the hypotheses to be tested is important. Evaluating the remaining steps of the scientific method will be covered in detail in the remaining chapters of this book.

While pure axiomatic studies are quite rare in applied statistical research, the method of deductive reasoning from some initial assumptions as a method for developing a hypothesis to use in the scientific method is common. This type of reasoning is also common in interpreting the results of the testing phase of the scientific method. Therefore, one will find that many studies contain aspects of both inductive and deductive reasoning.

The key questions in critically analyzing the use of empiricism in a study are:
1. Does the study use an axiomatic method, an empirical method, or both?
2. Do the basic assumptions of the analysis seem reasonable? Are the axioms self-evident? Are there weak axioms that could cause the entire analysis to be questionable?
3. To what extent is the scientific method used?
4. How would a multicultural approach to science be used for the same experiment?

This page titled 2.5: Critical Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.5.1: Labor Market Discrimination

Government economic data continue to show that significant racial inequity exists in the labor market of the United States with consistently higher unemployment rates and lower wages for African Americans when compared to white individuals. One of the potential factors related to this issue is whether a bias exists in the hiring practices of employers, and researchers are interested in the bias in the way that African American and white candidates are assessed for job opportunities.

One interesting study created (Bertrand and Mullainathan 2004) virtual resumes in response to help wanted ads in Chicago and Boston. Two types of resumes were sent, one with names that were perceived as uniquely African American and the other using names that were perceived as uniquely white. These names were chosen using demographic data from birth certificates and were vetted using surveys that specifically asked respondents what race they expected the name to relate to, if they were able to make such a determination. The resumes were carefully constructed to be believable with the correct level of experience for the advertised position. The results of the study showed the resumes that were perceived as African American received fewer callbacks from potential employers than the resumes that were perceived to come from white applicants.

This study is based on empiricism and an application of the scientific method. Within the context of the scientific method, the initial observation is that historically African American individuals have higher unemployment rates and lower wages when compared to white individuals. From this observation the researchers formed the hypothesis that employers may be biased when evaluating the resumes of potential employees, even when the race of the candidate is not explicitly known to the evaluator. To test this hypothesis the researchers created and sent out the virtual resumes, and it was observed that resumes that were designed to be perceived to come from African American applicants received fewer callbacks than those resumes that were designed to be perceived to come from white applicants. The conclusion is that there is a potential for bias when employers evaluate resumes, even when the race of the applicant is not explicitly revealed.

This page titled 2.5.1: Labor Market Discrimination is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.5.2: Same-Sex Marriage Laws and Hate Crimes

Between 2004 and 2018 support for legal same-sex marriages has more than doubled, partly due to the normalization of same-sex relationships in mainstream culture (Nikolaou 2022). Despite this trend, homophobic sentiments, which are often expressed as hate crimes, are still widespread and under-reported (Coffman et al. 2017). In fact, sexual orientation offenses are the third most reported type of hate crime in the United States, with verbal harassment and discriminatory practices ranking as the top two forms.

Recent research investigated whether advances in marriage equality have influenced criminal offenses against the LGBTQIA+ community. The psychological theory known as the legitimacy model states that policies that legitimatize the LGBTQIA+ community can lead to increased acceptance and improved inclusion in the society (Flores and Barclay 2016; Black et al. 2007). One study revealed increased support for the LGBTQIA+ community among participants exposed to favorable reports of the U.S. Supreme Court deliberations about same-sex marriage compared to participants exposed to unfavorable reports (Tankard and Paluck 2017). Other studies have also indicated that the implementation of same-sex marriage laws have improved the general attitude toward the LGBTQIA+ community (Flores and Barclay 2016; Hooghe and Meeusen 2013; Kreitzer et al. 2014). The hypothesis is then that same-sex marriage laws can result in changes in social attitudes and individual behaviors that may reduce victimization of the LGBTQIA+ community.

These differences are studied using the scientific method through the variation in the implementation of same-sex marriage laws prior to the Supreme Court ruling that legalized same-sex marriage in the United States. Data for hate crimes against the LGBTQIA+ community were studied using data provided by the Federal Bureau of Investigation through the National Incident Based Reporting System. The data show that enacting same-sex marriage laws are associated with a 12% decrease in hate crimes (Nikolaou 2022). This supports the hypothesis that marriage equality increases tolerance and acceptance and decreases social stigma, both of which can lead to lower rates of hate crimes.

This page titled 2.5.2: Same-Sex Marriage Laws and Hate Crimes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.5.3: Performance After Surgery in Baseball Players

In the world of Major League Baseball (MLB), pitchers are some of the most valuable members of a team, with their salaries being some of the highest in the league. Pitching can be difficult on the body, particularly the throwing arm, and it is not uncommon for some players to require surgery during their career. Of course, one of the major concerns of both the athlete and their organization, is whether the player can return to their previous level of performance after such a surgery. A recent study considered this question for pitchers who had received ulnar collateral ligament (UCL) surgery (Mastroianni et al. 2025).

The study considered 129 MLB pitchers who suffered UCL injuries and had primary UCL reconstruction or repair between November 2017 and November 2023. The performance after these surgeries was calculated in November 2024 for all pitchers who had at least two seasons of 100 pitches or more after their surgery. This resulted in 58 pitchers who were used for the study. The researchers gathered information about the body characteristics, game statistics, and pitch characteristics.

The results of the study are voluminous and somewhat complicated, but several broad trends were identified by the researchers, a couple of which will be highlighted here. Looking at the change of a statistic that measures how valuable a pitcher is to a team, the researchers concluded that a little fewer than half of the pitchers provided the same value to teams three seasons after surgery even though these players should be fully recovered from the surgery. When considering the physical characteristics of their pitches, it was observed that these characteristics improved each season over the three post-operative seasons observed and that at least 75% of the players returned to pre-injury performance after three seasons.

This is a study based on empiricism and the scientific method. The researchers hypothesized that the injury and surgery combination would influence the value of the pitchers to the team, and this was empirically observed in the data. The researchers point out that there are many limitations to this study. Primarily, some of the observed effects were quite small, so that even though a difference was observed, it was not clear that the difference would have a practical effect. Furthermore, the researchers used statistical definitions to determine the effectiveness of a pitcher. The value of a player to a team is a complicated and abstract concept that could be defined and measured in many ways.

In terms of the use of the scientific method, this study is based on the idea of induction and the uniformity of nature. The results were observed for a specific set of individuals who were involved in the study. The usefulness of the results would be to conclude that this trend would apply to other populations of players who experience the same set of circumstances. This conclusion is complicated, particularly since the nature of the position of pitcher in baseball and the technology used in medical interventions is always changing. If this study were done again a decade later, the results could be different.

This page titled 2.5.3: Performance After Surgery in Baseball Players is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.5.4: The Supreme Court and Personal Attitudes

Another interesting research project studied whether the potential Supreme Court ruling on same-sex marriage would change individuals' personal attitudes about the LGBTQIA+ community (Tankard and Paluck 2017). The authors of this study hypothesized that institutional decisions may change an individual's perception of social norms, and the impressions of what opinions or behaviors are common or considered desirable by society (Cialdini and Goldstein 2004). The idea is that individuals in a society use social norms as a guide to social behavior because they wish to be accurate in their social judgment and because they wish to avoid social rejection (Tankard and Paluck 2017; Cialdini and Goldstein 2004). The authors point out that previous research had demonstrated this type of effect with social behaviors concerned with energy conservation, voting, and alcohol consumption (Tankard and Paluck 2016).

Prior to the Supreme Court ruling, the study participants were invited to read a brief article about the likely outcome of the upcoming Supreme Court ruling on same-sex marriage. Each participant was randomly assigned to read either a version entitled "Supreme Court likely to rule in favor of gay marriage" or a version entitled "Supreme Court unlikely to rule in favor of gay marriage.". Participants then responded to a series of questions about demographics, other social viewpoints, and behaviors (Tankard and Paluck 2017).

This study is based on empiricism and the scientific method. The hypothesis that the Supreme Court ruling may change the attitudes of individuals is based on previous research that demonstrated such an effect in other similar situations. The hypothesis was tested experimentally and the result from this study suggests that perceptions of Americans' current and future support for same-sex marriage can be changed by a positive Supreme Court ruling on same-sex marriage. However, the research found that personal attitudes and feelings toward same-sex marriage and people in the LGBTQIA+ community did not shift in the wake of a potential ruling. The difference here is based on what an individual may accept as a societal norm against how they feel personally about the subject. That is, an individual may feel that it is acceptable for individuals of the same sex to have legally recognized marriages but may still feel that it is not morally right for them personally. The research showed that these personal feelings did not shift in the same way as the perception of societal norms did.

This page titled 2.5.4: The Supreme Court and Personal Attitudes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.5.5: Success Factors for Mobile Work in Health Care

In the days prior to the global pandemic in 2020 the concept of using mobile technology to provide health care was still somewhat new and was being studied by researchers as a means of improving access to health care. Health providers considered the potential of mobile health care for several tasks, such as getting regular updates about patients, access to physicians for consultation outside of the office, the access of medical of patients through their mobile devices, and the potential of having a central location for the storage and retrieval of medical documents (Ammenworth et al. 2000).

A recent investigation studied mobile work in health care by building a theoretical model based on propositions that arise from looking at published literature about mobile computing. The primary objective of the study was to identify key factors found in the literature that influenced the success of mobile work within the context of health care, using a measure of net benefit as the instrument of comparison. The authors noted that the benefits for mobile work in health care has many factors that influence the net benefit (Chatterjee et al. 2009).

The authors stated that mobile work in health care is successful when health care professionals can enter patient and medical information into centralized databases irrespective of their location and can communicate and coordinate actions with other health care workers using multiple channels. Using these ideas, the researchers concluded that system, information, and service quality, are factors that affect the successful use of mobile technology for health care workers. They also concluded that user satisfaction would also be associated with a net benefit.

Based on the existing literature of the subject, the authors proposed the following propositions for their study related to use and satisfaction, summarized here (Chatterjee et al. 2009):

1. The extent of data processing capability of a mobile device will positively affect use.
2. The extent of data processing capability of a mobile device will positively affect satisfaction.
3. The extent of information access will positively affect use.
4. The extent of information access will positively affect satisfaction.
5. The extent of communicability will positively affect use.
6. The extent of communicability will positively affect satisfaction.
7. The portability of a mobile device will positively affect use.
8. The portability of a mobile device will positively affect satisfaction.

Several other sets of propositions were developed to reflect ideas for the structure of tasks, urgency of information, mobility of health care workers, information complexity, and system support. These statements became the axioms upon which the results of the study were be inferred. The statements are axioms in the sense that they were developed in consultation with a large review of the mobile technology research literature. From these propositions a deductive reasoning structure was developed that was used to draw the conclusions for the study.

As the authors point out, this research is not completely deductive in nature. The authors use the term hypothetico-deductive logic for this type of research, a term and methodology that has been used by researchers in the field of information systems. As we have seen, a purely deductive approach uses already formulated theoretical propositions. These propositions are then used to deduce what results logically follow from the propositions. Using hypothetico-deductive logic, the theoretical propositions only temporarily support results consistent with the propositions. The propositions are then tested using empirical evidence. If the observed evidence is not found to be consistent with the propositions, then one can doubt the validity of the propositions.

The credibility of the propositions was explored using an approach that included researching two well-known extensive medical research databases. These databases were searched for articles including terms like “mobile technology” and “hand-held devices” that appeared between 2005 and 2008. The abstracts of the articles were examined, and irrelevant articles were excluded, which resulted in 70 research articles being included in the study. Each article was then examined within the context of each of the propositions to determine whether the article confirmed each proposition.

The review of the literature reveals mixed support for the propositions put forth by the authors. Some surprising results were that the level of data processing capability of the mobile device was not associated with amount of use or user satisfaction. One possible reason for this fact is that health care professionals at the time tended to use the devices for simple applications. Further, the extent of information access was not associated with amount of use or user satisfaction. The reason, once again, could be the simple tasks

that health care professionals used the devices for at the time. The overall analysis developed a small set of conditions that must be met for technology to be used and for user satisfaction to be high. These conditions include portability, reliability, and institutional support.

This type of analysis is interesting from the viewpoint in that it is essentially a deductive type of study with the exception that the propositions were proposed and then the veracity of the propositions were explored empirically. In pure deductive studies the propositions are taken to be universal truths, that every reasonable person can agree on without any analysis. In practice, propositions are based on experience, which is close to the type of study presented here. This study differs in that the propositions are taken to be reasonable but are then subjected to an empirical analysis to determine if they are reasonable in practical application.

This page titled 2.5.5: Success Factors for Mobile Work in Health Care is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.5.6: Improving Web Searches

Internet searches are designed to provide satisfactory results to the user. User satisfaction is usually defined in terms of whether the user finds the information that they were seeking. A major problem in measuring user satisfaction for web searches is that most users will be unwilling to provide direct feedback with each search. To aid with this problem a research study from 2005 considered implicit measures of user satisfaction, and whether they would provide a useful surrogate for directly measuring user satisfaction in web searches (Fox et al. 2005).

To determine if such implicit measures would be effective, the authors of the study examined the behavior of 146 internet users over a six-week period. Direct feedback was collected from these users on the satisfaction with individual search results as well as satisfaction with the entire search session. Through the use of a dialog box that popped up at the end of a search session, the user could select one response from a set of four possible responses: “I liked it,” “It was interesting but I need more information,” “I didn’t like it,” and “I did not get a chance to evaluate it.” The response “I liked it” was taken to mean that the user was satisfied with the search result. Implicit measures such as mouse and keyboard actions, whether the results were printed or were added to their favorites, and several other measures were recorded for each search as well. Sophisticated statistical modeling techniques where then used to determine if the implicit measures could be used to predict the direct feedback from the participants.

The researchers were able to find some interesting predictive results. For example, users who spent nearly a minute looking at a page that had abundant images and was in the top three results on the list were satisfied with the results nearly 90% of the time. Similarly, participants who spent less than a minute on a page, and exited the page were dissatisfied nearly 75% of the time.

This type of study is based on empiricism and the scientific method. The type of implicit measures that were selected for the study were based on previous observations in other contexts. The researchers used these propositions to form the empirical hypothesis that these activities would be useful in the current setting as well. The connection between the proposed implicit measures was then established by dividing the responses from the implicit measures in such a way that they predicted the outcome in the most efficient way. The high percentages associated with the results on the satisfaction scale then provide evidence that the implicit measures are useful in predicting the explicit measures. The researchers also took advantage of additional, more sophisticated methods to ensure that the results observed in the study were not simply due to chance but were likely due to authentic observed effects.

This page titled 2.5.6: Improving Web Searches is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.5.7: Customer Avoidance of Advertising on Social Media

Message stream advertising, advertisements that appear as part of a user's social media feed, has emerged as an important tool for marketing for businesses. This type of advertising is popular because it is often seen along with user-generated content (Campbell and Marks 2015). However, message stream advertising has the potential for consumer avoidance;, that is, the advertisement is ignored or scrolled past by the user. This is often due to skepticism about the advertisement, the intrusiveness of the advertisement on the individual's social media feed, anger about being interrupted in their social media feed, and the perceived clutter of the advertisements invading their social media feed (Tran 2017; Youn and Kim 2019; Fan et al. 2017).

Ad avoidance is defined as any action that a user may take to avoid or lessen exposure to an ad on a social media feed, such as scrolling past the unwanted ad, closing or skipping advertisement pages when possible, or by taking actions to consciously ignore the advertisement when it is encountered (Li et al. 2025). Using a conceptual framework that was developed by the authors of the study, a series of hypotheses were set and tested empirically using data gathered from the study. For example, the first hypothesis posed by the researchers is that the intrusiveness of the advertising has a positive effect on consumer's negative emotions. In other words, the researchers hypothesize that consumers will react negatively when they notice the advertising intruding on their social feed. A similar hypothesis is that ad intrusiveness has a negative effect on the consumer's perceived entertainment. Several other hypotheses were formulated for the research project that are too numerous to consider here. Within the scientific method we can assume that these hypotheses were developed based on previous observations. These hypotheses were then tested using new observations from the study.

After forming the research hypotheses, the researchers interviewed 438 users of a popular social media application where measures of both advertising intrusiveness and negative emotions for twenty example items that were presented to the participants. Using a statistical modeling technique the researchers were able to show that advertising intrusiveness is linked to negative emotions and advertising avoidance. The negative emotions included irritation, frustration, discouragement, and anger. Moreover, the study showed that users are more likely to increase their sense of control by discontinuing the advertisement when experiencing these negative emotions.

These results demonstrate the next phase of the scientific method, which is observation. These observations can now be used to form new theories and hypotheses about the behavior of users on social media and their reactions to intrusive advertising. The results may be used to change advertising methodology, thus creating new hypotheses about the effect of using these new methods, and the cycle begins anew.

This page titled 2.5.7: Customer Avoidance of Advertising on Social Media is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.6: Summary

The concept of using observation to learn about the world dates to ancient Roman philosophy. The Copernican Revolution demonstrated the potential of this method by developing models for the dynamic movement of objects through time and space. The philosophy of obtaining knowledge through observation was provided with an influential philosophical framework through the work of British philosophers in the 16th and 17th centuries, culminating in what became known as British Empiricism. The modern scientific method, an iterative process of observation, hypothesis, and testing, is used extensively in scientific research to this day, though many aspects of the process are still the subject of intense debate. Feminist empirical philosophers acknowledge that the social context of observations can have a fundamental effect on the scientific method. Feminist empiricism provides modifications to the modern scientific method that account for this effect.

This page titled 2.6: Summary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.7: About the Quote

Agatha Christie (1890–1976) was an English writer primarily known for her detective fiction and was the creator of the fictional detectives Hercule Poirot and Jane Marple. Her play *The Mousetrap* is the world's longest running play, and she is the best-selling fiction author of all time. She was made a Dame by Queen Elizabeth II for her contributions to literature. The quote at the beginning of the chapter is taken from the book The Mysterious Affair at Styles. See Figure **2.7.1**.



Figure **2.7.1**: Agatha Christie as a young woman. (Public Domain, Author Unknown)

This page titled 2.7: About the Quote is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.8: Practice and Exploration

1. In the chapter we postulated that “all cats are fuzzy.”. We observed that Felix is a cat and then concluded that Felix is fuzzy using a deductive argument. Now suppose that we are researching cats online and see a picture of a Sphynx cat, a breed of cat that is hairless, and is not fuzzy. Our argument has broken down. Why does our theory not work in this case, and what does this say about using a deductive argument in the scientific method?
2. As students enter a classroom the instructor of the class notices that all the students have a cell phone. When class begins the instructor remembers that they did not close their office door and hence needs to borrow a cell phone from one of the students to call the office manager of the department to ask them to close their office door. The instructor reasons that it does not matter which student they ask because they will have a cell phone. Is the instructor using a deductive or inductive argument to make this conclusion?
3. A chef is making a cake and has cracked three eggs from a carton. Each of the three eggs were good eggs and from this this chef concludes that all the eggs in the carton are good. Is the chef using a deductive or inductive argument to draw this conclusion?
4. Suppose that a chemist does an experiment in their lab showing that a certain chemical acts as a catalyst for a chemical reaction, that is, the chemical speeds up the rate of the chemical reaction. After checking their lab work to be sure they have done the experiment carefully, they order a large amount of this catalyst anticipating an increase of these chemical reactions in future experiments. Is the chemist using deductive or inductive reasoning?
5. While getting breakfast you notice that there is a sign at your favorite donut shop that states, “All of our donuts contain sugar.” You purchase a donut, and while giving it to your partner you tell them that “this donut contains sugar.” Are you using inductive or deductive reasoning to make this statement?
6. An employee at an automobile repair shop is told that one of the four tires on a pickup truck is low on air. The employee checks three of the tires and they seem fine and therefore concludes that the fourth tire is the one that is low on air. What type of reasoning is the employee using to make this conclusion?
7. On a weekend trip you and a friend hear the strains of live music and head to a local music venue. The next band comes on stage, and it is a local band that you have never seen and know nothing about. They play five songs, and all five songs are classic songs by the band The Grateful Dead. Your friend looks at you and says, “The next song will be a Grateful Dead Song.” Is your friend using inductive or deductive reasoning to make this conclusion?
8. In the evening after work you and your partner come home and notice that the lamp has been knocked over and broken. Your partner is not surprised and blames your two cats for knocking down the lamp. You, however, construct a scenario that involves intruders breaking into your house undetected, knocking over the lamp, not disturbing anything else in the house and leaving. If a bystander were to conclude that your partner is probably correct, what type of reasoning might they use to back up their conclusion?
9. During a long day at work you decide that it is time for some refreshment. You gather the quarters from your desk and head toward the closest soft drink machine. There are five people in front of you. One by one, each person inserts their money into the machine and after some pushing of buttons and an audible expression of anger from each person, they leave without their money or a refreshing soft drink. Now it is your turn. Will you put your money into the machine to attempt to get a soft drink? Why or why not? Why type of reasoning are you using in making your decision?

This page titled 2.8: Practice and Exploration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.9: Collaborative Activity

When one thinks of deduction, one often thinks of the fictional character Sherlock Holmes, a detective introduced by British author Sir Arthur Conan Doyle in the 1887 novel A Study in Scarlet. One of the reasons why Sherlock Holmes became so famous and so memorable is that the character was inspired by Joseph Bell, a surgeon at the Royal Infirmary of Edinburgh. Conan Doyle met Dr. Bell in 1877 and worked for him as a clerk. Dr. Bell was noted for drawing broad conclusions from minute observations. The character of Holmes is also known for making such conclusions, often to the surprise of those around him. In reading the stories, one finds that Holmes uses both deduction and induction in making his conclusions.

In this collaborative activity, you and your group will read a passage from the short story “The Red Headed League,” which was written by Conan Doyle and published in the Strand Magazine in August 1891. In the story, Holmes’ associate Dr. Watson arrives at their quarters to find Holmes talking to Jabez Wilson, a London pawnbroker. Holmes invites Watson to stay for the consultation at which point we pick up the story.

> I did not gain very much, however, by my inspection. Our visitor bore every mark of being an average commonplace British tradesman, obese, pompous, and slow. He wore rather baggy gray shepherd's check trousers, a not over-clean black frock coat, unbuttoned in the front, and a drab waistcoat with a heavy brassy Albert chain, and a square pierced bit of metal dangling down as an ornament. A frayed top hat and a faded brown overcoat with a wrinkled velvet collar lay upon a chair beside him. Altogether, look as I would, there was nothing remarkable about the man save his blazing red head and the expression of extreme chagrin and discontent upon his features.
>
> Sherlock Holmes's quick eye took in my occupation, and he shook his head with a smile as he noticed my questioning glances. "Beyond the obvious facts that he has at some time done manual labor, that he takes snuff, that he is a Freemason, that he has been in China, and that he has done a considerable amount of writing lately, I can deduce nothing else."
>
> Mr. Jabez Wilson started up in his chair, with his forefinger upon the paper, but his eyes upon my companion.
>
> How, in the name of good fortune, did you know all that, Mr. Holmes?" he asked. "How did you know, for example, that I did manual labor? It's as true as gospel, for I began as a ship's carpenter."
>
> "Your hands, my dear sir. Your right hand is quite a size larger than your left. You have worked with it and the muscles are more developed."
>
> "Well, the snuff, then, and the Freemasonry?"
>
> "I won't insult your intelligence by telling you how I read that, especially as, rather against the strict rules of your order, you use an arc and compass breastpin."
>
> "Ah, of course, I forgot that. But the writing?"
>
> "What else can be indicated by that right cuff so very shiny for five inches, and the left one with the smooth patch near the elbow where you rest it upon the desk."
>
> "Well, but China?"
>
> "The fish which you have tattooed immediately above your wrist could only have been done in China. I have made a small study of tattoo marks, and have even contributed to the literature of the subject. That trick of staining the fishes' scales of a delicate pink is quite peculiar to China. When, in addition, I see a Chinese coin hanging from your watch chain, the matter becomes even more simple."
>
> Mr. Jabez Wilson laughed heavily. "Well, I never!" said he. "I thought at first that you had done something clever, but I see that there was nothing in it after all."

For this activity have one member of your group read the passage out loud to the other group members. After reading the passage the group should identify each spot where Holmes draws a conclusion from what was observed about the appearance of Mr. Wilson. Discuss each conclusion, answering the following questions for each:

1. Is the conclusion based on inductive or deductive reasoning?

2. Are there other possible explanations for each conclusion? In other words, is it possible that Holmes could be wrong?
3. How does the role of reasoning to the best explanation play a part in the conclusions that Holmes makes?

This page titled 2.9: Collaborative Activity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

2.10: References

Aarts, Kees. 2007. “Parsimonious Methodology.” *Methodological Innovations Online* 2, no. 1, 2–10. https://doi.org/10.4256/mio.2007.0002

Ammenworth, E., Buchauer, A., Bludau, B., and Haux, R. 2000. “Mobile Information and Communication Tools in the Hospital.” *International Journal of Medical Informatics* 57, no. 1, 21–40. https://doi.org/10.1016/S1386-5056(99)00056-8

Bertrand, Marianne, and Sendhil Mullainathan. 2004. “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review* 94, no. 4, 991–1013. https://doi.org/10.1257/0002828042002561

Bigg, Charlotte. 2008. “Evident Atoms: Visuality in Jean Perrin’s Brownian Motion Research.” *Studies in History and Philosophy of Science Part A* 39, no. 3, 312–322. https://doi.org/10.1016/j.shpsa.2008.06.003

Bird, Sharon R., and Laura A. Rhoton. 2021. “Seeing Isn’t Always Believing: Gender, Academic STEM, and Women Scientists’ Perceptions of Career Opportunities.” *Gender and Society* 35, no. 3, 422–448. https://doi.org/10.1177/08912432211008814

Black, Dan A., Seth G. Sanders, and Lowell J. Taylor. 2007. “The Economics of Lesbian and Gay Families.” *Journal of Economic Perspectives* 21, no. 2, 53–https://doi.org/70.10.1257/jep.21.2.53

Campbell, C., and L. J. Marks. 2015. “Good Native Advertising Isn’t a Secret.” *Business Horizons* 58, no. 6, 599–606. https://doi.org/10.1016/j.bushor.2015.06.003

Casad, Bettina J., Jillian E. Franks, Christina E. Garasky, Melinda M. Kittleman, Alanna C. Roesler, Deidre Y. Hall, and Zachary W. Petzel. 2021. “Gender Inequality in Academia: Problems and Solutions for Women Faculty in STEM.” *Journal of Neuroscience Research* 99, no. 1, 13–23. https://doi.org/10.1002/jnr.24631

Cavallo, Ann M. L., and David McCall. 2008. “Seeing May Not Mean Believing: Examining Students’ Understandings and Beliefs in Evolution.” *The American Biology Teacher* 70, no. 9, 522–530. https://doi.org/10.2307/27669336

Chatterjee, S., S. Chakraborty, S. Sarker, S. Sarker, and F. Y. Lau. 2009. “Examining the Success Factors for Mobile Work in Healthcare: A Deductive Study.” *Decision Support Systems* 46, 620–633. https://doi.org/10.1016/j.dss.2008.11.003

Cialdini, Robert B., and Noah J. Goldstein. 2004. “Social Influence: Compliance and Conformity.” *Annual Review of Psychology* 55, no. 1, 591–621. https://doi.org/10.1146/annurev.psyc....090902.142015

Cobern, William W. 1994. “Comments and Criticism. Point: Belief, Understanding, and the Teaching of Evolution.” *Journal of Research in Science Teaching* 31, no. 5, 583–590. https://doi.org/10.1002/tea.3660310511

Coffman, Katherine B., Lucas C. Coffman, and Keith M. Marzilli Ericson. 2017. “The Size of the LGBT Population and the Magnitude of Antigay Sentiment are Substantially Underestimated.” *Management Science* 63, no. 10, 3168–3186. https://doi.org/10.1287/mnsc.2016.2503

Crivellaro, Fredrica, and Alessandra Sperduti. 2014. “Accepting and Understanding Evolution in Italy: A Case Study from a Selected Public Attending a Darwin Day Celebration.” *Evolution: Education and Outreach* 7, no. 1, 1–8. https://doi.org/10.1186/s12052-014-0013-4

Curd, Martin, and Stathis Psillos, Eds. 2014. *The Routledge Companion to Philosophy of Science*. Routledge. https://www.routledge.com/The-Routle...yURpni33uVOotA

DeVantier, L. M. 1992. “Rafting of Tropical Marine Organisms on Buoyant Coralla.” *Marine Ecology–Progress Series* 86, 301–301. https://doi.org/10.3354/meps086301

Diogo, Rui. 2020. “Quasi-Religious Belief in Darwin and Darwinism: “Straw-Men” Scientist Believers Everywhere.” *Current Molecular Biology Reports* 6, no. 2, 16–31. https://doi.org/10.1007/s40610-020-00127-y

Fan, S., Y. Lu, and S. Gupta. 2017. “Social Media In-Feed Advertising: The Impacts of Consistency and Sociability on Ad Avoidance.” *PACIS 2017 proceedings*, 190. https://aisel.aisnet.org/pacis2017/190

Flores, Andrew R., and Scott Barclay. 2016. “Backlash, Consensus, Legitimacy, or Polarization: The Effect of Same-Sex Marriage Policy on Mass Attitudes.” *Political Research Quarterly* 69, no. 1, 43–56. https://doi.org/10.1177/1065912915621175

Fox, Steve, Kuldeep Karnawat, Mark Mydland, Susan Dumais and Thomas White. 2005. “Evaluating Implicit Measures to Improve Web Search.” *ACM Transactions on Information Systems* 23, No. 2, 147–168. https://doi.org/10.1145/1059981.1059982

Friday, A. E. 1982. “Parsimony, Simplicity and What Actually Happened.” *Zoological Journal of the Linnean Society* 74, no. 3, 329–335. https://doi.org/10.1111/j.1096-3642.1982.tb01155.x

Fumerton, Richard. 2009. *Epistemology*. John Wiley & Sons. https://www.wiley.com/en-us/Epistemo...-9781405150354

Gilson, Etienne. 2009. *From Aristotle to Darwin and Back Again: A Journey in Final Causality, Species, and Evolution*. Ignatius Press. https://ignatius.com/from-aristotle-...-again-fadbap/

Goldstein, H., Poole, C. P., Safko, J. L. 2002. *Classical Mechanics*. Addison Wesley. https://www.google.com/books/edition...QgAACAAJ?hl=en

Hannam, James. 2011. *The Genesis of Science: How the Christian Middle Ages Launched the Scientific Revolution*. Simon and Schuster. https://www.simonandschuster.com/boo.../9781596982055

Harding, Sandra G. 1994. “Is Science Multicultural?: Challenges, Resources, Opportunities, Uncertainties.” *Configurations* 2, no. 2, 301–330. https://doi.org/10.1353/con.1994.0019

Harding, Sandra G. 1998. *Is Science Multicultural? Postcolonialisms, Feminisms, and Epistemologies*. Indiana University Press. https://iupress.org/9780253211569/is...multicultural/

Harrison, Peter, Ronald L. Numbers, and Michael H. Shank. 2011. *Wrestling with Nature: From Omens to Science*. University of Chicago Press. https://press.uchicago.edu/ucp/books...o11045791.html

Haw, Marc D. 2002. “Colloidal Suspensions, Brownian Motion, Molecular Reality: A Short History.” *Journal of Physics: Condensed Matter* 14, no. 33. 7769. https://doi.org/10.1088/0953-8984/14/33/315

Hildering, Peter, Luca Consoli, and Riyan van den Born. 2013. “Denying Darwin: Views on Science in the Rejection of Evolution by Dutch Protestants.” *Public Understanding of Science* 22, no. 8, 988–998. https://doi.org/10.1177/0963662512437328

Hooghe, Marc, and Cecil Meeusen. 2013. “Is Same-Sex Marriage Legislation Related to Attitudes Toward Homosexuality?” *Sexuality Research and Social Policy* 10, no. 4, 258–268. https://doi.org/10.1007/s13178-013-0125-6

Huemer, Michael. “When Is Parsimony a Virtue?” *The Philosophical Quarterly* 59, no. 235, 216–236. https://doi.org/10.1111/j.1467-9213.2008.569.x

Hundleby, Catherine. 2011. *Feminist Empiricism. Handbook of Feminist Research: Theory and Praxis*. Sage Publishing. https://doi.org/10.4135/9781483384740

Katsikadelis, John T. 2015. “Derivation of Newton’s Law of Motion Using Galileo’s Experimental Data.” *Acta Mechanica* 226, no. 9, 3195–3204. https://doi.org/10.1007/s00707-015-1354-y

Klug, Aaron. 1968. “Rosalind Franklin and the Discovery of the Structure of DNA.” *Nature* 219, 808–810. https://doi.org/10.1038/219808a0

Kreitzer, Rebecca J., Allison J. Hamilton, and Caroline J. Tolbert. 2014. “Does Policy Adoption Change Opinions on Minority Rights? The Effects of Legalizing Same-Sex Marriage.” *Political Research Quarterly* 67, no. 4, 795–808. https://doi.org/10.1177/1065912914540483

Letherby, Gayle. 2003. *Feminist Research in Theory and Practice*. McGraw-Hill Education. https://archive.org/details/feministresearch0000leth

Li, X., Z. Liu, Y. Chen, and A. Ren. 2025. “Consumer Avoidance Toward Message Stream Advertising on Mobile Social Media: A Stimulus-Organism-Response Perspective.” *Information Technology and People* 38, no. 1, 23–47. https://doi.org/10.1108/ITP-11-2020-0761

Locke, John. 1974. *The Empiricists*. Anchor. https://archive.org/details/the-empi...-berkeley-hume

Longino, Helen E. 2018. *The Fate of Knowledge*. Princeton University Press. https://press.princeton.edu/books/pa...dfizKNg6_xJXxf

Longino, Helen E. 1990. *Science as Social Knowledge*. Princeton University Press. https://press.princeton.edu/books/pa...zvAV_UaXForR61

Maiocchi, Roberto. 1990. "The Case of Brownian Motion." *The British Journal for the History of Science* 23, no. 3, 257–283. https://doi.org/10.1017/S0007087400043983

Marchesoni, F., and Hänggi, P. 2005. "Introduction: 100 Years of Brownian Motion." *Chaos* 15, 026101. https://doi.org/10.1063/1.1895505

Mastroianni, Michael A., Jennifer A. Kunes, Dany B. El-Najjar, Kyle K. Obana, Sohil S. Desai, Cole R. Morrissette , Frank J. Alexander, Alexander J. Rondon, David P. Trofa, Charles A. Popkin, William N. Levine, Christopher S. Ahmad. 2025. "Return to Performance After Ulnar Collateral Ligament Surgery in Major League Baseball Pitchers." *The American Journal of Sports Medicine* 53, no. 1, 104–114. https://doi.org/10.1177/03635465241289142

McNeal, Ann P. 1994. "Lifting the Veil: The Feminine Face of Science." *BioScience* 44, no. 3, 183–185. https://doi.org/10.2307/1312258

Nelson, Lynn. 2010. *Who Knows: From Quine to a Feminist Empiricism*. Temple University Press. https://archive.org/details/whoknowsfromquin0000nels

Nikolaou, Dimitrios. 2022. "Same-Sex Marriage Laws, LGBT Hate Crimes, and Employment Discrimination Charges." *Southern Economic Journal* 88, no. 3, 869–905. https://doi.org/10.1002/soej.12548

Okasha, S. 2002. *Philosophy of Science: A Very Short Introduction.* Oxford University Press. https://academic.oup.com/book/517

Popper, K. R. 1989. *Conjectures and Refutations.* Routledge. https://archive.org/details/karl-pop...ific-knowledge

Russell, Bertrand. 2013. *History of Western Philosophy: Collector's Edition*. Routledge. https://books.google.com/books/about...d=Gm_cCZBiOhQC

Schlesinger, George. "Confirmation and parsimony." *Induction, Probability, and Confirmation* 6, 324–342. https://www.jstor.org/stable/10.5749/j.ctttsz9j

Schwerter, Jacob and Lena Ilg. 2021. "Gender Differences in the Labour Market Entry of stem STEM Graduates." *European Journal of Higher Education* 13, no. 3, 1–19. https://doi.org/10.1080/21568235.2021.2010226

Simon, Herbert A. 2002. "Science Seeks Parsimony, not Simplicity: Searching for Pattern in Phenomena." In A. Zellner, H. A. Keuzenkamp, M. McAleer, eds. *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple*. Cambridge University Press. 32–72. https://doi.org/10.1017/CBO9780511493164.003

Skiba, Russell J., Robert H. Horner, Choong-Geun Chung, M. Karega Rausch, Seth L. May, And Tary Tobin. 2011. "Race Is Not Neutral: A National Investigation of African American and Latino Disproportionality in School Discipline." *School Psychology Review* 40, no. 1, 85–107. https://doi.org/10.1080/02796015.2011.12087730

Stewart, Caro-Beth. 1993. "The Powers and Pitfalls of Parsimony." *Nature* 361, 603–607. https://doi.org/10.1038/361603a0

Stewart-Williams, Steve, and Lewis G. Halsey. 2021. "Men, Women and STEM: Why the Differences and What Should Be Done?" *European Journal of Personality* 35, no. 1, 3–39. https://doi.org/10.1177/0890207020962326

Tankard, Margaret E., and Elizabeth Levy Paluck. 2016. "Norm Perception As a Vehicle for Social Change." *Social Issues and Policy Review* 10, no. 1, 181–211. https://doi.org/10.1111/sipr.12022

Tankard, Margaret E., and Elizabeth Levy Paluck. 2017. "The Effect of a Supreme Court Decision Regarding Gay Marriage on Social Norms and Personal Attitudes." *Psychological Science* 28, no. 9, 1334–1344. https://doi.org/10.1177/0956797617709594

Thagard, Paul, and Scott Findlay. 2010. "Getting to Darwin: Obstacles to Accepting Evolution by Natural Selection." *Science and Education* 19, no. 6, 625–636. https://doi.org/10.1007/s11191-009-9204-8

Tran, T. P. 2017. "Personalized Ads on Facebook: An Effective Marketing Tool for Online Marketers." *Journal of Retailing and Consumer Services* 39, 230–242. https://doi.org/10.1016/j.jretconser.2017.06.010

Van de Sande, Adje, and Christie Byvelds. 2015. *Statistics for Social Justice. A Structural Perspective*. Fernwood Publishing. https://cup.columbia.edu/book/statis...olitical%20and

Waugh, Joanne, and Roger Ariew. 2008. “The History of Philosophy and the Philosophy of Science.” In *The Routledge Companion to Philosophy of Science*, 43–53. Routledge. https://www.routledge.com/The-Routle...u7eQ_dFEMZjlA0

Youn, S., and Kim, S. 2019. “Understanding Ad Avoidance on Facebook: Antecedents and Outcomes of Psychological Reactance.” *Computers in Human Behavior* 98, 232–244. https://doi.org/10.1016/j.chb.2019.04.025

This page titled 2.10: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

CHAPTER OVERVIEW

3: Experiments

“It does not make any difference how beautiful your guess is. It does not make any difference how smart you are, who made the guess, or what his name is—if it disagrees with the experiment, it is wrong. That is all there is to it.”

Richard P. Feynman

- 3.1: What to Expect
- 3.2: Experiments
- 3.3: Designed Experiments
- 3.4: Observational Studies
- 3.5: Case Control Studies
- 3.6: Unethical Experiments
- 3.7: Critical Analysis
  - 3.7.1: Labor Market Discrimination
  - 3.7.2: Visual Exercise Environments and Treadmill Running
  - 3.7.3: Same-Sex Marriage Laws and Hate Crimes
  - 3.7.4: Effects of a New Special Formula Fertilizer on Potato Growth
  - 3.7.5: The Supreme Court and Personal Attitudes
  - 3.7.6: Sleep and College Students
  - 3.7.7: Gender and COVID-19 Outcomes
  - 3.7.8: Thought and the Growth of Fungi Cultures
  - 3.7.9: Risk Factors for Homelessness
- 3.8: Summary
- 3.9: About the Quote
- 3.10: Practice and Exploration
- 3.11: Collaborative Activity
- 3.12: References

This page titled 3: Experiments is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.1: What to Expect

This chapter discusses the types of experiments that are used to implement the scientific method in practice. The realities of the mechanisms involved, along with the form of the hypotheses being studied, determine the circumstances under which observations should be made to determine the truth of a hypothesis. In some cases, the proper form of an experiment is impossible, impractical, or unethical to use in practice, and the corresponding effect that this can have on the veracity of the conclusions based on using the scientific method need to be carefully investigated. This chapter also discusses instances where unethical experiments were used in the twentieth century. By the end of this chapter, you should be able to:

1. Discuss, define, and contrast designed experiments and observational studies.
2. Define factor, response, treatment group, and control group.
3. Identify factors by which an experiment can be considered unethical.
4. Critically analyze the experimental design used in a research article.

This page titled 3.1: What to Expect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.2: Experiments

In the previous chapter we considered the development of empiricism and the scientific method. The key step of the scientific method is the test step where observation is used to provide evidence about the truth of the hypothesis specified in the previous step. The method that produces these observations is called an experiment.

Definition: Experiment

An *experiment* is any process that produces observations that provide evidence that can be used to determine if a hypothesis is true or not.

There are many types of hypotheses and experiments. We will describe some of the more common types used in modern applications of statistics. In the most common formulation, researchers are interested in whether a certain condition is related to or associated with another condition. For example, a researcher might hypothesize that individual income is related to race or ethnicity. In this case individual income and race are the conditions that are of interest to the researcher in forming a hypothesis. Note that there is a natural direction to this hypothesis in that implicitly the idea that the race of an individual is at least associated with, to some extent, the income of the individual. To see the direction of this relationship, a researcher would hopefully not hypothesize that individual income influences the race of the individual. If it is observed that families of certain races tend to have lower measures of wealth than families of other races, then evidence has been observed that the race of a family is a condition that partially determines family wealth. As we shall see later in this book, this is only partial evidence of such a relationship.

In this type of formulation, race is known as a factor or a treatment, and family wealth is called the response. In the general formulation of these experiments, researchers are usually interested in how the response is related to the factor or factors observed in an experiment.

Definition: Response

In an experiment the response is the observed condition that is of interest to the researcher.

Definition: Factor or Treatment

In an experiment a *factor* or *treatment* is a condition that is hypothesized to have an effect on, or is associated with, the response.

Considering another example, suppose that a researcher hypothesizes that the political affiliation of an individual is a factor that partially determines the network that the individual trusts as their news source. In this case the response is the network of the news source of the individual and the factor is the political affiliation. Expanding on this example the researcher might also be interested in how gender and age also determine the network of the news source of an individual. For this example the response is still the network of the news source, but now there are three factors: political affiliation, gender, and age. Note that it is possible that an experiment may have more than one response as well, though we will not discuss these studies except in a few examples.

It can also happen that an experiment does not have any factors, and only responses. For example, a researcher might hypothesize that the average age of students at a community college is greater than 25 years. The researcher would only observe the response which in this case is age, and there are no factors. In some other cases researchers may observe more than one measurement but no directional relationship between the measurements is hypothesized. In this type of study, the researcher may be interested in how well the measurements mimic one another without hypothesizing that one condition is causing the other condition to behave in a certain way.

This page titled 3.2: Experiments is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.3: Designed Experiments

In many classical scientific experiments, we envision a scientist who sets up the conditions of the experiment and then observes what happens as a result. Within the framework of the scientific method, the conditions of the experiment are stated in the hypothesis. That is, the hypothesis usually states that if the conditions are set in a certain way, then a certain outcome will be observed. The purpose of the experiment is then to observe if the hypothesized outcome occurs, which would provide evidence that the hypothesis is correct. Alternatively, if the hypothesized outcome does not occur, then this would refute the hypothesis. In this type of experiment the conditions set by the researchers are the factors, and what is observed to occur is the response. This type of experiment is called a designed experiment.

Definition: Designed Experiment

A *designed experiment* is an experiment where the factors are controlled and set by the researcher prior to observing the outcome.

Designed experiments are common in scientific applications because the factors are usually relatively easy to control. For example, one can study the effect on the acceleration of a marble rolling down an inclined plane at different angles by setting the inclined plane at several angles and observing how long it takes the marble to get to the bottom. In social science research these experiments are less common, but they do occur. The most common designed experiments in social justice research are usually psychological experiments that attempt to assess how a stimulus, taken to be a controlled factor, affects the attitude of an individual about a social issue, which is taken to be the response. For example, researchers might measure the empathy of two groups of individuals after viewing two different versions of a documentary. Of interest would be to observe whether there is evidence that the viewpoint shown in the documentary influences the individuals observed in the study. Because the researchers control which version of the documentary each individual views, this is a designed experiment.

A related type of experiment compares individuals who are treated in a special way to those who receive a standard treatment. For example, a researcher may divide a group of students into two groups. One of the groups receives a standard study guide for the final exam in a specific course, while the second group of students is given an enhanced version of the study guide. The purpose of the study is then to observe whether the enhanced study guide improves student performance on the final exam. In a study using this type of framework, the group that receives the standard treatment is called the control group and the group that receives the special treatment is called the treatment group.

Definition: Control Group

In an experiment that compares a standard treatment to a special treatment, the group that receives the standard treatment is called the *control group*.

Definition: Treatment Group

In an experiment that compares a standard treatment to a special treatment, the group that receives the special treatment is called the *treatment group*.

The idea of this type of study is to compare the special treatment, usually a new or potential improvement on the standard treatment, to a standard or established treatment. By including individuals who are given the standard treatment (the control group) we have a baseline for comparison to those given the new or special treatment (the treatment group). Without this baseline observation it is impossible to determine if the special treatment is making any difference. For example, if the researcher wants to determine if an enhanced study guide improves student performance, a measure of student performance using the standard study guide must be used for comparison.

This idea can be expanded to compare several treatments to a control group, or several treatment groups to one another. So, for example, a researcher could compare several types of study guides to observe evidence about which of these approaches is the best, or if any of the approaches is better than the standard study guide.

A crucial question related to the use of a designed experiment is to determine how the individuals in a study are assigned to the treatment and control groups. Considering the example of comparing the two study guides for the final exam, the researchers could give the enhanced study guide to those students whose grades are in the bottom half of the class while giving the standard study guide to those students whose grades are in the top half of the class. This may seem like a good idea in that the enhanced study guide should help the students who are having trouble with the class. Unfortunately, from the perspective of the scientific method, this would be a poor choice. If, for example, the grades for the students who had the enhanced study guide were lower on average than for the students who had the standard study guide, then it would be impossible to know if this was because the enhanced study guide did not work or if it was because the students tended not to do well in the class to begin with. Issues with determining what individuals will be in the treatment group and which individuals will be in the control group will be considered in greater detail in the next section. The best way to choose the participants for each group is to do it based on a mechanism that is completely unrelated to student performance. The easiest method for implementing such a selection methodology, randomization, will be covered later in the book.

This page titled 3.3: Designed Experiments is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.4: Observational Studies

Many studies cannot be studied using designed experiments because the researcher does not control the conditions under which the experimental data is observed. There are several reasons for this. Sometimes it is simply not possible to control the factors of interest. If you are interested in animal behavior prior to an earthquake, it is not possible to schedule an earthquake of a specified magnitude to occur next week so that you can observe how the animals behave beforehand. In other cases, there are moral and social prohibitions to controlling factors. If you are interested in how long-term exposure to nicotine affects the likelihood of lung cancer, it is simply not ethical to expose individuals to nicotine for the purpose of such an experiment. Similarly, if we are interested in how the poverty rate is related to the likelihood of incarceration, it is not ethical to deprive a specified population from economic advantage in the interest of the experiment.

In these cases what are known as observational studies are used instead. In observational studies the factors thought to influence the outcome, along with the outcome, are observed at the same time. Implicit in this structure is that the researcher does not have control over these factors.

Definition: Observational Study

An *observational study* is an experiment where the researcher does not control or set the factors.

In observational studies the individuals in the study decide for themselves what their behavior will be, and the researchers simply observe the behavior. For example, studies on the effect of nicotine exposure through cigarette smoking are observational studies usually conducted over large populations where the presence of lung cancer and the amount of exposure to cigarette smoking is observed in individuals through a health survey. Many of the studies presented thus far are based on observational studies. For example, if researchers are interested in how the poverty rate is related to the likelihood of incarceration, then they might look at survey data that both observes the income and past incarceration of survey participants.

The main problem with observational studies is the difficultly of attributing what is observed in the experiment to any factor. This is particularly a problem when many possible factors may influence observations, and this problem may be compounded when the factors themselves are related or associated with one another. This problem is known as confounding.

Definition: Confounding

Two or more factors are *confounded* when the effect of one of the factors cannot be distinguished from the effects of the other factors on the outcome.

In the case of attempting to link lung cancer to cigarette smoking, the use of observational studies could only suggest cigarette smoking as a possible cause. For example, one could argue that there is possibly a genetic link to lung cancer that is responsible for the increased likelihood of contracting the disease, and this genetic condition is also linked to nicotine addiction. Hence, it is the genetic effect that is causing the lung cancer and the addiction to cigarette smoking. However, data from an observational study would clearly indicate that nicotine exposure and the incidence of lung cancer are related, and concluding that there is a link between the two would be tempting.

Early in the study of linking the incidence of lung cancer to cigarette smoking, even famous statisticians such as Joseph Berkson and Sir Ronald A. Fisher did not believe that there was a link (Freedman et al. 2007). Only very carefully designed studies can be used to try to separate out these types of confounding problems. Today it is a known fact in the scientific community that there is a link between cigarette use and the likelihood of developing lung cancer (Witschi 2001; Proctor 2012; Musk and De Klerk 2003; Doll 1998; Wynder 1997; Proctor 2004; de Groot et al. 2018).

Observational studies should be implemented to ensure that confounding is as small a problem as possible, though it must be acknowledged that confounding will always be an issue in these types of studies. One common practice is to limit the scope of the study to reduce the number of factors that could have confounding problems. This is because any factors that differ between comparison groups have the potential to be the cause of any observed differences. For example, in a study of the effect of poverty on health, the scope of the study may be limited to individuals in a certain neighborhood since there may be geographical issues,

such as environmental factors, that could also affect health. One could well imagine that a neighborhood with a lower income level might also be closer to an industrial area where pollution may be a problem, while a comparatively more affluent neighborhood might not have that problem. If a difference in health is observed between the two areas, it cannot be determined whether the exposure to pollution or the poverty would be to blame.

A major problem is that it is not always clear what potential confounding factors could be present, and it may be quite difficult to control for these factors even when they are known. As an example, consider the Coronary Drug Project, a carefully designed study to evaluate drugs and their potential for the prevention of heart attacks (Coronary Drug Project Research Group 1973). To reduce potential confounding problems the study focused on 8,341 middle-aged men with heart trouble. One potentially helpful drug included in the study was the cholesterol reducing drug clofibrate. Over the range of the study the mortality rate of people not taking any drug, the control group, and the mortality rate of people taking clofibrate was roughly the same, so it appeared that the drug was not helpful. However, there was a problem with the data in that some individuals in the study did not take most of the drug, and hence were not participating. A closer look at the data revealed that those who took less of the drug had higher mortality rates than people who took more of the drug. Does this mean the drug works? Maybe not, as it is possible that those who took the drug lived generally healthier lifestyles than those who did not. These issues make it very difficult to determine if the drug was effective or not (Coronary Drug Project Research Group 1980; Freedman et al. 2007).

Further examples of confounding problems can be considered when assessing treatments for substance abuse. To judge whether a treatment is effective, changes in an individual’s behavior, over and above what would have occurred without treatment, is usually used to quantify change attributable to treatment (Pierce et al. 2017; Prochaska et al. 1993). This is usually accomplished by comparing the overall behavior of groups of individuals who had the treatment to other groups of individuals who did not have the treatment. The effectiveness of a treatment is often quantified by comparing outcomes between periods when subjects are being treated with periods when they are not. One potential problem is that individuals with substance problems often cycle in and out of treatment, and therefore the application of the treatment varies with time (Condron et al. 2013). In these cases, for example, this can cause a confounding issue with past substance abuse, in that the effectiveness of the current treatment cycle could very well be influenced by the frequency and type of substance abuse in the patient's history as well as the number and types of treatments they had received in the past (Pierce et al. 2017).

Considering research in social justice, many studies point to an apparent relationship between race and domestic violence. In surveys, studies of calls to the police, and case studies of women’s shelters and emergency rooms consistently find a disproportionate share of African American offenders and victims (Stets 1991). The inherent conclusion from this research includes the disturbing undertone that there is a link between race and domestic violence. This research often avoids addressing the issue of why these studies show these types of results (Benson et al. 2003, 2004; Benson and Greer 2002).

One study considered this problem with a more sophisticated analysis that was based on the idea that a variety of social forces destroyed the social fabric of African American urban communities, leading to a situation in which a large proportion of the African American population are trapped in socially isolated urban areas with high levels of unemployment. Such environments may feed into a culture that is more tolerant of violence and crime (Wilson 2012; Sampson and Wilson 2020). If this hypothesis is true, then it may follow that an association between domestic violence and race could be explained by the fact that African Americans tend to live in areas that whose social fabric has been destroyed by poverty and unemployment. Therefore, race is confounded with the local economic climate.

To explore this idea, census and survey data were used to carefully compare the domestic violence rates of African Americans and whites who live in comparable economic situations (Benson et al. 2003). The data showed that the domestic violence rates were similar for both races in the same economic environment. This research supports the hypothesis that it is the economic environment, and not race, that is related to domestic violence rates.

This page titled 3.4: Observational Studies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.5: Case Control Studies

Many observational studies are based on looking at the current conditions of a population, and tracing back possible causes in the past. For example, a researcher might hypothesize that children who are exposed to domestic violence at home are more likely to be incarcerated later in life. The researcher is interested in comparing the outcomes of children who were not exposed to domestic violence, the control group, to children who were, the treatment group. One way to do this study would be to survey individuals who had been exposed to domestic violence and a similar number of individuals who had not been exposed to domestic violence in their childhood. While this is a perfectly valid method for conducting such a study, there is a practical problem: The percentage of incarcerated individuals is generally quite small, and therefore it is unlikely that any incarcerated individuals will be observed in the study, regardless of whether they were exposed to domestic violence. Hence, the comparison between the two groups would not be useful. One way to solve this problem is to simply observe as many people as is possible. The practical problem with this solution is the expense and time involved. A much better approach is to design the study so that incarcerated individuals are forced to be observed.

For this second approach the researchers could decide to simply interview many individuals who have been incarcerated and several individuals who have not. The researchers could then compare the rate at which individuals who were incarcerated were exposed to domestic violence compared to those individuals who were not incarcerated. In this version of the study, things have been turned around. Instead of comparing rates of incarceration for those who were exposed to domestic violence against those who were not, the researchers would be comparing the rates of domestic violence for those who were incarcerated against those who were not, even though the researchers are still interested in whether exposure to domestic violence affects incarceration rates. This type of study is known as a case-control study.

Definition: Case Control Study

A *case control study* is an observational study that observes data about past differences between two groups.

A case control study gathers evidence about the differences between two groups, but in a less direct way than a controlled experiment. Because the assignment of groups is not controlled by the researcher, and the study is therefore observational, the interpretation of case-control studies is less straightforward than a designed experiment (Shadish et al. 2002). Potential confounding factors can present real challenges when interpreting the results of case-control studies. In the example of studying the effect of exposure to domestic violence, the observation that more of the individuals who were incarcerated were exposed to domestic violence could be caused by some other factor associated with both outcomes. For example, it could be the case that children in households that were more exposed to criminal activity are also more likely to experience domestic violence. In this case it could be argued that the exposure to criminal activity early in life could be the cause of higher incarceration rates, and not the exposure to domestic violence. In this case the exposure to domestic violence would also be the result of exposure to criminal activity.

This page titled 3.5: Case Control Studies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.6: Unethical Experiments

There is a history of government entities performing unethical and harmful studies on its own citizens to avoid the difficulties encountered with observational studies. One of the most famous is the Tuskegee Syphilis Study, which was a study of medication for syphilis that was jointly administered by the United States Public Health Services and Tuskegee University (Reverby 2012). Amid growing concern over medication options and effectiveness for syphilis in the late stages of the disease, this study was set up to see what happened when men with this condition were left untreated and if there were differences in the disease by race. The participants were all African American men, of whom some had syphilis. Some subjects in the study who had the disease were left untreated to serve as a comparison group. Additionally, the sex partners of the men were not traced, and few were treated. Even after penicillin proved effective for syphilis, the men in the comparison group were left untreated. Many articles that appeared in the medical literature reporting on the results of the study referred to the men as volunteers.

After the public was made aware of the study in 1972, there was a federal investigation, Senate hearings, and a lawsuit. Eventually medical and health care for the survivors and their syphilis-positive wives and children were provided with compensation. Additionally, federal rules on informed consent and human subject protections were enacted. President Bill Clinton offered a formal apology in 1997. While the researchers involved with the study did not control the sexual habits of the men, nor did they infect any of the men with syphilis, they did control whether the men received treatment for the disease—so this study is not completely an observational study. That this study was conducted under these conditions is a stark example of systemic racism in the United States (Park 2017). Some authors have even connected the design of the experiment to the eugenics movement of the early twentieth century in the United States (Lombardo and Dorr 2006). The long-term damage of this experiment extends beyond those who were participants in the study. Efforts to aid many African American communities is met with distrust because of this experiment (Gamble 1997; McCallum et al. 2006; Corbie-Smith 1999).

Another astounding example of an outright violation of human rights in an empirical study was perpetrated by American psychiatrists, psychologists, and neurosurgeons throughout the second half of the twentieth century in a series of studies housed at leading medical schools with funding from the Central Intelligence Agency (CIA) and the U.S. military. These experiments involved the use of brain electrodes, LSD, hypnosis, the development of biological, chemical, and nonlethal weapons, the implantation of false memories, and creation of amnesia. Many of these experiments were conducted on individuals without documented consent (Ross 2007).

Lysergic acid diethylamide (LSD) was first synthesized in 1938, and its psychoactive properties were discovered in 1943. Such psychedelic substances were thought to have potential therapeutic uses (Carhart-Harris and Goodwin 2017). Shortly after these developments the CIA began sponsoring research into the use of LSD based on fears that it could be used as a biological weapon during the cold war with the potential of using the drug for mind control, “brainwashing,”, and in interrogation (Dyck 2008).

After World War II, the CIA recruited German scientists to help develop technologies for use against the Soviet Union in the cold war, including individuals who had been accused of human rights violations. One of these scientists brought LSD to the attention of the CIA in 1948, and its use was soon tested as an enhanced interrogation technique against Soviet spies. In 1953 the CIA project MKUltra was established using some of these same scientists. Public and private research institutions, including many universities, hospitals, and prisons, were provided with funding from the CIA under this program to conduct human experimental research with LSD on civilians, prisoners, and patients (Dyck 2008). While many of the records of this project were destroyed, some remained and were brought to light during hearings before the U.S. Senate in 1975 (Disbennett 2014).

It is notable that at the time of these experiments, the United States had enacted and agreed to the Nuremberg Code, which forbade such practices (Shuster 1997). During the same time as some of these experiments, the Declaration of Helsinki was adopted by the World Medical Association, which set out ethical principles for physician researchers to follow when conducting human medical trials (Carlson et al. 2004; Rickham 1964). It specified that the goal of producing new medical findings should never take precedence over the well-being of research participants and that researchers are responsible for the protection of participants (Rickham 1964).

Hence, these early research studies were found to have clearly disregarded ethical guidelines agreed on by the government of the United States. There is also evidence that these studies have a disturbing racial element to them. An examination of archival documents, public testimony, interviews, and oral histories from participants of the early LSD studies at the U.S. National Institute

of Mental Health's Addiction Research Center in Lexington, Kentucky and the National Institutes of Health Clinical Center in Bethesda, Maryland details the conditions under which the experiments took place. At the Addiction Research Center in particular, African American prisoners with substance abuse problems were routinely harmfully exploited in these LSD studies (Campbell and Stark 2017; Strauss et al. 2021).

A recent review found that participants in these studies were mainly recruited from either prisons or hospitals and that in each case individuals of color were over-represented when compared to the corresponding state demographics (Strauss et al. 2021). These studies often used high, frequent, and prolonged dosing without any informed consent. When race was reported, the incarcerated participants were usually African American. The severity of the treatment in the study was also sometimes based on race. In one study, African American participants were given more than double the dose of LSD compared with white participants. Moreover, white participants endured only eight days of LSD administration, less than one -tenth of the African American participants. Astoundingly, some participants were even given drugs like heroin in exchange for their participation in the study. Other studies exploited, dehumanized, and outright tortured female, African American, schizophrenic, and cognitively disabled subjects. Even with the incredible risks that the participants were exposed to, the design of many of the studies was so poor and inconsistent that no meaningful scientific conclusions could be reached (Strauss et al. 2021).

It should be noted that instances of unethical experimentation are not isolated to the past nor to medical experimentation. Modern examples of unethical studies include what are known as social pressure manipulations, which include political studies that explicitly have the intent to change outcomes, generate feelings of group conflict, or pursue activist and partisan goals (McDermott and Hatemi 2020). One study targeted approximately half of the registered African American voters in a southern state with a history of racial inequality. This study included subjecting some of this population to a racially charged message. The study resulted in a reduction in minority turnout in a real election (Nickerson and White 2013). These types of studies fail to inform the individuals in the population that they are part of a study, and hence the participants have not consented to participation.

Social media has provided a platform for low-cost, rapid, and targeted social experimentation. A recent emotional -contagion study considered how emotional states can be transferred between individuals, leading others to experience the same emotions without their awareness. This study was conducted on hundreds of thousands of people without any informed consent on a popular social media platform by researchers at Cornell University. The experiment manipulated almost 700,000 individuals who were exposed to emotional expressions on the social media platform. The researchers tracked how this manipulation led to changes in the posting behaviors of the participants. That is, the emotions of individuals were purposely manipulated to observe how they would respond. These researchers did not obtain any ethics approval for the study (Kramer et al. 2014; McDermott and Hatemi 2020).

Other examples include resource allocation studies where some subjects are given large allocations of financial resources so that their behavior can be compared with those not given such resources. In many cases the gains are short term, and no thought is given to how the subjects will respond when the additional resources ended or what secondary effects such arbitrary financial windfalls for some in a society may have when others are denied help. In other studies, people are subjected to violent stimuli to study societal polarization and sectarianism. There is often no follow- up on what the long-term consequences are to the subjects in the study or for society. Other examples include studies where researchers manipulate dating site algorithms to create intentional mismatches. The researchers then observing what happens to the corresponding paired individuals.

All these studies go against some of the most basic concepts of ethical research practices. First, study participants should be able to give informed consent before they are used for research purposes, and secondly, participants should not be harmed by participation in the study.

The studies considered in this section were from the United States. Of course, unethical studies have been, and continue to be performed worldwide.

This page titled 3.6: Unethical Experiments is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.7: Critical Analysis

When critically analyzing a study in terms of its use of experiments, the main analysis concerns whether the study is based on a designed experiment, where the researchers control the environment of the subjects, or an observational study, where the subjects are simply observed. Many studies, particularly those with a social or medial context, will have at least some part of them that are observational, though some factors may be controlled to some extent. The key is to consider what is being controlled, what is being observed, and what confounding issues may affect how the results can be interpreted.

The key questions in critically analyzing the use of an experiment in a study are:

1. Is the study a designed experiment, an observational study, or does it contain elements of both types of studies?
2. What are the potential confounding problems that the study may have?
3. Are there ethical questions about how the experiment was performed?

This page titled 3.7: Critical Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.7.1: Labor Market Discrimination

Consider the study introduced earlier (Section 2.5.1) which used virtual resumes to demonstrate that resumes designed to be perceived as coming from African American applicants received fewer callbacks from potential employers than resumes designed to be perceived as coming from white applicants (Bertrand and Mullainathan 2004). This study demonstrated bias in the way that employers evaluated resumes, even when the race of the applicant was not explicitly revealed to the evaluator.

This study was based on a designed experiment. The researchers controlled the perceived race conveyed by the names on the resumes, as opposed observing how employers reacted to real resumes that happen to have been sent to them. Conducting the study in this way had the advantage that the researchers were able to carefully control the level of the experience on the resumes, as well as other factors that could eliminate many confounding issues from the study. Observational studies of this same topic have problems with confounding in that the experience of the applicants, the appearance of the resume, and other potentially important factors are not controlled. In such cases it would be impossible to conclude that the behavior of the employers was due solely to the perceived race of the applicant, and not one of these other factors.

The authors of this study did a good job of pointing out possible remaining confounding problems. One potential problem was that employers may not be evaluating the race of the applicant based on the name but may really be evaluating the social background of the applicant. This assumes that the perceived African American names indicate individuals from disadvantaged backgrounds. In this case the employer would be discriminating based on the perceived social background of the applicants instead of race. The researchers performed a more nuanced analysis of the data and argue against this confounding problem. Another potential confounding issue may be that the perceived African American names are less well known and may appear “odd” to human resource managers. In this case the discrimination would be against odd sounding names. Again, the researchers pointed out that they used more common names and provided additional statistical analysis to support the idea that this confounding problem should not be an issue.

The authors pointed out some other potential problems with the study that are obliquely related to confounding problems. The first was that a callback from an employer does not necessarily mean that the applicant was being seriously considered for the job or that they would be hired once their race was revealed. Secondly, not all African American applicants have names that would be perceived as being African American, and hence once again the study did not reveal what would happen when these applicants' race was revealed at some point in the application process. Combining these issues, the study seemed to address whether these resumes would be rejected in their first evaluation based on the perceived race of the name.

The authors emphasized the techniques that they went through to eliminate as many possible confounding issues as possible. For example, the resumes were designed to be both believable and to vary qualifications as little as possible between the perceived African American and white named resumes. The main problem with this study, which the authors emphasized throughout, was that the perception of race, and the exhibition of a corresponding bias in the employment process is a very sophisticated system that is difficult to summarize with a simple experiment. Therefore, it was very difficult to pinpoint exactly why the differences in the callback rates was observed. Overall, this study certainly provided evidence of callback bias for potential employment of the basis of the perceived race of the applicant, though the issues involved may be more complicated than could be addressed by this study.

This page titled 3.7.1: Labor Market Discrimination is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.7.2: Visual Exercise Environments and Treadmill Running

Jogging and walking are examples of physical activities often undertaken by individuals who wish to increase their cardiovascular health. These activities can normally be done in many outdoor locations, but because of weather and safety issues individuals often opt to exercise inside using a stationary treadmill instead. Many individuals using a treadmill do so while watching television, streaming programs, or listening to music. Some research has shown that the type of visual and auditory stimulus used by these individuals can influence their performance during their exercise session. For example, research shows that listening to loud or fast music can encourage treadmill runners to run faster and longer (Bharani et al. 2004; Tennenbaum et al. 2004; Edworthy and Waring 2006). The environment that physical activity takes place in can also influence well-being. Immersion in the natural environment has been associated with lowered blood pressure, reduced perceived exertion, improved mood, and general enjoyment of the activity, even when viewed using static images (Duncan et al. 2014; Pretty et al. 2005, 2007; Akers et al. 2012; Wooler et al. 2015; Focht 2009; Plante et al. 2007; Jones et al. 2014).

One such study concerns the use of images or video of natural environments. Several theories have been proposed to explain why exposure to natural environments can enhance emotions. For example, it has been proposed that natural environments have a restorative effect on the ability of the brain to focus. Other theories suggest that we have an automatic unconscious response to nature (Kaplan and Kaplan 1989; Bratman et al. 2012; Johnson 1994). Of particular interest is the potential difference between a static environment, one for which the visual stimulus is constant and not changing, and a dynamic environment where the visual stimulus is constantly changing (Yeh et al. 2017).

The study was based on thirty individuals, 12 of whom were female and 18 of whom were male, who were recruited and gave informed consent to participate in the study. Twenty-four of the recruited individuals regularly exercised at least 150 minutes per week, while the remaining participants exercised irregularly or participated in light exercise routines. The research protocol using these human subjects was approved by the ethics committee at the university where the research was performed (Yeh et al. 2017).

The research team designed two visual nature images, one static and one dynamic. The dynamic image was based on a video filmed from the point of view of a cyclist proceeding through the paths of a local botanical garden. The static image consisted of one of the frames of the video. A third environment consisted of the personal preference of the participant, which included watching television or movies or viewing a static image. A treadmill was placed in this environment and the distance and duration of the exercise session was intentionally shielded from the participant.

Each participant exercised on the treadmill under each of the three conditions at a speed set by the individual. Participants were told to run at a comfortable speed. There was a seven-day gap between the test for each condition to avoid possible effects from previous sessions. Each session lasted twenty minutes. The heart rate was monitored for each participant along with the distance covered. Additionally, the Sport Emotion Questionnaire was used to measure the emotions of each participant five minutes before and immediately after each observed run. The Sport Emotion Questionnaire is a standard questionnaire used to measure happiness, anxiety, dejection, anger, and excitement and has been used effectively in other studies (Jones et al. 2005; Arnold and Fletcher 2014).

The results from the study showed that self-selected entertainment was associated with higher running distances when compared to the static image during the exercise period. There was no detectable difference in running distances between the dynamic image and the self-selected entertainment. Higher heart rates were achieved with self-selected entertainment than either of the dynamic or static images. Additionally, there was some indication that participants felt happier when using the static and dynamic images instead of the self-selected entertainment, while the remaining measures of anxiety, dejection, anger, and excitement were similar for all three groups.

This study is based on a designed experiment in that the researchers decided what stimuli that individuals would be subjected to, as well as the other conditions of the experiment. This study is a little different in that each participant was observed in each of the three groups. This is not uncommon for studies involving human subjects where there is usually a large amount of variation between the physical and mental attributes of the subjects. If the subjects were simply divided into three groups, then any observed differences between the groups would be confounded with any possible differences between the subjects themselves. Because each subject was observed with each treatment, the same group of subjects was observed, so differences between the groups are more likely due to the different treatments.

This page titled 3.7.2: Visual Exercise Environments and Treadmill Running is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.7.3: Same-Sex Marriage Laws and Hate Crimes

Consider the study introduced in Section 2.5.2 that considers the incidence of hate crimes against the LGBTQIA+ community when same-sex marriage laws were enacted in the United States. This research was based on an observational study as the authors of the study obviously did not control any of the factors that would influence the rate at which hate crimes were observed. The authors of the study pointed out several potential confounding problems. The main such problem was that societal views may be changing regardless of the implementation of same-sex marriage laws. For example, the authors pointed out that changes in population demographics, anti-hate crime legislation, and population migration may also influence the rate of hate crimes against the LGBTQIA+ community. Migration could possibly be in response to same-sex marriage legislation. The authors of the study attempted to address these issues using statistical techniques.

The impact of this study is difficult to assess as the social constructs associated with societal trends, acceptance, and violent crime are complicated. It is certainly believable that legislative or judicial progress in the marriage rights of the LGBTQIA+ community could have some positive influence on individuals who have moderate or slight bias against those in the LGBTQIA+ community. It may be more difficult to accept that those with very strong feelings, those who would act violently against another individual, would be changed significantly by these factors. Indeed, there is a secondary explanation that could be as plausible in that the court cases could coincide with or even follow changes in societal attitudes about the issue. This is partially evidenced by the fact that many states legalized same-sex marriage, and that this type of legislative effort usually reflects changing attitudes that are already present in society. Therefore, the legislation could be in response to changing social attitudes about the LGBTQIA+ community.

This page titled 3.7.3: Same-Sex Marriage Laws and Hate Crimes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.7.4: Effects of a New Special Formula Fertilizer on Potato Growth

The potato is a highly adaptable staple food that is rich in nutrients and is the fourth largest crop grown globally (Xu et al. 2025). In some areas of the world such as China, potatoes are strategically important for economic development and food security (Naqqash et al. 2016; Ali et al. 2021; Devaux et al. 2014). Because of the importance of this crop, researchers are always looking for effective methods to increase yields and enhance plant characteristics. A recent study considered the use of a new special fertilizer on the growth and yield of potato crops in the region of Xinjiang located in northwest China, a region that suffers from low potato crop yield.

The experiment was conducted at the Baicheng Agricultural Experiment Station, which has a temperate climate characterized by cold winters and cool summers. Before the experiment, soil samples were collected from the field used in the experiment. The samples were analyzed for their chemical and physical properties and were used to develop three of the fertilizer treatments used in the study. The experiment was conducted using a potato variety called Xisen 6 with five treatments: a control group where no fertilizer was applied; a group using a conventional fertilizer, and three more groups each using specific fertilizers formulated by the researchers. The field where the experiment took place was divided into 15 plots. The control group and each of the five treatments were randomly assigned three plots each.

During four of the stages of potato growth, three randomly selected plants from each treatment plot were measured for plant height and stem diameter, with the average values for each plot representing the overall growth of the plants in that plot. Other important agronomy measurements were taken during the appropriate stages of growth. These measurements include relative chlorophyll content, the above-ground dry matter accumulation, and photosynthetic data. At the end of the growing season all potatoes in each plot were harvested and weighed to determine the total yield, which was then converted into standardized yield. The number of marketable potatoes and marketable weight per plant was also recorded. Finally, the marketable rate and economic benefit were calculated. Additionally, the total sugar content and amount of Vitamin C was measured.

The study provided evidence that one of the fertilizers formulated by the researchers improved the agronomic traits and photosynthetic performance of the potato crop. Specifically, the researchers observed an increase in both yield and economic efficiency, suggesting that the fertilizer formula has significant potential for improving fertilizer efficiency in potato production in arid regions. However, the researchers caution that the large-scale promotion of the fertilizer should be based on local agronomic practices.

This study was based on a designed experiment where potato plots were randomly assigned to one of the treatments or the control. The randomization scheme used for assigning the plots to the treatments is a very important part of the design of this study. Randomization is used to reduce issues with confounding. Even though the researchers may have attempted to make the plots as uniform as possible, some plots could have been more fertile or have slightly better growing conditions than some of the other plots. This discrepancy created a serious problem by confounding the plot and the treatments. By randomly selecting which plots got which treatments, the researchers broke up this confounding problem by ensuring that the treatment selection for each plot was not in any way related to how well the potato plants might naturally grow in the plot. As we shall see later in the book, this type of randomization scheme is very common in experiments of this type.

This page titled 3.7.4: Effects of a New Special Formula Fertilizer on Potato Growth is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.7.5: The Supreme Court and Personal Attitudes

Consider the study of the effect of a supreme court ruling about same-sex marriage and personal attitudes from Section 2.5.4. In that study participants used an article about the upcoming supreme court ruling on same-sex marriage. The researchers randomly assigned each participant to either read an article that indicated that the supreme court was likely to rule in favor of same-sex marriage or an article that indicated a ruling against same-sex marriage. Because the researchers decided who would read each article, this study was based on a designed experiment.

Because this study was based on a controlled experiment, fewer confounding factors were immediately obvious. The main weakness of this experiment was that it was based on a hypothetical situation that may or may not correspond to reality. Indeed, one suggested confounding factor in this case was the potential effect of how the participants obviously knew that they were part of an experiment. One can easily imagine that individuals who hold viewpoints that are perceived as unpopular or controversial may feel the need to temper their responses in an experimental situation such as this, particularly when expressing internal perceptions. That is, participants who do not approve of same-sex marriage may be more likely to state that they would support it to some extent to appear less controversial. The authors of the study addressed this problem by pointing out that participants appeared to express both positive and negative reactions to each of the potential rulings, but that this confounding effect could not be ruled out (Tankard and Paluck 2017).

Another related potential problem is that this type of experiment may cause participants to personally understand the issue to an extent that they otherwise might not (Egan and Citrin 2011; Unger 2008; Tankard and Paluck 2017). Therefore, the authors of the study pointed out that the results from the experiment only approximated what might happen with a real-world event.

This page titled 3.7.5: The Supreme Court and Personal Attitudes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.7.6: Sleep and College Students

Sleep deprivation is becoming a major public health challenge. The Centers for Disease Control and Prevention reported that more than one third of Americans sleep less than the recommended amount (Liu 2016). Sleep deprivation has been linked to chronic health problems as well as decreases in cognitive skills, decision making, productivity, and physical performance (Gibson and Shrader 2018; Heissel and Norris 2018; Giuntella and Mazzonna 2019).

One study considered the role of economic incentives to promote healthy sleep habits. The study was conducted using observations on students from the University of Oxford and the University of Pittsburgh (Avery et al. 2025). Each student was given a wearable device that collected data on their sleep habits. The experiment was conducted first over an eight-week period at Oxford, and then over a fifteen-week period at the University of Pittsburgh. These time lengths were chosen because they corresponded to the term lengths for each school. For the first two weeks, baseline data was collected from the wearable devices for everyone in the study. This baseline data allowed the researchers to establish the usual behavior of the individuals before any incentives are introduced. With this baseline data researchers were able to determine if the incentives were associated with a change in the behavior of the individuals. Surveys and incentives started in the third week of the study, and the wearable devices were monitored until the end of the study period.

Individuals in the treatment group of the study received financial incentives for meeting their sleep targets. Each night they achieved their targets they were rewarded by a nominal amount ($3.75 in U.S. dollars or £2.5 Great Britian pounds). Similarly, each night that the participant failed to meet their targets, they were penalized by the same amount. The study also provided incentives for continued participation. Individuals in the control group did not receive any financial incentives.

A total of 508 individuals provided useful data for the study. The researchers reported that some of the data provided by the participants were not usable either because they found the wearable device uncomfortable or they dropped out of the study for other reasons. Importantly, the researchers investigated whether the group of individuals who did not provide usable data for the study were different in some key demographic way from those who did provide useful data. The researchers found no differences that they felt would introduce bias in the overall results of the study.

Some parts of this study did not have a control group, and some of the analysis pooled all the data regardless of whether there was a control group. For the discussion here, only the results based on comparing the control group to the treatment group will be used. For these data, it was found that subjects receiving incentives were more likely to sleep the recommended number of hours. Using statistical techniques, the researchers were able to show that these results hold even when accounting for differences between individuals.

An interesting question that the researchers considered was whether the incentives produced an effect only while they were in effect, or if the incentives produced a lasting effect. The researchers followed the subjects for a few weeks after the incentives stopped, and the data from this part of the study showed that the effects on sleep persisted to some extent in the weeks following the termination of treatment.

This study was based on a designed experiment, but it is somewhat more complicated than some of the other experiments we have encountered in that there is a time element, and the wider application of the study did not have a control group. As the authors of the study pointed out, the time element was crucially important because an effect may last for a short time. If the effect is short-term, then the conclusion is that the economic incentives did not provide lasting change and the effect may disappear as soon as, or slightly after, the incentives stop. Another possible conclusion was that when the subjects experience this lifestyle change, it leaves a positive lasting effect on their habits. Unfortunately, without a follow-up study of these same subjects, it is impossible to know if these individuals adopted better sleep habits on any long-term scale. Note that the subjects obviously knew they were taking part in a sleep study and may have kept good sleeping habits as long as the data were being collected, knowing that someone would be watching their progress even after the economic incentives ended. Many of these individuals may have kept up healthier sleep habits in an unconscious effort to please the researchers.

This page titled 3.7.6: Sleep and College Students is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.7.7: Gender and COVID-19 Outcomes

In the early stages of the COVID-19 pandemic, researchers noted a pattern of much higher mortality among men when compared to women (Grasselli et al. 2020; Jin et al. 2020). This trend led researchers to hypothesize the existence of biological sex-related factors in susceptibility to the virus. These results suggested the possibility of sex-specific clinical and public health interventions (Gebhard 2020; Peckham et al. 2020; Pivonello et al. 2021; Scully et al. 2020; Bunders and Altfeld 2020; Hussain et al. 2020; Klein et al. 2020; Takahashi et al. 2020; Wolfe et al. 2021).

However, a later study argued that the variation that was attributed to sex observed in COVID-19 data is rather likely explained by variations in data collection practices, gender- (not sex) based patterns in health behaviors, occupational exposures, and health conditions. Further, the gender-based effect may vary according to other issues such as socioeconomic status, geography, and race or ethnicity (Danielsen et al. 2022). This is of interest to researchers because the shift of the focus from sex to gender implies that there may not necessarily be biological factors involved in the difference.

The study analyzed weekly data from April 27, 2020, through the week of May 10, 2021, of sex categorized COVID-19 cases and mortality for the United States. The data were collected from state public health websites. The data suggested moderately higher COVID-19 mortality rates for men. It should first be noted that in the United States men had higher mortality rates than women before the pandemic. Interestingly, the data provided little evidence that COVID-19 changed this dynamic. The authors of the study argued that this could indicate that the disparity in COVID-19 outcomes between men and women did not have to do with a specific, sex-linked male vulnerability to COVID-19, but reflects a set of complex biosocial factors as well as the systemic health and social inequities of men and women in different socioeconomic groups.

The conclusions of this study were based on an observational study as the researchers did not control any of the factors associated with sex or COVID-19. The authors of the study pointed out that the problem was essentially confounding because of the unequal rates of preexisting health conditions between women and men across various social groups. For example, heart conditions are associated with poorer COVID-19 outcomes and mortality (Bae et al. 2021; Madjid et al. 2020). At the same time the prevalence is not evenly distributed across age and sex, and in fact older men tend to have higher rates of cardiovascular disease (James et al. 2018; Mosca et al. 2011). This means that the higher COVID-19 mortality rates may reflect the fact that these preexisting health conditions vary across sex, and the high prevalence of the preexisting health conditions for men drove up the mortality rate from COVID-19. In the observed data, this effect could not be distinguished from a sex-specific biological effect on COVID-19 mortality.

This page titled 3.7.7: Gender and COVID-19 Outcomes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.7.8: Thought and the Growth of Fungi Cultures

Since the spiritual movement of the mid-nineteenth century, there has been a substantial amount of public interest in the study of alleged psychic phenomena, including extrasensory perception, telepathy, teleportation, precognition, clairvoyance, and psychokinesis. Along with this interest many intellectuals performed investigations to study these phenomena using the scientific method. A study conducted in 1968 considered the question of whether individuals could use thought processes to retard the growth of fungi (Barry 1968).

In a laboratory the fungus Rhizoctonia solani was cultivated in petri dishes with an inoculation made in the center of the dishes the evening before the experimental sessions. The dishes were then placed in a warm incubator with intermittent lighting 12 hours out of the following 24. The experimental conditions were closely controlled in that all possible precautions were taken to ensure that all the petri dishes were subjected to the same conditions so that the only varying factor that could have influenced the growth of the fungus was the psychokinetic one. The authors of the study report that the purity of the fungus, the composition of the culture, the relative humidity, the temperature, and lighting were carefully regulated.

The experimental sessions took place in a room adjoining the incubator where each subject was assigned five experimental dishes and five control dishes. The subjects were asked to exert their metal ability on their experimental dishes, trying to retard the growth of the fungus in them while disregarding the controls. The control dishes were kept in the room for the same length of time so that they were subjected to the same temperature conditions. Each subject was kept approximately 1.5 meters from the dishes. Each session lasted fifteen minutes.

The experimental results were obtained by outlining, on a sheet of thin paper, the boundary of the fungi growth by someone who did not know the aim of the experiment and who did not know which dishes were the control dishes. The outlines were cut out and weighed under conditions of constant temperature and humidity. If the experimental total for the treatment group was less than the control, the session was counted as a success; otherwise, the session was counted as a failure.

Out of 39 trials, there were 33 successes, 3 failures, and 3 where there was no difference between the control and treatment groups. These results can be assessed in terms of what would be expected if there was no effect on the treatment group, where one would expect half of the trials to be successful and half of the trials to be failures. The researchers were able to use statistical calculations to conclude that if there were no effect, then one would observe the results found in the experiment about 1 in 1,000 times. The results led the investigators to conclude that something else must have caused the retardation in fungus growth, namely the treatment condition.

This controlled experiment used a treatment and a control group. According to the details of the experimental design as reported by in the research article, the researchers took every precaution to attempt to treat the fungus cultures used for the treatment group the same as those used for the control group. This control ensured as few confounding factors as possible, and any observed differences should only be attributable to the treatment, which in this case was the thought energy of the participants in the study. The results as reported in the paper are more detailed than are reported here, and it seems that many of the positive results were attributed to a few subjects used in the study, which could indicate that these individuals had a particular gift in this area.

What is interesting is the way the data were treated. Notice that instead of comparing the amount of reduction in the growth of the fungus, any reduction was treated as a success. For most skeptics of this type of research the amount of growth reduction would be very relevant information in determining the strength of the conclusions from the research project. Certainly, when a researcher manipulates the observed data in such a way that not all the information from the experiment is used, it is suggestive of what a more in-depth analysis of the data would reveal. Hence, this example is a research study where the design of the experiment was quite carefully thought out, but the way the results were used creates questions about the results.

This page titled 3.7.8: Thought and the Growth of Fungi Cultures is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.7.9: Risk Factors for Homelessness

Homelessness is a chronic and growing problem in the United States that can be partially attributed to labor market changes, inadequate supply of low-cost housing, and cuts in income assistance programs (Burt and Cohen 1989; Rossi 1991). Other individual risk factors such as race and ethnicity, gender, age, and substance abuse have also been cited as contributing factors (Susser et al. 1993; Shelton et al. 2009). A study from 1997 uses previous research on adolescent factors for negative psychiatric outcomes in adulthood to link these factors to homelessness (Herman et al. 1997). The hypothesis was also based on previous observations that these factors are associated with adult homelessness, and the study sought to strengthen these conclusions (Susser et al. 1993; Koegel 2011).

This study analyzed data from an earlier study based on observations from 1,507 adults. Those observed in the first study were contacted and observed again. In the initial observations, 169 individuals reported that they had been homeless at some point in their adult lives. A second group of observations from the first study was constructed of individuals who had not reported being homeless but were comparable to the group that reported homelessness in terms of other factors often associated with homelessness. Both groups were asked to report their assessments of the quality of their care during childhood in terms of factors such as lack of parental care, physical abuse, and sexual abuse.

After carefully controlling for these additional factors using statistical techniques, the authors of the study found that adverse events in adolescence were strongly associated with periods of homelessness as adults. This is an observational study based on a case-control design. The authors were interested in the effect of factors that occurred in adolescence and their effect on homelessness. A standard approach would consider two groups of adult subjects: those who had the adolescent factors and those who did not. Periods of adult homelessness would then be observed and compared for each group. However, this study grouped adults by whether they have experienced periods of homelessness and looked at the adolescent factors in each group, resulting in the case-control design.

This page titled 3.7.9: Risk Factors for Homelessness is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.8: Summary

The observational portion of the scientific method can be classified by how much control the researcher has over the conditions of the study. In designed experiments researchers have almost absolute control over the conditions of the experiment. These experiments can provide the most useful results, though they are often impractical in practice due to ethical or economic issues. In other cases, experimental conditions are impossible to control in practice. Experiments where researchers observe experimental conditions and observed responses simultaneously, and the conditions of the experiment are not controlled, are called observational studies. These types of experiments have the added difficulty in that confounding can be a problem, where the potential effect of the experimental conditions and the response cannot be distinguished from one another. These problems can seriously weaken the impact of the conclusions from the experiment.

Unethical experimentation has occurred in the United States, and the personal and social impacts of these experiments have been devastating.

This page titled 3.8: Summary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.9: About the Quote



Figure **3.9.1**: Richard Feynman was an American physicist who received the Nobel Prize in Physics in 1965. (Public Domain)

Richard Feynman (1918–1988) was an American physicist who received the Nobel Prize in Physics in 1965 (Figure **3.9.1**). Feynman primarily worked with quantum mechanics, which studies the interaction of subatomic particles. He worked on the development of the atomic bomb during World War II and became well known as a member of the Rogers Commission, the panel that investigated the Space Shuttle Challenger disaster. He is also credited with pioneering the field of quantum computing and nanotechnology. Feynman is also known for his many popular books on physics and science. The quotation at the beginning of the chapter is taken from a video lecture on the scientific method recorded in 1965.

This page titled 3.9: About the Quote is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.10: Practice and Exploration

1. A professor is interested in knowing whether the use of a free tutoring center in a department is helpful with student grades. At the end of the semester the professor gives the students in their class a short survey, and one of the questions asks whether the students used the tutoring center or not. It turns out that the grades for the students who used the tutoring center are higher, on average, than the grades of the students who did not use the center. Explain why this is an observational study and point out some possible confounding problems that could affect how the results of the study are interpreted. Identify the response, the factors studied, and the steps in the scientific method used to make conclusions about the issue. Is there a treatment group and a control group? What are they identified as? Are there other explanations as to why the average grades might be different?
2. Professor Plum is interested in whether the color of the paper used during an exam influences student performance. At the midterm exam in their class, the students who were seated in the first ten rows were given the midterm printed on yellow paper, while the students who were seated in the last ten rows where given the same midterm printed on green paper. The results from the exam showed that the students who took the exam on yellow paper had higher grades on average than the students who took the exam on green paper. Explain why this is or is not an observational study. Whether it is observational or not, suggest some confounding problems that this study might have. Identify the response, the factors studied, and the steps in the scientific method used to make conclusions about the issue. Is there a treatment group and a control group? What are they identified as? Are there ethical implications associated with this study? If so, how could the ethical implications be addressed?
3. One study considered demographic disparities in the nutritional quality of packaged food purchases in the United States between 2008 and 2018. This study used data based on the Nielsen *U.S. Homescan Consumer Panel* for which households self-report demographic measures and are instructed to scan barcodes on all purchases. The researchers then used the bar codes to match the purchases with specific food items, from which the nutritional value of the purchases could be determined. The study found that nutritional disparities existed between different income and education levels with the largest disparities for education level. Is this an observational study or a designed experiment? Identify the response, the factors studied, and the steps in the scientific method used to make conclusions about the issue. Is there a treatment group and a control group? What are they identified as? Disparities were also found between African American and white households, while no disparities were found between white and Hispanic households. Identify at least three possible confounding effects for this study. For each confounding effect, explain why you think it would occur and if it would be possible to control for the effect with a study designed in a different way.
4. Demographics of high school students have been changing rapidly while the demographics of high school teachers have remained relatively steady. A great deal of research has demonstrated that racial parity between students and teachers increases student performance and success. While it is estimated that by 2029 students of color will make up 60% of the student population, the teacher population has been predominately white. One research article examined how racial parity between students and teachers affects class attendance (Gottfried et al. 2022). In this study the data collection methodology states:

   > *This study relies on administrative high school data from a high school district in California to examine the influence of a student–teacher racial and ethnic match on absenteeism. The data was obtained directly from the district's central office.*

   Is this an observational study or a designed experiment? Identify the response and the factors that are being studied and the steps in the scientific method used to make conclusions about the issue. Is there a treatment group and a control group? What are they identified as? The authors of the study also state that the students' and teachers' names were coded. Why do you think this was done?
5. Consider a situation where student retention has become an important issue at a large Midwestern university. The problem of interest is that a sizable percentage of those who enroll as first-year students do not complete a degree at the university within seven years. An on-campus researcher hypothesizes that student cultural experiences during their first year enables students to become more connected with the local university community, and that those students would be less likely to leave. Detail how a designed experiment could be used to study this hypothesis. What would be some major problems with implementing this type of study? Would there be possible ethical issues? Detail how this hypothesis could be studied using a traditional observational

study and alternatively a case-control study. What are the potential problems with each type of approach? What are some confounding factors that could present a problem?

6. A political scientist is interested in how political affiliates may differ in their response to scientific news reports. To study this idea, they compose a fake scientific report that suggests a common treatment for a contagious disease is just a sugar pill that is used to bolster the profits of large pharmaceutical companies and that the disease normally goes away on its own. The report is then released to several bloggers of different political affiliations, and the response is observed. Is this an observational study or a designed experiment? What are the ethical implications of this study?
7. Using an online news service, find a news article about an issue that is important to you. Does the article use any empirical analysis to convey information about the issue? What was observed, and how were the observations interpreted? Was the empirical investigation based on a designed experiment or an observational study? Identify the response, the factors studied, and the steps in the scientific method used to make conclusions about the issue. Is there a treatment group and a control group? What are they identified as? What possible confounding issues are contained in the empirical analysis? Are there any ethical implications connected with the study? Do you trust the empirical analysis and the conclusions of the article? Why or why not?
8. Using your library's database research tool, find an academic research article about an issue that is important to you. Does the article use any empirical analysis to convey information about the issue? What was observed, and how were the observations interpreted? Was the empirical investigation based on a designed experiment or an observational study? Identify the response, the factors being studied, and the steps in the scientific method used to make conclusions about the issue. Is there a treatment group and a control group? What are they identified as? What possible confounding issues are contained in the empirical analysis? Are there any ethical implications connected with the study? Do you trust the empirical analysis and the conclusions of the article? Why or why not?

This page titled 3.10: Practice and Exploration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.11: Collaborative Activity

In this collaborative activity your group will run a designed experiment that will investigate how the method used for rolling a die possibly affects the outcomes of the die rolls. The purpose of this activity is to participate in an experiment and experience first-hand the possible problems associated with experimentation and data observation.

For this collaborative activity you will need the following materials:

- A regular six-sided die.
- A flat surface on which the die can land during the experiment.
- A hardcover book that can be used as a short ramp.
- A cup suitable for shaking the die.

Before starting the activity, you will need to assign members of your group to the following roles:

- A data manager who will oversee collecting the data from the experiment.
- A person in charge of rolling the die.

The roles can be shared when there are additional members in the group.

The experiment:

- The first part of the experiment will consider rolling the die using the method most often employed by game players: holding the die in the palm of one's hand and rolling it on a table. Be sure the table is clear, so the die does not encounter any obstructions. If the die rolls off of the table, do not use that observation and roll again. Using this method the die roller should roll the die 50–100 times (depending on how much time is allotted for the activity). The data manager should keep track of each of the rolls using the data sheet shown below.

Roll	Palm	Book	Cup	Roll	Palm	Book	Cup
1				51			
2				52			
3				53			
4				54			
5				55			
6				56			
7				57			
8				58			
9				59			
10				60			
11				61			
12				62			
13				63			
14				64			
15				65			
16				66			
17				67			
18				68			
19				69			
20				70			
21				71			
22				72			
23				73			
24				74			
25				75			
26				76			
27				77			
28				78			
29				79			
30				80			
31				81			
32				82			
33				83			

Roll	Palm	Book	Cup	Roll	Palm	Book	Cup
34				84			
35				85			
36				86			
37				87			
38				88			
39				89			
40				90			
41				91			
42				92			
43				93			
44				94			
45				95			
46				96			
47				97			
48				98			
49				99			
50				100			

- The next part of the experiment will consider what happens when the die is rolled in a very deliberate way. In this case a hard cover book should be placed on a table. The book should be larger than a usual paperback novel. With the book closed, the die should be place on the top cover near the edge opposite of the spine with the face with a single dot as the top face. Next, slowly open the book cover until the die slides down the cover and rolls on the table. Using this method the die roller should roll the die 40–80 times (depending on how much time is allotted for the activity). The data manager should keep track of each of the rolls using the data table.
- The final part of the experiment uses the cup to roll the die. For each roll, the die should be placed into the cup and shaken before rolling the die out onto the table. Once again, the die roller should roll the die 40–80 times with the data manager keeping track of the rolls on the data table.
- The table shown below provides an example of how the data table should look at the end of the experiment. For the example, we used 25 rolls of the die using each method. For the first method, the first roll was 5, following by a roll of 1, and then a roll of 4. The data was recorded in the first three rows of the column labeled “Palm”. The remaining numbers in the column correspond to the remaining observations from the example experiment using the “Palm” method of rolling the die. Similarly, the first three observed rolls for the “Book” method were 6, 5, and 3, and the first three observed rolls for the “Cup” method were 6, 3, and 5, with the remaining observations for each method continuing down the corresponding columns.

Roll	Palm	Book	Cup
1	5	6	6
2	1	5	3
3	4	3	5
4	3	1	1
5	4	4	5
6	6	2	3
7	5	5	3
8	4	2	4
9	6	5	3
10	4	2	4
11	6	4	6
12	3	1	2
13	4	4	6
14	3	2	3
15	3	2	4
16	3	4	1
17	6	4	1
18	1	5	2
19	2	4	4
20	1	4	6
21	2	5	6
22	3	2	6
23	2	4	6
24	4	2	3
25	3	6	2

Calculations:

- Once the observations have been collected, construct a table like the one shown in below. The calculations contained in this table will allow us to investigate how the three different methods of rolling the die affect the observed outcomes of the rolls.

	Outcomes					
Method	1	2	3	4	5	6
Palm						
Book						
Cup						

- For the first calculation, count how many times each of the die faces is observed for the first method of rolling the die, and report these counts in the row labeled “Palm".
- Next, count how many times each of the die faces is observed for the second method of rolling the die, and report these counts in the row labeled “Book”.
- Finally, count how many times each of the die faces is observed for the third method of rolling the die, and report these counts in the row labeled “Cup”.
- The table shown below contains the results of these calculations that have been done for the example data shown earlier. In the first column of the row labeled “Palm,” we have entered a 3 for the example data. If you look down the “Palm” column of the table below, you will see that we observed a roll of 1 three times. Similarly, there were three times when a roll of 2 was observed in that column, and seven times that a roll of 3 was observed in that column. The remaining entries of the table are computed in a similar manner.

	Outcomes					
Method	1	2	3	4	5	6
Palm	3	3	7	6	2	4
Book	2	7	1	8	5	2
Cup	3	3	6	4	2	7

Questions:

1. Suppose the method of rolling the die has absolutely no effect on the behavior of the die in terms of how often each of the faces is observed. What do you think the results observed in the table showing the number of oucomes in each category would look like? Would they always be the same? How do the results that you observed in your experiment compare to what would be expected?
2. Regardless of the data that was observed during your experiment, do you believe that the rolling method influences the behavior of the die? What is the reason for these beliefs? How do your personal feelings about the methods compare to what you observed? Are you surprised by the results?
3. Were there any practical problems that you encountered while doing the experiment? What were they? What types of changes could be made in the experiment to make it run easier?
4. Can you think of anything about the experiment that might produce bias in the results? Are there any possible confounding factors that could be problematic when interpreting the results of the experiment? How might these problems be addressed?

This page titled 3.11: Collaborative Activity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

3.12: References

Akers, A., J. Barton, R. Cossey, P. Gainsford, M. Griffin, D. Micklewright. 2012. “Visual Color Perception in Green Exercise: Positive Effects on Mood and Perceived Exertion.” *Environmental Science and Technology* 46, 8861–8866. https://doi.org/10.1021/es301685g

Ali, M. M. E., S. A. Petropoulus, D. A. F. H. Selim, M. Elbagory, M. M. Othman, A. E.-D. Omara, M. H. Mohamed. 2021. “Plant Growth, Yield, and Quality of Potato Crop in Relation to Potassium Fertilization.” *Agronomy* 11, 675. https://doi.org/10.3390/agronomy11040675

Arnold, R., and D. Fletcher. 2015. “Confirmatory Factor Analysis of the Sport Emotion Questionnaire in Organisational Environments.” *Journal of Sports Sciences* 33, no. 2, 169–179. https://doi.org/10.1080/02640414.2014.955520

Avery, Mallory, Osea Giuntella, and Peiran Jiao. 2025. "Why Don’t We Sleep Enough? A Field Experiment Among College Students." *Review of Economics and Statistics* 107, no. 1, 65-77. https://doi.org/10.1162/rest_a_01242

Bae, Sung A., So Ree Kim, Mi-Na Kim, Wan Joo Shim, and Seong-Mi Park. 2021. “Impact of Cardiovascular Disease and Risk Factors on Fatal Outcomes in Patients with COVID-19 According to Age: A Systematic Review and Meta-Analysis.” *Heart* 107, no. 5, 373–380. https://doi.org/10.1136/heartjnl-2020-317901

Barry, J. 1968. “General and Comparative Study of the Psychokinetic Effect on a Fungus Culture.” *Journal of Parapsychology* 32, 237–243. https://www.proquest.com/openview/e9...ar&cbl=1818062

Benson, Michael L., and Greer L. Fox. 2002. *Economic Distress, Community Context and Intimate Violence: An Application and Extension of Social Disorganization Theory, Final Report*. U.S. Department of Justice, National Institute of Justice. https://www.ojp.gov/pdffiles1/nij/grants/193434.pdf

Benson, Michael L., Greer L. Fox, Alfred DeMaris, and Judy Van Wyk. 2003. “Neighborhood Disadvantage, Individual Economic Distress and Violence Against Women in Intimate Relationships.” *Journal of Quantitative Criminology* 19, no. 3, 207–235. https://doi.org/10.1023/A:1024930208331

Benson, Michael L., John Wooldredge, Amy B. Thistlethwaite, and Greer Litton Fox. 2004. “The Correlation Between Race and Domestic Violence Is Confounded with Community Context.” *Social Problems* 51, no. 3, 326–342. https://doi.org/10.1525/sp.2004.51.3.326

Bertrand, Marianne, and Sendhil Mullainathan. 2004. “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review* 94, no. 4, 991–1013. https://doi.org/10.1257/0002828042002561

Bharani, A., A. Sahu, and V. Mathew. 2004. “Effect of Passive Distraction on Treadmill Exercise Test Performance in Healthy Males Using Music.” *International Journal of Cardiology* 97, no. 2, 305–306. https://doi.org/10.1016/j.ijcard.2003.05.048

Bratman, G. N., J. P. Hamilton, and G. C. Daily. 2012. “The Impacts of Nature Experience on Human Cognitive Function and Mental Health.” *Annals of the New York Academy of Sciences* 1249, no. 1, 118–136. https://doi.org/10.1111/j.1749-6632.2011.06400.x

Bunders, Madeleine J., and Marcus Altfeld. 2020. “Implications of Sex Differences in Immunity For SARS-COV-2 Pathogenesis and Design of Therapeutic Interventions.” *Immunity* 53, no. 3, 487–495. https://doi.org/10.1016/j.immuni.2020.08.003

Burt, Martha R., and Barbara E. Cohen. 1989. *America's Homeless: Numbers, Characteristics, and Programs that Serve Them*. Urban Institute Press.

Campbell, Nancy D., and Laura Stark. 2017. “Making Up ‘Vulnerable’ People: Human Subjects and the Subjective Experience of Medical Experiment.” In *Towards a Revival of Analytical Philosophy of History*. Krzysztof Brzechczyn, Editor. 225–253. Leiden, Netherlands: Brill. https://doi.org/10.1093/shm/hkv031

Carhart-Harris, Robin L., and Guy M. Goodwin. 2017. “The Therapeutic Potential of Psychedelic Drugs: Past, Present, and Future.” *Neuropsychopharmacology* 42, no. 11, 2105–2113. https://doi.org/10.1038/npp.2017.84

Carlson, Robert V., Kenneth M. Boyd, and David J. Webb. 2004. “The Revision of the Declaration of Helsinki: Past, Present and Future.” *British Journal of Clinical Pharmacology* 57, no. 6, 695–713. https://doi.org/10.1111/j.1365-2125.2004.02103.x

Condron, Dennis J., Daniel Tope, Christina R. Steidl, and Kendralin J. Freeman. 2013. “Racial Segregation and the Black/White Achievement Gap, 1992 to 2009.” *The Sociological Quarterly* 54, no. 1, 130–157. https://doi.org/10.1111/tsq.12010

Corbie-Smith, Giselle. 1999. “The Continuing Legacy of the Tuskegee Syphilis Study: Considerations for Clinical Investigation.” *The American Journal of the Medical Sciences* 317, no. 1, 5–8. https://doi.org/10.1016/S0002-9629(15)40464-1

Coronary Drug Project Research Group. 1973. “The Coronary Drug Project: Design, Methods, and Baseline Results”. *Circulation* 47, no. 3s1, I–1. https://doi.org/10.1161/01.CIR.47.3S1.I-1

Coronary Drug Project Research Group. 1980. “Influence of Adherence to Treatment and Response of Cholesterol on Mortality in the Coronary Drug Project.” *New England Journal of Medicine* 303, no. 18, 1038–1041. https://doi.org/10.1056/NEJM198010303031804

Danielsen, Ann Caroline, Katharine M. Lee, Marion Boulicault, Tamara Rushovich, Annika Gompers, Amelia Tarrant, Meredith Reiches, Heather Shattuck-Heidorn, Luke Weisman Miratrix, and Sarah S. Richardson. 2022. “Sex Disparities in COVID-19 Outcomes in the United States: Quantifying and Contextualizing Variation.” *Social Science and Medicine* 294, 114716. https://doi.org/10.1016/j.socscimed.2022.114716

de Groot, Patricia M., Carol C. Wu, Brett W. Carter, and Reginald F. Munden. 2018. “The Epidemiology of Lung Cancer.” *Translational Lung Cancer Research* 7, no. 3, 220. https://doi.org/10.21037/tlcr.2018.05.06

Devaux, A., P. Kromann, and O. Ortiz. 2014. “Potatoes for Sustainable Global Food Security.” *Potato Research* 57, 185–199. https://doi.org/10.1007/s11540-014-9265-1

Disbennett, B. M. 2014. “An Analysis of CIA and Military Testing of LSD on Non-Consenting US Service Members and Recovery Through the VA Disability System.” *Tennessee Journal of Race, Gender, and Social Justice* 3, no. 2, 173–202. https://doi.org/10.70658/2693-3225.1058

Doll, Richard. 1998. “Uncovering the Effects of Smoking: Historical Perspective.” *Statistical Methods in Medical Research* 7, no. 2, 87–117. https://doi.org/10.1177/096228029800700202

Duncan, M. J., N. D. Clark, S. L. Birch, J. Tallis, J. Hankey, E. Bryant, and E. L. Eyre. 2014. “The Effect of Green Exercise on Blood Pressure, Heart Rate, and Mood State in Primary School Children.” *International Journal of Environmental Research and Public Health* 11, 3678–3688. https://doi.org/10.3390/ijerph110403678

Dyck, Erika. 2008. *Psychedelic Psychiatry: LSD from Clinic to Campus*. John Hopkins University Press. https://www.press.jhu.edu/books/titl...u2LQ20GEVx-pFj

Edworthy, J., and H. Waring. 2006. “The Effects of Music Tempo and Loudness Level on Treadmill Exercise.” *Ergonomics* 49, no. 15, 1597–1610. https://doi.org/10.1080/00140130600899104

Egan, Patrick J., and Jack Citrin. 2011. “The Limits of Judicial Persuasion and the Fragility of Judicial Legitimacy.” Institute of Governmental Studies. Retrieved from https://escholarship.org/uc/item/2gh262w3.

Focht, B. C. 2009. “Brief Walks in Outdoor and Laboratory Environments: Effects on Affective Responses, Enjoyment, and Intentions to Walk for Exercise.” *Research Quarterly for Exercise and Sport* 80 no. 3, 611–620. https://doi.org/10.1080/02701367.2009.10599600

Freedman, David, Robert Pisani, and Roger Purves. 2007. *Statistics*. Norton Publishing. https://wwnorton.com/books/Statistic...roduct-details

Gamble, Vanessa Northington. 1997. “Under the Shadow of Tuskegee: African Americans and Health Care.” *American Journal of Public Health* 87, no. 11, 1773–1778. https://doi.org/10.2105/AJPH.87.11.1773

Gebhard, Catherine, Vera Regitz-Zagrosek, Hannelore K. Neuhauser, Rosemary Morgan, and Sabra L. Klein. 2020. “Impact of Sex and Gender on COVID-19 Outcomes in Europe.” *Biology of Sex Differences* 11, no. 1, 1–13, 2020. https://doi.org/10.1186/s13293-020-00304-9

Gibson, Matthew, and Jeffery Shrader. 2018. “Time Use and Labor Productivity: The Returns to Sleep.” *The Review of Economics and Statistics* 100, no. 5, 783–798. https://doi.org/10.1162/rest_a_00746

Giuntella, Osea, and Fabrizio Mazzonna. 2019. “Sunset Time and the Economic Effects of Social Jetlag: Evidence from US Time Zone Borders.” *Journal of Health Economics* 65: May, 210–226. https://doi.org/10.1016/j.jhealeco.2019.03.007

Gottfried, Michael, J. Jacob Kirksey, and Tina L. Fletcher. 2022. “Do High School Students with a Same-Race Teacher Attend Class More Often?” *Educational Evaluation and Policy Analysis* 44, no. 1, 149–169. https://doi.org/10.3102/01623737211032241

Grasselli, Giacomo, Massimiliano Greco, Alberto Zanella, Giovanni Albano, Massimo Antonelli, Giacomo Bellani, Ezio Bonanomi, Luca Cabrini, Eleonora Carlesso, Gianpaolo Castelli, et al. 2020. “Risk Factors Associated with Mortality Among Patients with COVID-19 in Intensive Care Units in Lombardy, Italy.” *JAMA Internal Medicine* 180, no. 10, 1345–1355. https://doi.org/10.1001/jamainternmed.2020.3539

Herman, D. B., E. S. Susser, E. L. Struening, and B. L. Link. 1997. “Adverse Childhood Experiences: Are They Risk Factors for Adult Homelessness?” *American Journal of Public Health* 87, no. 2, 249–255. https://doi.org/10.2105/AJPH.87.2.249

Heissel, Jennifer A., and Samuel Norris. 2018. “Rise and Shine: The Effect of School Start Times on Academic Performance from Childhood Through Puberty.” *Journal of Human Resources* 53, no. 4: 957–992.

Hussain, Aneela N., Fazal Hussain, and Shahrukh K. Hashmi. 2020. “Role of Testosterone in COVID-19 Patients—A Double-Edged Sword?” *Medical Hypotheses* 144, 110287. https://doi.org/10.1016/j.mehy.2020.110287

James, Spencer L., Degu Abate, Kalkidan Hassen Abate, Solomon M. Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. 2018. “Global, Regional, and National Incidence, Prevalence, and Years Lived with Disability for 354 Diseases and Injuries for 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017.” *The Lancet* 392, no. 10159, 1789–1858. https://doi.org/10.1016/S0140-6736(18)32279-7

Jin, Jian-Min, Peng Bai, Wei He, Fei Wu, Xiao-Fang Liu, De-Min Han, Shi Liu, and Jin-Kui Yang. 2020. “Gender Differences in Patients with COVID-19: Focus on Severity and Mortality.” *Frontiers in Public Health* 8, 152. https://doi.org/10.3389/fpubh.2020.00152

Jones, L., C. I. Karageorghis and P. Ekkekakis. 2014. “Can High-Intensity Exercise be More Pleasant? Attentional Dissociation Using Music and Video.” *Journal of Sport and Exercise Psychology* 36, no. 5, 528–541. https://doi.org/10.1123/jsep.2013-0251

Johnson, A. 1994. “Love of Life.” *Bioscience* 44, no. 5, 363–365. https://doi.org/10.2307/1312386

Jones, M. V., A. M. Lane, S. R. Bray, M. Uphill, and J. Catlin. 2005. “Development and Validation of the Sport Emotion Questionnaire.” *Journal of Sport and Exercise Psychology* 27, no. 4, 407–431. https://doi.org/10.1123/jsep.27.4.407

Kaplan, R. and S. Kaplan. 1989. *The Experience of Nature: A Psychological Perspective*. Cambridge University Press. https://archive.org/details/experienceofnatu00kapl

Klein, Sabra L., Santosh Dhakal, Rebecca L. Ursin, Sharvari Deshpande, Kathryn Sandberg, and Franck Mauvais-Jarvis. 2020. “Biological Sex Impacts COVID-19 Outcomes.” *PLoS Pathogens* 16, no. 6, e1008570. https://doi.org/10.1371/journal.ppat.1008570

Koegel, P., E. Melamid and M. A. Burnam. 2011. “Childhood Risk Factors for Homelessness Among Homeless Adults.” *American Journal of Public Health* 85, 1642–1649. https://doi.org/10.2105/AJPH.85.12.1642

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. “Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks.” *Proceedings of the National Academy of Sciences* 111, no. 24, 8788–8790. https://doi.org/10.1073/pnas.1320040111

Liu, Yong. 2016. “Prevalence of Healthy Sleep Duration Among Adults – United States, 2014.” *Morbidity and Mortality Weekly Report* 65, no. 6, 137–41.

Lombardo, Paul A., and Gregory M. Dorr. 2006. “Eugenics, Medical Education, and the Public Health Service: Another Perspective on the Tuskegee Syphilis Experiment.” *Bulletin of the History of Medicine* 80, no. 2, 291–316. https://doi.org/10.1353/bhm.2006.0066

Madjid, Mohammad, Payam Safavi-Naeini, Scott D. Solomon, and Orly Vardeny. 2020. “Potential Effects of Coronaviruses on the Cardiovascular System: A Review.” *JAMA Cardiology* 5, no. 7, 831–840. https://doi.org/10.1001/jamacardio.2020.1286

McCallum, Jan M., Dhananjaya M. Arekere, B. Lee Green, Ralph V. Katz, and Brian M. Rivers. 2006. “Awareness and Knowledge of the US Public Health Service Syphilis Study at Tuskegee: Implications for Biomedical Research.” *Journal of Health Care for the Poor and Underserved* 17, no. 4, 716–733. https://doi.org/10.1353/hpu.2006.0130

McDermott, Rose and Peter K. Hatemi. 2020. “Ethics in Field Experimentation: A Call to Establish New Standards to Protect the Public from Unwanted Manipulation and Real Harms.” *Proceedings of the National Academy of Sciences* 117, 48, 30014–30021. https://doi.org/10.1073/pnas.2012021117

Mosca, Lori, Elizabeth Barrett-Connor, and Nanette Kass Wenger. 2011. “Sex/Gender Differences in Cardiovascular Disease Prevention: What a Difference a Decade Makes.” *Circulation* 124, no. 19, 2145–2154. https://doi.org/10.1161/CIRCULATIONAHA.110.968792

Musk, Arthur William and Nicholas Hubert De Klerk. 2003. “History of Tobacco and Health.” *Respirology* 8, no. 3, 286–290. https://doi.org/10.1046/j.1440-1843.2003.00483.x

Naqqash, T., S. Hameed, A. Imran, M. K. Hanif, A. Majeed, and J. D. van Elsas. 2016. “Differential Response of Potato Toward Inoculation with Taxonomically Diverse Plant Growth Promoting Rhizobacteria.” *Frontiers in Plant Science* 7, 144. https://doi.org/10.3389/fpls.2016.00144

Nickerson, David W., and Ismail K White. 2013. “The Effect of Priming Racial In-Group Norms of Participation and Racial Group Conflict on Black Voter Turnout: A Field Experiment.” *Unpublished manuscript*. https://polisci.osu.edu/sites/polisc...ct.3.15.13.pdf

Park, Jinbin. 2017. “Historical Origins of the Tuskegee Experiment: The Dilemma of Public Health in the United States.” *Korean Journal of Medical History* 26, no. 3, 545–578. https://doi.org/10.13081/kjmh.2017.26.545

Peckham, Hannah, Nina M. de Gruijter, Charles Raine, Anna Radziszewska, Coziana Ciurtin, Lucy R. Wedderburn, Elizabeth C. Rosser, Kate Webb, and Claire T. Deakin. 2020. “Male Sex Identified by Global COVID-19 Meta-Analysis as a Risk Factor for Death and ITU Admission.” *Nature Communications* 11, no. 1, 1–10. https://doi.org/10.1038/s41467-020-19741-6

Pierce, Matthias, Graham Dunn, and Tim Millar. 2017. “Confounding in Longitudinal Studies in Addiction Treatment Research.” *Addiction Research and Theory* 25, no. 3, 236–242. https://doi.org/10.1080/16066359.2016.1247812

Pivonello, Rosario, Renata S. Auriemma, Claudia Pivonello, Andrea M. Isidori, Giovanni Corona, Annamaria Colao and Robert P. Millar. 2021. “Sex Disparities in COVID-19 Severity and Outcome: Are Men Weaker or Women Stronger?” *Neuroendocrinology* 111, no. 11, 1066–1085. https://doi.org/10.1159/000513346

Plante, T. G., C. Gores, C. Brecht, J. Carrow, A. Imbs, and E. Willemsen. 2007. “Does Exercise Environment Enhance the Psychological Benefits of Exercise for Women?” *International Journal of Stress Management* 14, no. 1, 88–98. https://doi.org/10.1037/1072-5245.14.1.88

Pretty, J., J. Peacock, M. Sellens, and M. Griffin. 2005. “The Mental and Physical Health Outcomes of Green Exercise.” *International Journal of Environmental Health Research* 15, 319–337. https://doi.org/10.1080/09603120500155963

Pretty, J., J. Peacock, R. Hine, M. Sellens, N. South and M. Griffin. 2007. “Green Exercise in the UK Countryside: Effects on Health and Psychological Well-Being, and Implications for Policy and Planning.” *Journal of Environmental Planning and Management* 50, no. 2, 211–231. https://doi.org/10.1080/09640560601156466

Prochaska, James O., Carlo C. DiClemente, and John C. Norcross. 1993. “In Search of How People Change: Applications to Addictive Behaviors.” *Addictions Nursing Network* 5, no. 1, 2–16. https://doi.org/10.1037/0003-066X.47.9.1102

Proctor, Robert N. 2004. “The Global Smoking Epidemic: A History and Status Report.” *Clinical Lung Cancer* 5, no. 6, 371–376. https://doi.org/10.3816/CLC.2004.n.016

Proctor, Robert N. 2012. "The History of the Discovery of the Cigarette–Lung Cancer Link: Evidentiary Traditions, Corporate Denial, Global Toll." *Tobacco control* 21, no. 2, 87–91. https://doi.org/10.1136/tobaccocontrol-2011-050338

Reverby, Susan M. 2012. “Ethical Failures and History Lessons: The US Public Health Service Research Studies in Tuskegee and Guatemala.” *Public Health Reviews* 34, no. 1, 1–18. https://doi.org/10.1007/BF03391665

Rickham, P. P. 1964. “Human Experimentation. Code of Ethics of the World Medical Association. Declaration of Helsinki.” *British Medical Journal* 18, no. 2, 177. https://doi.org/10.1136/bmj.2.5402.177

Ross, Colin A. 2007. “Ethics of CIA and Military Contracting by Psychiatrists and Psychologists.” *Ethical Human Psychology and Psychiatry* 9, no. 1, 25–34. https://doi.org/10.1891/152315007780493799

Rossi, P. H. 1991. *Down and out in America: The Origins of Homelessness*. University of Chicago Press. https://press.uchicago.edu/ucp/books...bo3774726.html

Sampson, Robert J. and William J. Wilson. 2020. “Toward a Theory of Race, Crime, and Urban Inequality.” In *Crime, Inequality and the State*, 312–325. Routledge. https://www.taylorfrancis.com/chapte...-julius-wilson

Scully, Eileen P., Jenna Haverfield, Rebecca L. Ursin, Cara Tannenbaum and Sabra L. Klein. 2020. “Considering How Biological Sex Impacts Immune Responses and COVID-19 Outcomes.” *Nature Reviews Immunology* 20, no. 7, 442–447. https://doi.org/10.1038/s41577-020-0348-8

Shadish, William, T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin. https://www.google.com/books/edition...AAAAMAAJ?hl=en

Shelton, K. H., P. J. Taylor, A. Bonner, and M. Van Den Bree. 2009. “Risk Factors for Homelessness: Evidence from a Population-Based Study.” *Psychiatric Services* 60, no. 4, 465–472. https://doi.org/10.1176/ps.2009.60.4.465

Shuster, Evelyne. 1997. “Fifty years later: The significance of the Nuremberg code.” *New England Journal of Medicine* 337, no. 20, 1436–1440. https://doi.org/10.1056/NEJM199711133372006

Stets, Jan E. 1991. “Cohabiting and Marital Aggression: The Role of Social Isolation.” *Journal of Marriage and the Family* 53, no. 3, 669–680. https://doi.org/10.2307/352742

Strauss, Dana, Sara de la Salle, Jordan Sloshower, and Monnica T. Williams. 2021. “Research Abuses Against People of Colour and Other Vulnerable Groups in Early Psychedelic Research.” *Journal of Medical Ethics* 48, 728–737. https://doi.org/10.1136/medethics-2021-107262

Susser, Ezra, Robert Moore, and Bruce Link. 1993. “Risk Factors for Homelessness.” *Epidemiologic Reviews* 15, no. 2, 546–556. https://doi.org/10.1093/oxfordjournals.epirev.a036133

Takahashi, Takehiro, Mallory K. Ellingson, Patrick Wong, Benjamin Israelow, Carolina Lucas, Jon Klein, Julio Silva, Tianyang Mao, Ji Eun Oh, Maria Tokuyama, et al. 2020. “Sex Differences in Immune Responses that Underlie COVID-19 Disease Outcomes.” *Nature* 588, no. 7837, 315–320. https://doi.org/10.1038/s41586-020-2700-3

Tankard, Margaret E. and Elizabeth Levy Paluck. 2017. “The Effect of a Supreme Court Decision Regarding Gay Marriage on Social Norms and Personal Attitudes.” *Psychological Science* 28, no. 9, 1334–1344. https://doi.org/10.1177/0956797617709594

Tennenbaum, G., R. Lidor, N. Lavyan, K. Morrow, S. Tonnel, A. Gershgoren, J. Meis, and M. Johnson. 2004. “The Effect of Music Type on Running Perseverance and Coping with Effort Sensations.” *Psychology of Sport and Exercise* 5, no. 2, 89–109. https://doi.org/10.1016/S1469-0292(02)00041-9

Unger, Michael A. 2008. “After the Supreme Word: The Effect of McCreary County v. ALCU (2005) and Van Orden v. Perry (2005) on Support for Public Displays of the Ten Commandments.” *American Politics Research* 36, no. 5, 750–775. https://doi.org/10.1177/1532673X08315856

Wilson, William Julius. 2012. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press, 2012. https://press.uchicago.edu/ucp/books...o13375722.html

Witschi, Hanspeter. 2001. “A Short History of Lung Cancer.” *Toxicological Sciences* 64, no. 1, 4–6. https://doi.org/10.1093/toxsci/64.1.4

Wolfe, Jeannette, Basmah Safdar, Tracy E. Madsen, Kinjal N. Sethuraman, Bruce Becker, Marna Rayl Greenberg, and Alyson J. McGregor. 2021. “Sex-_or Gender-Specific Differences in the Clinical Presentation, Outcome, and Treatment of SARS-COV-2.” *Clinical Therapeutics* 43, no. 3, 557–571. https://doi.org/10.1016/j.clinthera.2021.01.015

Wooler, J. J., J. Barton, V. F. Gladwell, and D. Mickelwright. 2015. “Occlulsion of Sight, Sound and Smell During Green Exercise Influences Mood, Perceived Exertion and Heart Rate.” *International Journal of Environmental Health Research* 26, 267–280. https://doi.org/10.1080/09603123.2015.1109068

Wynder, Ernst L. 1997. “Tobacco as a Cause of Lung Cancer: Some Reflections.” *American Journal of Epidemiology* 146, no. 9, 687–694. https://doi.org/10.1093/oxfordjournals.aje.a009342

Xu, F., A. Meng, Y. Liu, J. Li, and N. Wu. 2025. “Effects of New Special Formula Fertilizer on Potato Growth, Yield, and Fertilizer Utilization Efficiency.” *Plants* 14, 627. https://doi.org/10.3390/plants140406

Yeh, H.-P., J. A. Stone, S. M. Churchill, E. Bymer, and K. Davids. 2017. “Physical and Emotional Benefits of Different Exercise Environments Designed for Treadmill Running.” *Environmental Research and Public Health* 14, no. 7, 752. https://doi.org/10.3390/ijerph14070752

This page titled 3.12: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

CHAPTER OVERVIEW

4: The Nature of Populations

"Without data, you're just another person with an opinion."

—W. Edwards Deming

4.1: What to Expect
4.2: Defining a Population
4.3: Practicality
4.4: Established Data
4.5: Defining Populations in Litigation
4.6: Critical Analysis
4.6.1: The Presidential Election of 1936
4.6.2: Young Women in the Justice System
4.6.3: Phone Use Policies in Schools in England
4.6.4: Same-Sex Marriage and LGBT Discrimination
4.6.5: Sampling Animals in Populations
4.6.6: Land Cover and Racial Segregation
4.6.7: Tobacco Retailer Density
4.7: Summary
4.8: About the Quote
4.9: Practice and Exploration
4.10: Collaborative Activity
4.11: References

This page titled 4: The Nature of Populations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.1: What to Expect

Statistical studies are often centered on groups of individuals in society that may be affected in some way due to one or more of their characteristics. These groups are called populations. It is important in these studies that the population of interest be precisely defined to match the interests of the researcher. This chapter will discuss and define populations along with some important considerations and characteristics of these populations.

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

- Define and discuss populations, data, variables, and parameters.
- Consider the importance of the practicality of studying certain populations.
- Consider what established sources of data may be helpful to a research study.
- Consider the importance of correctly defining a population in discrimination litigation.

This page titled 4.1: What to Expect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.2: Defining a Population

When researchers consider an issue they would like to study, there is a group of individuals who are the focus of the study. In the theory of statistics, these groups are called *statistical populations*, or simply *populations*. For example, if a researcher is interested in comparing the average income across self-identified genders for all adults in a large city, then the population consists of all adults in the city. Similarly, if a researcher is interested in the potential presidential voting profiles of all eligible voters for the next presidential election in the United States, then the population consists of every individual in the United States who has or can register to vote prior to the next presidential election. Populations can also consist of physical objects or other living beings that are not people. For example, a study might consider the number of grocery stores that offer fresh fruit and vegetables in each neighborhood of a city to determine if access to fresh produce is associated with the income level of the neighborhoods. In this case, the population of interest is all the grocery stores in the city. In other cases, the population may not consist of individuals or physical objects at all, but rather an idealized set of observations that could be made under the right conditions. These hypothetical or theoretical populations may even have an infinite number of members in the population. For example, it might be of interest to compare the waiting times of emergency room patients at urban, suburban, and rural hospitals. These waiting times do not physically exist—they are observations that will be made of an event that will happen later. Nevertheless, this population may be of interest and could be theoretically infinite, and it relates to many important issues in our modern world.

Definition: Population

A *population* consists of all individuals or items that are of interest in a statistical study.

It is crucial in the development of a statistical study that the population of interest be defined precisely. The importance of this precise definition is because the population is the source of the observed data that will be used to learn about the issue of interest. The defined population should only include individuals or items that are of interest, while containing all the individuals or items of interest. The definition of the population of interest is guided by the question or hypothesis about the topic that the researcher is interested in investigating. In most cases researchers will use an iterative process of refining the relationship between the research question and the definition of the population many times before settling on a precise statement of each.

As an example, suppose that a researcher is interested in studying the association between immigration status and yearly income. The population would include everyone in the United States. The associated population is quite large and complex, and there are many questions that arise as one considers the problem under consideration. What is meant by yearly income? Certainly, individual income could be observed, but how would that vary within family units? Some members of a family may be of an age that would make work outside the family possible, but they may be tasked with watching after children or siblings to support others in the family who work. As such, many in the population potentially have no income but support those who bring income into the family unit. In this situation it is also relevant that the income from the people who work outside the family would be shared between the family members, whereas a single person with the same income would support only themselves. At this point some refinement of the research question seems to be in order.

The researcher may decide to focus on the income of individuals who work outside the home. By making this refinement, the population is then focused on individuals in the United States who work outside the home. The research question is further narrowed down to consider the individual income of workers employed outside the home. Now a time element becomes important. Some individuals may be working currently but may be unemployed later, and vice versa. In this case the researcher may refine the population to be those employed at the time of the study. The research question then is refined to study the relationship between the yearly income of individuals who were employed at the time of the study. Further refinements in the population may be considered as the research question is refined further.

The population in surveys usually refers to the actual physical individuals or items of interest. To learn from this population, characteristics and measurements are observed from these individuals. For example, in a study of consumer credit ratings, the population of interest may be all individuals living within a specified set of zip codes between 18 and 65 years old. When individuals from this population are observed for the study, we may be interested in several characteristics of the individuals. For example, we might determine the credit rating of the individual, along with other demographic data like race, age, employment status, and gender identity. Collectively these measurements are known as data.

Definition: Data

Measurements taken on individuals or items in a population are called *data*. A single measurement is called a *datum*.

Many measurements may be taken on each observed individual from the population. Each distinct measurement taken on the individuals in a population is called a variable.

Definition: Variable

A single measurement that is taken on everyone observed in a population is called a *variable*.

Some practitioners simply define variables as quantities that vary, but this simplistic definition is unsatisfactory from a practical viewpoint, as not all variables vary between individuals and, more importantly, it is not a mathematically sound definition (Polansky 2011). In the previous example, the measurements obtained from the observed individuals from the population are the credit rating, race, age, employment status, and gender identity of the observed individuals. Taken collectively these observations and variables constitute the data observed for the study. The measurements of credit rating of the individuals in the population are referred to as observations of the credit rating variable. The other variables in the study correspond to the observations of race, age, employment status, and gender identity.

Once a population and variables are defined, the variables' quantities need to be determined. In the previous example we are interested in how income, as defined above, varies with immigration status. We might decide that we could answer this question definitively by comparing the average income of everyone in the population aggregated across each possible immigration status. These averages are examples of parameters, quantities that can be computed from the known values of the variables in a population.

Definition: Parameter

A *parameter* is a quantity or characteristic that can be computed when the whole population has been observed.

The key idea here is that the value of a parameter is only known when the variables on every individual or item in a population is observed. In the example, if we observed everyone in the defined population along with their income, we could compute the average value of the income variable within each immigration status. We could not compute these values unless everyone in the population could be observed, and the income and immigration status of each individual could also be observed.

As another example, let us consider a small population consisting of two small classes of graduate students in advanced study at a university. There are seven students in the first class where a traditional instructional approach was taken, and ten students in a second class where the same material was taught but from a diverse perspective with respect to gender. The instructor of these classes is interested in the possible effect that the gender diverse instructional method may have on the final grades in the classes. The instructor is also interested in comparing the grades across gender identities in the classes. Hence, we have a population of seventeen students with two variables: final grade and gender identity. The population is reported in Table 4.1 below.

Table 4.1. A population of seventeen students in two classes. The student names are censored with labels. The classes are identified as traditional (T) or gender diverse (D). All students in the study identified their genders as female (F), male (M), or nonbinary (N).

Student	Class	Gender Identity	Grade
A	T	M	68
B	T	N	57
C	T	F	74
D	T	M	78
E	T	F	64
F	T	F	85
G	T	M	96
H	D	M	56
I	D	N	83
J	D	F	68
K	D	F	77
L	D	M	89
M	D	N	70
N	D	F	75
O	D	N	72
P	D	M	71
Q	D	F	87

For this study we are interested in comparing the average grades across the classifications of gender identities and the two classroom experiences. Therefore, we can take these averages as our parameters. The population is small, and since we know the gender identity and final grade for everyone, we can compute the parameter values directly. Recall that an average grade is computed by adding up all the scores and then dividing by the number of scores. Therefore, as an example, the average final grade for students who identify as female in the traditional course is $(74 + 64 + 85) \div 2 = 74.3 \approx 74$ where we have rounded to the nearest whole number. The symbol $\approx$ is used to indicate that the value following the symbol is an approximate value, in this case due to rounding the value 74.3 to 74. The remaining averages are given in Table 4.2.

Table 4.2. The population parameters of seventeen students in two classes, rounded to the nearest whole number. The classes are identified as traditional (T) or gender diverse (D). All students in the study identified their genders as female (F), male (M), or non-binary (N).

Gender Identity	Class	Average Grade
F	T	74
M	T	81
N	T	57
F	D	77
M	D	72
N	D	75

We will not attempt to draw any conclusions from this hypothetical study as this population is incredibly small, and any conclusions would probably not extend to larger populations. The important conceptual idea at this point is to make the connection

between the members of the population, the variables, and the parameters.

This page titled 4.2: Defining a Population is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.3: Practicality

Section 4.2 considered a hypothetical study of the differences in income across classifications of immigration status. As the discussion progressed, the definition of the population of interest was refined in the interest of practicality. Practicality is a key issue in statistical studies. Populations can be large, complex, and members of the population may be difficult to contact. Practitioners who study populations also have limited practical means due to many unavoidable constraints such as cost, time, and accessibility. Additionally, as we will see later, it is important that each member of the population is accessible for observation to the researcher, even though the researcher may only look at a small part of a population.

Refining a population definition has a consequential effect on the applicability of the results of the study. In the hypothetical study of the differences in income across classifications of immigration status, it was indicated that the study covered the population of residents in a specific city. When the research study is completed, the results of the study would apply only to those individuals who were in the population of interest. Therefore, if the study concluded that a certain immigration status has a negative effect on income, those conclusions could not automatically be assumed true for other cities—unless it could be argued that the populations and the social and employment environments of the two cities are roughly the same.

Similarly, in the study of the effect a specific teaching approach has on students with various gender identities, the results are narrowly focused on what happened in those two classes. Attempting to generalize this study to the application of these teaching methods is difficult, as changing any of the teaching parameters may influence the student performance. It is quite difficult for a study of this nature to fully attribute any observed differences in the performance of the students to the teaching methodology unless great care is taken to control for as many differences between the two classes as is possible. For example, one could argue that there was a difference in performance because the instructor believed in one of the approaches more than the other and did a more competent job for that class. Therefore, attempting to imply that the results support the general notion that a change in teaching method would result in better student performance is tenuous at best, unless a more substantial justification accompanies the analysis.

With the applicability of the results in mind, there is a fundamental interest in considering as large a population as is possible so that the results of the corresponding study will have the widest applicability. For example, in the hypothetical example of income and immigration status, a study that considered the entire population of the United States would have wide applicability. If that type of study were possible, the results could be applicable across the country and could be the basis for legislative and social policy action. In the hypothetical study of teaching styles, a population that is focused on multiple instructors, departments, class subjects, and even universities would have the potential for generalizing the applicability of the results to broader populations.

So, what prevents researchers from always considering the largest possible population with the largest possible applicability for every study? The problem, as we will study in greater detail later, is that researchers need to have the ability to observe data from each of the individuals or items in the population. We will usually not observe all the members of a population, but we need to be able to have that opportunity.

The size and location of a population can influence the potential observability of the members of the population. The hypothetical study of income and immigration status can serve as a good example of the problems that can occur with large populations. First consider the study income and immigration status with the intention of considering the entire population of the United States. What we need to keep in mind is that we would at least, hypothetically, need to be able to observe all the members of this population. So, how could we do that? Do we have a list of everyone employed, their immigration status, and how to get in contact with them? Unless we work for a large government agency, it would be difficult to imagine that we do. Even if we did have an official list of those employed in the United States, would it really include everyone? Again, this is very unlikely, as there are many who work in jobs that are paid in ways that are not easily traceable even by the government. Therefore, we may not really know about these individuals and, most importantly, these workers may be people with certain immigration statuses. Even more importantly, because of the nature of this type of work, the pay may be lower than that of the rest of the population. Unfortunately, this is the very question we are attempting to answer. Hence, if these individuals are systematically left out of the study, then we may get a skewed idea about the incomes and how they relate to immigration status.

Getting in touch with individuals who would be unofficially employed, perhaps with an immigration status that would make them less likely to share personal information, requires a very detailed knowledge of the area and of the people that work in the area.

Gathering data from these individuals also requires establishing a trust-based relationship. This type of knowledge and these relationships are time consuming and resource intensive to develop. Hence, unless a researcher has access to local knowledge, along with the resources and relationships required to work with such populations, a study with broad based applicability may not be feasible in practice.

In this case, researchers may elect to work with a smaller population, or at least a more local population to save on resources and time. While this tactic has a fundamental effect on how widely the results of the study can be interpreted, the alternative is often not to do the study at all. In this case localized results may be the best option for obtaining some useful information with regards to the study hypotheses.

This page titled 4.3: Practicality is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.4: Established Data

Many studies take advantage of established datasets that are available either publicly or accessed with permission from the owners of the data. In this way researchers can employ the resources of much larger organizations for their studies.

The United States federal government is constantly gathering large amounts of data on the economic conditions, education, crime, and health issues of those living within the United States and its territories. Much of this data can be accessed quite easily, though some of the observations will be blinded in some way to protect the privacy of those included in the data. For example, names and Social Security numbers are not available, as well as other specific information that would identify residents. Nonetheless, these databases provide invaluable information about the individuals in the population. These data have been gathered by a dedicated group of professionals using the vast number of resources and associated infrastructure available to the federal government. The availability of this data allows individuals and researchers to take advantage of their expertise and resources for their own research.

The United State Census Bureau has a wealth of information that is collected every ten years for the constitutionally mandated census in the United States. Much of the collected data is available in summaries, reports, and dynamic data tools which can be found at www.census.gov. Observation level data is also available to academic and research organizations by requesting a special key and using specialized data extraction software. It is important to note that some of the data is censored so that individuals cannot be identified from the data that is supplied. Specifically, as mandated by federal law, the following is stated on the Census Bureau website:

> Federal law [13 U.S.C. §§ 8 and 9] requires the Census Bureau to publish data in a manner that does not permit use of the data alone, or in combination with other available information, to identify any particular respondent to a Census Bureau survey. The identification of a person or establishment or disclosure of information collected from that person or establishment in the course of a Census Bureau survey violates the assurances of confidentiality provided by federal law.

These types of restrictions are rarely prohibitive in research studies as the need to identify specific individuals in such a population is rarely necessary. After 50 years, full data are released to the public.

Additionally, the United States government runs the data.gov website, which is a collection of data that is available to anyone. The website was launched in late May 2009 by the Federal Chief Information Officer of the United States, Vivek Kundra, with the purpose of improving public access to datasets generated by the executive branch of the federal government. By 2022, the site listed almost 350,000 datasets that could be accessed online, including data from the federal government, state governments, city governments, county governments, and universities. Among the federal government data, agencies such as NOAA (National Oceanic and Atmospheric Administration), NASA (National Aeronautics and Space Administration), the Department of Commerce, and the Department of the Interior have data accessible from the site.

Using government data of this sort has many advantages for researchers. Access to the data is free and has been vetted by government agencies to ensure quality. For many studies much of the data comes from surveys performed by the Department of Education, the Department of Justice, and the Census Bureau. The Department of Education has a specific section dedicated to civil rights data that includes district level demographic data. Another section is dedicated to data on student outcomes. The National Center for Education Statistics was created as a part of the Department of Education by congressional mandate to “collect, collate, analyze, and report complete statistics on the condition of American education” (nces.ed.gov/about). The center's website contains additional educational data along with a tool that can link researchers from a dataset to published and peer-reviewed research articles that have analyzed that data.

The U.S. Department of Justice is a large government entity that maintains data through its agencies such as the Bureau of Justice Statistics, Federal Bureau of Investigation, Federal Bureau of Prisons, National Institute of Justice, Office of Juvenile Justice and Delinquency Prevention, and the U.S. Trustee Program. A complete list of agencies can be found at justice.gov, many of which provide data. These resources are valuable to researchers studying issues related to crime and the judicial system.

From the Bureau of Justice Statistics website (bjs.ojp.gov) data is available by topic, including data concerning corrections, courts, the federal justice system, and victim of crimes. For example, under the topic of victims of crime, the site includes a city-level

survey of crime victimization and citizens' attitudes, a national survey of victim service providers, and emergency room statistics on intentional violence. Reports can also be found on the site.

The Federal Bureau of Investigation (FBI) hosts the Uniform Crime Reporting Program (UCR), which can be used to generate reliable crime statistics (www.fbi.gov/services/cjis/ucr). The UCR Program includes data from more than 18,000 city, university and college, county, state, tribal, and federal law enforcement agencies, who voluntarily submit crime data. The databases include information on hate crime statistics, use of force, and suicide.

The Department of Commerce (data.commerce.gov) maintains a site for data from entities including the Bureau of Economic Analysis, the Bureau of Industry and Security, the Economic Data Administration, and the Minority Business Development Agency. The data included on this site includes observations on all types of economic activity.

Other established data may be available through universities or other nonprofit academic agencies. For example, it is becoming more commonplace that data availability is a prerequisite for the publication of research articles. This is due in part to the idea that the research community should be able to verify the results of a research study. This access helps maintain the integrity of the research area. Furthermore, there has been an effort to make publicly funded research data available to the public. As with census data, these datasets may censor some observations to protect the privacy of those who participated in a study or survey.

This page titled 4.4: Established Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.5: Defining Populations in Litigation

While there are many fields that have interesting questions about population definition, the field of litigation, particularly when applied to discrimination cases, provides a particularly interesting example of the types of questions that need to be considered when defining a population. While such questions are obviously not relevant to all situations when defining a population, the type of questions considered may be relevant in many other fields.

Discrimination litigation requires specialized experts in both law and statistical procedures. While some of the relevant statistical methodology will be reviewed later, defining the proper population for such studies is crucial. The final arbiter of what the relevant population is for such cases is the judge overseeing the case, and this decision is often made after receiving evidence.

The population used in discrimination litigation can be crucial to the outcome of the case. Lawyers and professional statisticians who specialize in expert witness testimony for these types of cases must be able to help gather evidence in relation to these issues for many potential populations. They must also be able to argue why a specific population should be used over a different choice. As Elaine Shoben, a professor of law who specializes in discrimination cases, stated, “A (statistical) expert's superb analysis on legally irrelevant data is useless” (Shoben 1986).

Discrimination cases often deal with employment and therefore much of the concern is over potential applications to a position. In these cases, a court may decide to compare the pool of applications for a position and compare it to those who were hired. In other cases, a court will compare the general population of all possible applicants versus characteristics of the employer's workforce, or the court will consider the population of potentially qualified applications to a position. This latter case can be used when the job has special qualifications, such as a certification or an advanced degree, or requires certain physical attributes, such as being able to lift a certain amount.

Definition: Applicant Flow

The *applicant flow* for a position is the collection of individuals who apply for a job or promotion.

A court will often compare the race, gender, ethnicity, and sexual orientation of those in the application flow against those who were hired. Until relatively recently the guidelines for determining if a particular group was discriminated against was based on a rule known as the 80% rule. Using this rule the hiring rate of any group cannot be less than 80% of the rate of another group (Shoben 1986). For example, if employees are classified by gender as female or male, the rate at which females are hired must be at least 80% of the rate at which males are hired. Therefore if 35% of the female applicants are hired and 70% of the male applicants are hired, the hiring rate of female applicants is $35 \div 70 = 0.5$, or 50% of that for male applicants, yielding evidence of discrimination. On the other hand, if 60% of the female applicants are hired and 70% of the male applicants are hired, the hiring rate of female applicants is $60 \div 70 = 0.86$, or 86% of that for male applicants. In this case these results do not provide evidence of discrimination. This is a controversial and arbitrary measure, and alternative methods have been proposed (Boardman 1979; Bobko and Roth 2004; Sobol and Ellard 1988; Gastwirth and Miao 2009; Hauenstein et al. 2013; Ismail and Kleiner 2001).

The applicant flow may not be acceptable as a comparative population as it may not be representative of those in the general population who would like to apply for the position. Potential applicants may be excluded by the recruitment methods of a company, producing a biased applicant flow. The reputation of a company may also affect the applicant flow. For example, a company may have a history of discriminatory practices or a poor working environment for those who are members of certain social groups. In these cases, the relevant population may consist of those who would apply for the job if there were no barriers to enter the applicant flow.

If the applicant flow cannot be used, then courts may turn to considering the general population of workers. The assumption in this case is that the composition of the general workforce is representative of the population of workers who want the job in question. This can be problematic in that not all the individuals in the general work force may be interested in all jobs, and some jobs may require special training or certification which individuals in the population may not have.

The Supreme Court addressed this issue in the case of *Hazelwood School District v. United States*, which concerned the hiring of African American teachers in a Missouri school district in the suburbs of Saint Louis. A district court ruling compared the ratio of African American teachers to white teachers to the corresponding ratio of African American students to white students in the

school. This decision was appealed to the Court of Appeals, which rejected this comparison and instead used hiring statistics for teachers in the geographic area, which includes the Saint Louis metropolitan area. This move limited the population to those who were qualified for a job that required a teaching certificate. This decision was appealed to the Supreme Court as the school district argued that the Saint Louis school districts had enacted hiring practices that were designed to specifically increase the number of African American teachers. There were many different and complicated legal questions relevant to the case, and the Supreme Court sent the case back to the district court but established that the relevant comparison population is those who possess the required skills for the position—in this case, a teaching certificate.

General populations may be refined in other ways. One of the most common refinements is based on geographical areas for employment opportunities that are unlikely to attract potential employees outside of a local area. In these cases, the local work force may be used for comparison rather than a more global work force. Localization may also be weighted for surrounding areas by the percentage of applications coming from each local area (Shoben 1986). As populations and the work force can shift with time, the periods considered in litigation can be an important aspect of a case as well.

An interesting issue is that the refining of populations in these ways can have consequences in terms of statistical theory and interpretation. In fact, these methods can be used to manipulate the data analysis, and legal officials may not always be aware of this tactic (Kahneman et al. 1982). Some of these issues will be discussed later.

In cases where a work requirement is challenged, the issues of the relevant population changes slightly. Here the main question is whether a work requirement is discriminatory as opposed to the hiring process of the employer. In these cases, the pool of applicants is not the correct population for comparison purposes because applicants who may have been discriminated against by the work requirement have already been excluded from the population. Hence, the focus must be on the population of potential employees who might wish to apply for the job. This was the case in the Supreme Court case *Dothard v. Rawlinson*, where a claim that a height requirement for Alabama prison guards disproportionately excluded women. The defendants in the case sought to use the applicant pool, but this idea was rejected because the height requirement was known to the applicants, and therefore those who did not fit the requirement did not apply. To see why this is a problem, suppose that half of the population of possible applications identified as women. If there was no hiring bias, one would expect that roughly half of those hired would be women. Now suppose that 50% of the women in the population of potential applicants meet the height requirement while 90% of the remaining potential employees meet the height requirement. It can be shown that about 36% of the applicant pool will identify as female. If the defendant was allowed to use the applicant pool, they could argue that only 36% of the hired worked force should be female.

Figure **4.5.1** demonstrates how this calculation is accomplished using a small population of 20 individuals. In this population, there are 10 potential applicants who identify as female, and 10 applicants who do not. In the left column of the figure are shapes representing the potential applicant pool, with squares representing the female potential applicants and circles representing the remaining potential applicants. The potential applicants that satisfy the height requirement are indicated in blue, and those who do not are indicated in red. Because half of the population of potential female applicants satisfies the height requirement, five of the ten squares are indicated in blue. Because 90% of the remaining population satisfies the height requirement, nine of the ten circles are indicated in blue. Only the applicants who satisfy the height requirement will apply for the job, and so only the blue shapes proceed to the applicant pool. Therefore, in this simple example, the applicant pool will have five applicants who who identify as female and nine other applicants. Hence, there are 5 individuals in the applicant pool. Computing the percentage of female applicants is then accomplished using the calculation

$$\frac{5}{5+1} \times 100\% = \frac{5}{14} \times 100\% = 36\%.$$

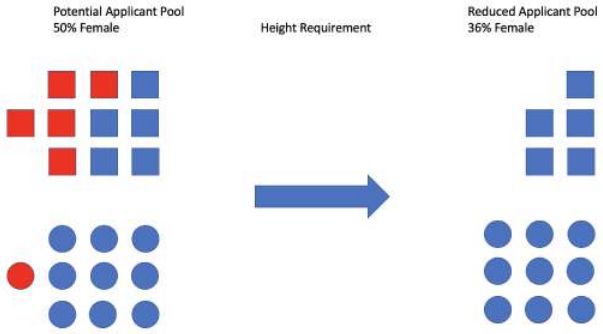


Figure **4.5.1**: An example showing how a height requirement reduces the percentage of female applicants for employment. In the left column shapes represent the potential applicant pool, with squares representing the female potential applicants and circles representing the remaining potential applicants. The potential applicants that satisfy the height requirement are indicated in blue, and those who do not are indicated in red. (Public domain image created by Alan M. Polansky)

In the case of *Dothard v. Rawlinson* the Supreme Court used national demographic figures to consider the case. The court ruled in an 8 to 1 decision that the height restriction was discriminatory. The basis of this decision was because the defendants had not proven that the height standard was necessary for effective job performance. Essentially it was found that the defendants were using the height requirement—which was not necessary for job performance—to exclude women from the applicant pool.

This page titled 4.5: Defining Populations in Litigation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.6: Critical Analysis

When considering a critical analysis of a research study, it is important to first consider the population being studied. A good research study will clearly identify the population, including details on the size of the population, its accessibility to the researchers, and information on why the population is relevant to the research study.

If the researchers are accessing the population themselves, and not relying on an established database, then they should consider the relation between the authors and the population. A small research group considering a large and complex population may be questionable, unless there are specific reasons, such as funding from a government agency, that would allow them to have the resources to conduct such research. Such problems are somewhat alleviated using internet-based surveys, which are very inexpensive and relatively simple to deploy. However, as we shall discuss later, internet-based surveys can be misleading, unless they are carefully designed.

Larger research groups and institutions may have additional resources and experience that allow them to study large and complex questions more easily. That does not necessarily mean that you can automatically trust studies from these types of groups, and you will still need to carefully consider how well the population is defined and the methods used to access the population. Correspondingly, you should not automatically distrust research performed by smaller and perhaps less established research groups. You might find that they have worked very hard to define a population for which they have access. The key is to assess, to the best of your ability, how well the population is defined and whether it appears that the researchers had reasonable access to the population.

If the researchers are using an established dataset, you should first assess whether the population defined by those who obtained the data is well defined. The emphasis in this case is whether the population described by the originators of the data matches the population of interest in the study. If these two population definitions do not match, then the authors of the study are not studying the population that they claim to be studying. In many cases, the population of interest for the study is a subset of the population used to establish the dataset. The researchers will use only a part of the established dataset, excluding cases that are not within their defined population. This type of sub-setting of data is common and is usually used to reduce the number of potential confounding factors in the study. However, when assessing such a study, one should consider what bias might result from using a subset of the population. Subsets of a population can have very different characteristics than the entire population. Hence, the authors should provide a clear explanation as to why a subset of the data has been used, and what limitations and biases may result.

Next, if the observations are based on an established dataset, one should critically assess the source of the data. The source can be very important in establishing a critical analysis of a study. Data obtained from government institutions have usually been gathered using peer-reviewed techniques. Additionally, federal and state governments usually have efficient resources to obtain reliable data. Government agencies also have access to government databases to which private agencies will not have access. Private foundations and companies will also sometimes provide useful data, but one should be careful in assessing these types of studies. You should carefully consider the possibility that a private entity, even if it is a nonprofit organization, might have specific motives for presenting or gathering data in a certain way. It is also worthwhile to check the affiliations of the authors of a study to see if they are in any way connected with the private organization that obtained the data. The main concern with this type of data is assessing whether the source of the data has potential interests in the outcome of the study.

One of the important concepts central to the modern scientific method is reproducibility. That is, other scientists should be able to replicate the results of any experiment on their own. Such reproducibility can be assured by granting public access to data, allowing other researchers to explore the same data and replicate any reported analyses of the data. Recently many research journals began to require either public access to research data, or instructions on how the data can be obtained. Access is not always possible, as some data may be proprietary or may contain sensitive personal information. While you may not be interested in reproducing the results shown in a research article on your own, studies based on publicly available data have an additional level of validity in that the data is readily available to the research community.

Another advantage of assessing studies that use publicly available data is that there may be other research reports that used the same data. One can look for studies that either support or refute the study you are considering. Even if the other studies do not address the same research questions, they may provide additional insight into how the data was obtained and what potential issues may be contained in the data.

In terms of the population, one should also be sure that the conclusions of a study match with the population that is being studied. For example, the results from a study of the economic habits of individuals in a certain neighborhood of a city would generally not be applicable to individuals in other neighborhoods. If the methodology and analysis of the study have been carefully designed and properly executed, the results will be relevant to the population that was used to observe the data. Attempting to make the same conclusions about larger or different populations is an example of extrapolation and is very dangerous.

Definition: Extrapolation

Extrapolation refers to making conclusions for populations that are larger than the population that was observed.

Extrapolation may be warranted when it can be shown that other populations have very similar characteristics to the one that was studied. This is something that the researchers need to carefully document and justify before those conclusions should be taken as reliable.

Most peer-reviewed academic journals insist that researchers include a section that explains the potential limitations of their research. Many of these limitations are often based on what population was observed. These limitations should be carefully considered as they are the key to understanding how universal the results may be, and where further research is required.

In summary, the key questions in critically analyzing the population in a study are:

1. Is the population well defined?
2. Do the researchers have the resources to access the population?
3. Does the population of an established dataset match the researcher's population of interest?
4. Do the sources of the data have potential interests in the outcome of the study?
5. Has a subset of an established set of data been used? Why? Are there potential biases that could result from sub-setting the data?
6. Have the authors clearly stated the limitations of the conclusions of the study based on the definition of the population?
7. Is the data publicly available?
8. Do other research reports use the same data source?
9. What are the conclusions of those research reports?
10. Are the conclusions relevant to the same population where the data was observed? Are the researchers attempting to extrapolate? Is the extrapolation justified?
11. What are the limitations of the study? Where is further research required?

This page titled 4.6: Critical Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.6.1: The Presidential Election of 1936

A classic example of the dangers of not carefully considering the population of interest comes from the results of the United States presidential election of 1936. Alfred M. Landon was the Republican candidate and Franklin D. Roosevelt was the Democratic candidate. The United States was still amid the Great Depression and Roosevelt was running on a platform that the social programs of the New Deal should be continued to help those who were unemployed and living in poverty. While Landon accepted much of the New Deal, he considered it to be hostile to business, wasteful, and inefficient. Landon also accused Roosevelt of acquiring so much power that he was subverting the Constitution. The candidates also differed on issues of foreign policy, with Landon having a somewhat isolationist attitude.

The *Literary Digest* magazine conducted a very large survey of approximately 2.4 million individuals in one of the largest political polls ever conducted (Figure **4.6.1.1**). The magazine had successfully predicted the outcome of every election since 1916. Over the course of several issues, the magazine updated the results of this survey and concluded that Landon would easily win the election, a surprising conclusion to the political experts of the time. Some partial results from this survey are shown from a page in the October 24, 1936, issue of the magazine in Figure **4.6.1.2**. In an astounding defeat for the magazine, Roosevelt easily won the election by securing 62% of the population vote and 523 of the 531 possible electoral votes. The results were devastating to the integrity of the magazine, which went bankrupt shortly afterward.
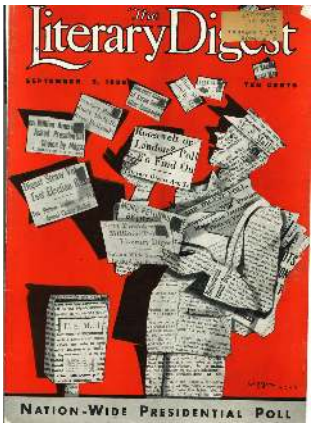


Figure **4.6.1.1**: Cover of the September 5, 1936, cover of *Literary Digest*, touting their nationwide presidential poll (Photograph by the author).



Figure **4.6.1.2**: Partial results of the nationwide presidential poll from the October 24, 1936, issue of *Literary Digest* (photograph by the author).

A large survey would be generally thought to provide reliable results but, in this case, there is a difference between the population that the researchers thought they were observing and the population they were actually observing. To implement the survey, the magazine mailed questionnaires to ten million people using addresses from telephone books and club membership lists (Freedman et al. 2007). Hence, the actual population studied corresponded to individuals who had telephones and belonged to clubs. During the Great Depression this population was overwhelmingly wealthy and supported Landon. This differed significantly from the

general population who supported Roosevelt to a greater extent. It is this difference that caused the major error in their poll and the ultimate demise of the magazine.

Interestingly, George Gallup (Figure **4.6.1.3**), who would become famous for the well-known Gallup poll, correctly predicted the outcome of the election by observing from a population whose makeup was closer to that of the actual voting population for the 1936 presidential election. In fact, Gallup also successfully predicted what the results of the *Literary Digest* poll would be well in advance of the publication of their results using a much smaller survey.



Figure **4.6.1.3**: George Gallup (1901–1984) was a pioneer of survey techniques and inventor of the Gallup poll, a successful survey for measuring public opinion (public domain image).

This page titled 4.6.1: The Presidential Election of 1936 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.6.2: Young Women in the Justice System

The attitudes of young people who become involved with the justice system may provide important insights into their chance of success when they re-enter society. One study of this issue was based on observations and in-depth interviews with 36 young women at a public residential drug treatment facility and juvenile detention center. Most of the study's participants were in juvenile justice facilities multiple times before, were legally adults, and had come from problematic homes. Many had been exposed to violence. Most of the participants were in the facility on either drug charges or probation violations connected to drug use and saw substance use as a “problem” for them. The interviews generally showed that the participants perceived that they had little support to help them when they left the facility (Myers 2017).

The population that the researcher studied in this case is quite specifically the young adult female residents of the facility at the time of the study. While the study does provide important and detailed insights into the perceptions of these young women, it would be difficult to generally extend these results beyond the facilities included in the study, particularly because the facilities were run at the county level. Even within the same state, rehabilitation options, access to counseling, and employment assistance may differ from county to county and even from facility to facility.

This page titled 4.6.2: Young Women in the Justice System is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.6.3: Phone Use Policies in Schools in England

The rate of adolescent mental health problems has been increasing, especially in Western countries (Blakemore 2019). Mental health problems are associated with higher rates of disruptive behavior, school absences, lower educational attainment, and sleep problems (Patton et al. 2016; Colizzi et al. 2020; Finning et al. 2019; Orchard et al. 2020; Zimmermann-Sloutskis et al. 2010; Agnafors et al. 2021). At the same time there has been an increase in the use of smartphones and social media, and these influences have become an important aspect of adolescents' daily lives (Orben 2020). Many research studies have suggested that there is an association between phone usage, social media usage, and adolescent health and well-being (Sanders et al. 2024). However, this association is not as simple as one might expect. In fact, moderate phone and social media use can be advantageous for many health and behavioral outcomes while at higher levels of use, the reverse effect tends to be observed (Sanders et al. 2024; Przybylski et al. 2020; Yan et al. 2017; Viner at al. 2019; Stiglic and Viner 2019; Roser et al. 2016; Carter et al. 2024a; Carter et al. 2024b). There have been some questions about the strength of some of these associations, and it has been suggested that the weakness of the conclusions from these studies may be due the presence of confounding factors such as gender and age (Orben et al. 2022).

There has also been a growing trend for the use of phones to be prohibited in schools, with the United Nations reporting that one in four countries have laws that that prohibit phone use during the school day in public schools (UNESCO 2023). England has provided guidance recommending prohibiting phones in schools but left the issue of prohibition to the school's discretion. Therefore, many schools have devised their own policies that restrict phone use. (Randhawa et al. 2024). Restrictive phone policies are based on the popular assumption that prohibiting phone use in schools will improve mental health and well-being, educational attainment, and reduce problematic use and levels of disruptive behavior (Goodyear 2025; Xia et al. 2025; Wood et al. 2023).

One study sought to investigate whether there is an association between school phone policies, adolescent phone use behaviors, mental health, well-being, and other related outcomes (Goodyear 2025). The study focused on comparing schools that do not permit phone use with schools that do. Three main research objectives were of interest to the researchers. The first objective was to determine whether there was a difference in mental well-being, anxiety and depression, sleep duration, time spent in physical activity, classroom disruptive behavior, attainment, and prevalence of problematic phone and social media usage between schools that allow phone use and schools that do not. The second objective was to determine if there was a difference in phone and social media use and duration of use within the school day, and if there was a difference in motives for phone use between schools that allow phone use and schools that do not. The third objective was to determine if there was an association between phone use and mental well-being, anxiety, depression, sleep duration, time spent in physical activity, classroom disruptive behavior, attainment, and prevalence of problematic use (Goodyear 2025).

The study began by considering over 1,300 secondary schools which were inside a 100-mile radius of the study recruiting center located in Birmingham, England. In this region 1,245 schools had restrictive phone policies, and 96 schools had permissive phone policies. These schools were recruited via email and of the restrictive schools 51 indicated interest in participating in the study while 20 of the permissive schools indicated interest. In the end 20 restrictive schools and 10 permissive schools participated in the study. From these schools 817 students at the restrictive schools and 406 students from the permissive schools provided data for the study. The authors of the study indicated that ethical approval for the study was approved by the University of Birmingham's Science, Technology, Engineering, and Mathematics Research Ethics Committee.

This was a comprehensive study, and many types of data were collected. The observation of interest was the mental well-being of the students as measured using the Warwick–Edinburgh Mental Well-Being Scale, which has been theoretically studied and has been used widely in research projects (Waite and Atkinson 2021). Secondary health and education-related outcomes included measures of anxiety, depression, physical activity, sleep, classroom behavior, academic attainment, and problematic social media use. Observations on smartphone and social media use and motives for social media use were also collected. Finally, demographic variables were collected from each student.

Given the complexity of the study and the data that was collected, the statistical analyses associated with the conclusions from the study were complicated. However, we will give a summary of the results of the study. The researchers found no difference in adolescent mental well-being, anxiety, depression, problematic social media use, sleep, physical activity, attainment, and disruptive behavior between students attending schools that permitted phone use compared to students attending schools that restricted phone use.

However, the observations from the study did indicate that the students who spent longer times on phones and on social media had, on average, lower mental well-being, higher rates of anxiety, depression, disruptive behavior, and problematic social media use, and less sleep and physical activity. Hence, the researchers concluded that the study provides further evidence of the adverse consequences from increased smartphone and social media use. Of particular interest was the fact that while students at restrictive schools did spend less time on their phone during the school day, this did not translate into observable differences in the adverse consequences listed above. The researchers concluded that phone use during the school day probably did not constitute a significant portion of total phone use for the students during the week.

The population of interest in this study is school age adolescents in England or possibly the wider United Kingdom. As a practical matter the researchers could not study this entire population directly. First and foremost, the study was geographically concentrated on an area around the University of Birmingham. It is not clear from the research article why the geographic area was used, but one can imagine that this simplified the process of enumerating and contacting schools for the study. Technically, one can argue that the results of the study should only apply to schools and students in this region, unless one can argue that there is a universality in the observations of these students that goes beyond geographic boundaries and could apply to all schools and students in England or the United Kingdom. Essentially one would need to argue that the social, health, and mental behavior of students in the region around Birmingham should be the same as, for example, the students in the regions around London, Glasgow or Belfast.

The more serious concern with the population of interest for this study is the high percentage of schools that were contacted but chose not to participate in the study. While the study contacted over 1,300 schools, only 30 participated in the study. The issue of the low participation is very closely connected to understanding the population studied. As we discussed above when we considered the geographic region the schools were selected from, one must consider the question of whether those schools that did participate in the study differ in some important way from those schools that did not. If it can be determined that the reason that some schools did not participate is not associated with the social, health, and mental behavior of students at those schools, then it is possible that the population could be considered all the schools from the geographic region. Otherwise, if the reason that the schools did not participate in the study is associated somehow with the social, health, and mental behavior of students at those schools, then the population of interest would be participating schools in the region. Unfortunately, there may be no way of knowing whether such an association exists since no data was directly observed from the schools that did not participate.

The authors of the study addressed the results in terms of the population in their discussion of the applicability of their conclusions by looking at what was known about the schools that did not participate. They conceded that the recruited schools differed in several ways that may impact generalizability of the findings, but concluded that despite some limitations, the results of the study were likely to be applicable across the United Kingdom. Interestingly, the researchers also pointed out that there were also some differences in the permissive and restrictive sample of schools that could cause some confounding problems. However, these differences were specifically accounted for in the statistical analysis. Finally, the researchers noted that there may be other school characteristics that could have an impact on the outcomes, possibly having a confounding effect.

This page titled 4.6.3: Phone Use Policies in Schools in England is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.6.4: Same-Sex Marriage and LGBT Discrimination

In an earlier section we considered a study that explored whether legal advances in marriage equality influence criminal offenses against the LGBTQIA+ community (Nikolaou 2022). This study considered hate crime data against those in the LGBTQIA+ community following the legalization of same-sex marriage in some states. The data on hate crimes used in the study was obtained from the National Incident Based Reporting System of the FBI's Uniform Crime Reporting program, which collects data on criminal offenses committed against individuals and property. This database provides details on whether a criminal offense is motivated by the offender's bias against the victim's race, gender, religion, disability, sexual orientation, or gender identity. The authors note that only criminal offenses with sufficient evidence that the crimes are motivated by bias are reported as hate crimes and are included in the data (Nikolaou 2022).

According to the FBI website, law enforcement agencies representing about 49% of the population in the United States report data to the National Incident Based Reporting System. This database is one of the most complete and accessible crime-related databases (Strom and Smith 2017). It is the basis for many research studies that consider either the effect of crime on communities or the incidence of crime on social groups. It has been suggested that care should be used when attempting to study national trends based on the data (Maxfield 1999; Addington 2006; McCormack et al. 2017) because large urban areas are underrepresented in the database (Finkelhor and Ormrod 2000). Therefore, while the authors represent the results as a general summary of the trends in the United States, the actual population studied corresponds to the law enforcement jurisdictions that report to the National Incident Based Reporting System. This does not mean that the study is not useful, or that the results should be dismissed. For the evaluator of this research, it simply means that one should keep in mind that the data recorded in the database may not fully reflect what is going on in the United States as a whole.

This page titled 4.6.4: Same-Sex Marriage and LGBT Discrimination is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.6.5: Sampling Animals in Populations

Nearly all research animal biologists have, at some point, sampled animals from the wild. These researchers strive to obtain samples of individuals from a given population by employing strategies that minimize the possibility for systematic biases. Some strategies include sampling at varying times during the day and in many different areas. Researchers also take care to minimize selectivity for animals of a particular size, age, or maturity (MacKenzie and Kendall 2002; Askey et al. 2007; Cox et al. 2024). However, biases due to the individual characteristics of each animal can affect the likelihood of being captured or detected in a biological study. This is an important issue because it is crucial for biologists to observe animals from the correct population, otherwise conclusions based on the observed data could be misleading (Biro 2013).

For example, consider a biological study of fish in a local lake. Of interest is to get a good idea of how much the average bass weighs. To perform the study the researchers get a net a drag it across the lake. When the net is pulled up the bass are weighed, and the fish are released. The crucial question is whether the bass collected using this method are taken from the whole population of bass in the lake, or if there is some bias in the collection method. Consider, for example, if the holes in the net were rather large. Then one could easily imagine that smaller bass could easily slip through the net and would tend not to be caught. The results from the study would then be biased, concluding that the average bass in the lake weighs more than the true value.

Research has shown that some animals are more likely to be captured or detected than others in a population due to behavioral differences (Crowcroft and Jeffers 1961; Burnham and Overton 1978; Wilson et al. 1993; Gérard et al. 1994; Palencia et al. 2021). In fact, these behavioral differences may not be dependent on the size, age, or maturity of the animals in question (Sih et al. 2004; Reale et al. 2007; Sih and Bell 2008). Active, bold, or exploratory animals are likely to be captured or detected, though this is not universal (Boon et al. 2008; Biro and Dingemanse 2009; Carter et al, 2012; Garamszegi et al. 2009; Guillette et al. 2009; Gabriel and Black 2010; Harrison 2025; Johnstone et al. 2021; Pellegrini et al. 2010; Wilson et al. 2011).

Bias with respect to animal behavior is a concern for any biologist studying animal physiology or ecology because animal behavior directly affects catchability, or whether an individual is observed, and animal behavior is often associated with the traits of interest (Stamps and Groothuis 2010). This can intoduce bias in the conclusions of studies of animal populations. Animal behavior capture bias can also affect the strength of the associations observed between the traits of interest and the associated animal behavior (Garamszegi et al. 2009; Carter et al. 2012).

One study directly researched animal behavior capture bias in a unique way (Biro 2013). In this study four natural lakes with no fish in them and no natural or human predators were stocked with rainbow trout that had three different growth rates (slow, intermediate, and fast). At the end of the season, fish were sampled from the lakes using experimental nets designed to eliminate bias due to the size of the fish. The researcher predicted that the fast-growing fish, independent of their size, would be more likely to be captured in nets. This prediction is based on research that shows that activity rates, boldness, and growth rates are positively associated in fish and other animals (Stamps 2007; Biro and Stamps 2008; Guenther 2018). Hence, the researcher predicted that the activity and the boldness of the faster growing fish would cause them to be more easily captured.

The data collected from the study indicated that the fast-growing fish were most likely to be captured in all lakes included in the study. In two of the lakes the likelihood of being captured increased with growth rate, while in the remaining lakes, the intermediate-growing fish had the lowest likelihood of being captured. Using statistical techniques, the researchers were able to show that these conclusions held even when the accounting for potential confounding effects. The authors further noted that fish were more likely to be captured in two of the lakes but that the data indicated that the mean growth rates in all the lakes were similar. Therefore, while it was easier to capture fish in some of the lakes, this effect is not confounded with the growth rate of the fish.

This study emphasizes the importance of carefully considering the practical aspects of observing individuals from a population. Even though the observation process was very carefully constructed so that there was no size bias in the fish that were sampled, the resulting observations did not correspond to all the fish in the associated lakes. Instead, the true population was fish that were likely to be caught because of their behavior patterns. As the researchers show, these fish can differ in many important ways from the fish in the entire population.

This page titled 4.6.5: Sampling Animals in Populations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.6.6: Land Cover and Racial Segregation

In the United States extreme heat is responsible for about 20% of natural hazard deaths (Borden and Cutter 2008). Some studies have concluded that there are large racial disparities in heat-related deaths while other studies have not found any difference (Greenberg et al. 1983; Jones et al. 1982; Kaiser et al. 2007; O'Neill et al. 2005; Schwartz 2005; Ramlow and Kuller 1990; Weisskopf et al. 2002; Ellis et al. 1975). The lack of vegetation, particularly trees, in urban areas may contribute to these disparities (Uejio et al. 2011; Hart and Sailor 2009; Oke 1982, 1989). Urban trees are known to be helpful for preserving a healthy urban environment. Some of the benefits include providing shade on hot days, reducing the need for wastewater treatment, reduced air pollution, and reduced noise pollution. In fact, urban tree coverage has been associated with lower overall mortality rates and positive health outcomes (Givoni 1991; Scott et al. 1999; Keim et al. 2006; Hwang et al. 2011; Nowak and Greenfield 2010; Samara and Tsitsoni 2011; Mitchell et al. 2011; Dadvand et al. 2012).

In the United States many studies have found racial and ethnic disparities in the presence of urban trees (Heynen 2006; Landry and Chakraborty 2009; Lowry et al. 2012; Ogneva-Himmelberger et al. 2009; Perkins et al. 2004; Zhang et al. 2008). These studies have tended to focus on single metropolitan areas. A national study from 2013 provided the first nationwide evidence of these trends (Jesdale et al. 2013). The presence of trees and other types of land cover were obtained from the 2001 National Land Cover Dataset (Homer et al. 2004). This data contains information at the census block level. According to the U.S. Census Bureau, census blocks are areas bounded by visible features such as roads, streams, and railroad tracks, and by non-visible boundaries such as property lines and city and county limits. Population data, household income, and home ownership were obtained from the 2000 U.S. census.

One interesting issue in this study is that the presence of tree cover in a census block may be uneven, and a resident of a census block may not live in the area where the trees are located. This effect may be especially true in rural areas, and for that reason the researchers only considered residents of metropolitan areas with census block population densities of more than 2,000 people residing within one square kilometer. The analysis was further restricted by the available data. That is, an area could be used if both tree cover information, demographic, and economic data were available. Therefore, the researchers restricted their analysis to residents for whom block-group level poverty information was available, and who identified as either Hispanic, non-Hispanic white, African American, or Asian (Jesdale et al. 2013).

The question of what population is being studied here is quite interesting. From the onset the basis of the study is a meeting of two major sources of data: the 2001 National Land Cover Dataset and the 2000 U.S. Census. Both databases contain a large amount of information. The National Land Cover Dataset contains land coverage information over the entirety of the United States using data based on information obtained by satellite imaging. The U.S. Census also encompasses the entire United States, so the linking of this data would potentially include the entire population, though the data is aggregated by census blocks. Because of the technical issues described above, the researchers then reduced the scope of the study to densely populated urban areas. The second restriction came from the fact that while the census is meant to encompass the entire United States, extended data is not collected in all areas; hence only the areas that collected the demographic and economic information that was relevant to the study were used. Therefore, from one viewpoint the population is restricted only to densely populated urban areas that were studied by the U.S. Census for extended demographic and economic data. This is a difficult population to define, but there is some help. As will be discussed later in the book, there is a way to choose the census blocks in a way so that they represent what would occur in the entire population of the United States. The methods used by the Census Bureau are selected to ensure this, so we can conclude that the population studied for this research corresponds to densely populated urban areas in the United States.

This page titled 4.6.6: Land Cover and Racial Segregation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.6.7: Tobacco Retailer Density

A substantial amount of research has found associations between the density of tobacco retailers in an area and smoking rates among the adults and youth in those areas (Valiente et al. 2021; Magid et al. 2020). While there are thought to be several reasons for these associations, greater exposure to the tobacco retail environment is considered a major factor (Waddell et al. 2016; Watson et al. 2018). Additionally, there are disparities across neighborhoods, with some studies finding that the density of tobacco retailers increased with the proportion of families living in poverty and with the proportion of African American residents (Kong et al. 2020; Adibe et al. 2019; Chaiton et al. 2013; Rodriguez et al. 2014). Additionally, the Centers for Disease Control and Prevention (CDC) estimate that the prevalence of cigarette smoking is higher among adults living below the poverty level compared with those living at or above the poverty level (CDC 2017). A few studies have found even more interesting interactive associations between the density of tobacco retailers and cigarette smoking rates in that the association may depend on neighborhood characteristics such as poverty (Farley et al. 2019; Cantrell et al. 2015; Golden et al. 2020).

Some additional research examined the association between disparities in the density of tobacco retailers and the disparities in the prevalence of tobacco use among youth and adults (Glasser et al. 2022). The study focused on the state of Ohio, where the current rate of cigarette smoking among adults is higher than the national rate, and almost half of the counties in the state exceed the national average for proportion of residents living below the poverty line (Creamer et al. 2019; Glasser et al. 2022).

To obtain information about cigarette smoking among adults in Ohio, the study used data from the 2017 Ohio Behavioral Risk Factor Surveillance System, an annual random-digit dial telephone survey, as well as the 2019 Ohio Medicaid Assessment Survey, a mail, web, and random-digit dial telephone surveys tracking health care (Ohio Department of Health and Wellness 2015; Berzofsky et al. 2019). Random-digit dialing is a methodology of selecting individuals for a survey that essentially calls random valid phone numbers. This method then can access anyone with a phone number, whether it is publicly available or not. These types of government backed surveys are usually carefully designed and validated so that they reach as much of the population as possible, with adjustments for issues like whether people answer the phone and whether people are willing to take the survey. In the research article, the authors state that the data are “a state-representative sample of adults" (Glasser et al. 2022). This means that the authors of the study consider the population to be all adults in the state of Ohio.

This page titled 4.6.7: Tobacco Retailer Density is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.7: Summary

Populations correspond to the collection of individuals, items, or measurements that are of interest to a researcher. It is very important that populations be well defined and practical for a study to be both practical and have valid conclusions. Deciding on a population to study is usually an iterative process that both refines the population and the associated research hypotheses. Government institutions can be an excellent source of population data due to their vast resources, access to government databases, and access to research professionals. These established collections of data are usually reliable and thoroughly vetted. In discrimination litigation the definition of the population of potential applicants is of great importance. The correct population is usually decided by the presiding judge in such cases, and the population used can have a fundamental effect on the outcome of the corresponding litigation.

This page titled 4.7: Summary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.8: About the Quote

William Edwards Deming (1900–1993) was an American business consultant who excelled in many fields, including industrial engineering, management, and statistics (See Figure 4.8.1). He helped develop the sampling techniques still used by the United States Census Bureau and the Bureau of Labor Statistics. He is also known as the father of the quality assurance movement and was influential in post-WWII Japan. He is credited with revolutionizing manufacturing and industry in Japan.

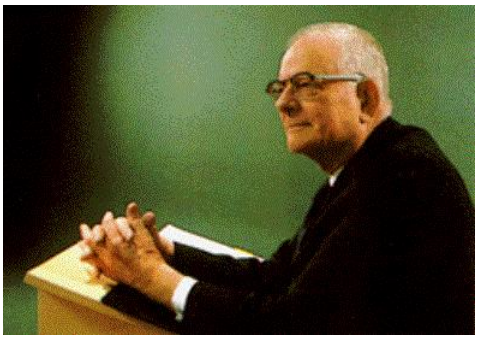


Figure 4.8.1: William Edwards Deming (1900–1993) was an American business consultant who excelled in many fields. (Public Domain Image)

This page titled 4.8: About the Quote is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.9: Practice and Exploration

1. The administration at a small college is interested in exploring how their current students feel about immigration issues, and in particular how gender and race might be associated with certain opinions about immigration issues. They would like to know what percentage of each race and gender combination support immigrant rights on campus. What is the population for this study? What variables do you think the study should include? What are the parameters of interest for this study? What issues of practicality might be of concern when conducting this study? Do you think the corresponding study will have limitations? What will they be?
2. The president of the sociology club at a large university is interested in sponsoring an on-campus movie to help celebrate Black History Month. During a previous meeting of the club, three movies were selected as possible candidates for the event. The president of the club would like the event to be a success and is therefore worried about which movie to pick. The club decides to use a small survey to try to determine which movie would have the largest audience. Describe several populations the club might consider studying. How accessible are these populations? Do you think these populations could be studied from a practical viewpoint? Which of these populations do you think the club would be most interested in?
3. Suppose that you are planning a research project that would investigate the association between the demographics of neighborhoods and their corresponding economic conditions. What established online databases maintained by the United States federal government do you think might be helpful with this research project? What types of data do you think you might be able to find in these databases? Can these databases be accessed by the public or is special access required?
4. A small business has ten employees and is interested in their opinion of their working conditions. In this business there are two places where employees can interact with customers, either inside the office or at a kiosk outside the office. The business is worried that weather conditions make working outside in the kiosk less appealing to the employees. The following data is obtained from talking with each of the employees.

Employee	Inside	Outside
1	6	5
2	10	7
3	10	6
4	10	8
5	6	2
6	2	1
7	2	3
8	5	2
9	10	4
10	9	9

   In the data each worker rated, on a scale from 1 to 10, the working conditions inside and outside the business. For this rating scale a higher number indicates better working conditions. Of interest is the average rating for working outside subtracted from the average rating for working inside. Identify the variables in this population along with the parameter of interest. What is the value of the parameter? What does the value of the parameter say about the working conditions inside compared with the working conditions outside?
5. Consider a court case where the plaintiffs allege that individuals who, subsequent to becoming employed at a company, identified themselves as having a nonbinary gender identity have not been subsequently promoted to higher positions in the company at the same rate as other employees. What would you consider the correct population to be for this type of litigation? Explain why you identified this population.

6. Suppose that you are working for a friend who is running for election to a position in the local county government, which is in a rural county with a population of about 10,000. Your friend is a young, dynamic individual who has a keen interest in helping the young population in the area. Because of your interest in data, you are put in charge of studying the political affiliations and viewpoints of young people. Describe several populations you might consider studying and how the views of the corresponding populations might be of interest to your friend. How accessible are these populations? Do you think these populations could be studied from a practical viewpoint? Which of these populations do you think your friend would be most interested in?
7. Consider a court case where an employer is accused of having a racial bias in their hiring practices for entry level positions that require a high school level education. The employer is a regional company that extends through several counties of their home state. What would you consider the correct population to be for this type of litigation? Explain why you identified this population.
8. Many people collect things like vinyl records, baseball cards, stuffed animals, sneakers, and the like. If you have such a collection, think about that collection as a population. If you do not collect anything, think about a collection a friend or family member might have. Suppose that you wanted to summarize what is in the collection without having to describe each item separately. What would be the variables that one might be interested in and why? What population characteristics might one be interested in and why?
9. An employer advertises for a position at a loading dock where applicants are required to be able to stack boxes onto a truck of a certain height, a requirement that is mentioned in the employment advertising. The employer is later sued by a group that states that the employer has biased hiring practices. In the past five years no applicants who have identified as female have been hired. In the case, the employer argues that the physical requirement for the position is the reason that no female applicants have been hired. If you were the lawyer for the plaintiffs, what population would you identify as being the correct relevant population of potential employees for this position? Why would you choose this population?
10. While searching for medical information on the dangers of vapor-based tobacco products, you come across a study that describes the use of vape products across different age groups, race and ethnic groups, and gender identities. The research is from a well-respected group of academic researchers. What type of government resources would have been helpful to the researchers in gathering data of this subject? What problems might they have encountered while trying to find their data and identify the relevant population? What limitations do you think the study might have?
11. A city has recently invested in city-run youth centers—places where young people can go and play sports, play games, study, and relax away from home during non-school days. The founders of the program would like to assess if there is an association between the presence of these centers and the crime rates in the surrounding neighborhoods. Determine several ways that the population of interest can be identified for this problem. What type of government resources could have been helpful to the researchers in gathering data on this subject? What problems might they encounter while trying to find their data and identify the relevant population? What limitations do you think the study might have?
12. Using an online news service, find a news article about an issue important to you that is based on an empirical analysis. Identify, as clearly as you can, the population that the authors of the news article have used to observe their data and gather information. Are the data they obtained based on information from government sources and databases? Are there stated limitations to the population in the new article? What are they? Are there other limitations in identifying the population that are not stated in the news article but which you think might present a problem? Why? Based on these answers, how reliable and universal do you think the conclusions of the news article are? Can you think of any ways to improve the research?
13. Using an online source or your library, find an academic research article about an issue that is important to you. Identify, as clearly as you can, the population that the authors of the article have used to observe their data and gather information. Are the data they obtained based on information from government sources and databases? Are there stated limitations to the population in the article? What are they? Are there other limitations in identifying the population not stated in the article but which you think might present a problem? Why? Based on these answers, how reliable and universal do you think that the conclusions of the article are? Can you think of any ways to improve the research?
14. Using an online source or your library, find information on a legal case that dealt with employment discrimination. What event occurred that gave rise to the case? If possible, determine the population used to decide the merits of the case. What was the result?

This page titled 4.9: Practice and Exploration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.10: Collaborative Activity

The presidential election of 1936 is not the only example of a very public mishap of presidential polling. Probably an even more famous example occurred a dozen years later with the election of 1948. This election was between the incumbent Harry Truman and New York governor Thomas Dewey. Three polls covered the election: Crossley, Gallup, and Roper. As election day neared, all three polls declared that Dewey would win the election. The predicted percentage of the popular vote for each of the four candidates is given in the table below (Freedman et al. 2007; Mosteller et al. 1949).

	Predictions		
Candidate	Crossley	Gallup	Roper
Truman	45	44	38
Dewey	50	50	53
Thurmond	2	2	5
Wallace	3	4	4

When election day arrived, Truman received about 50% of the vote while Dewey received only about 45% of the vote. In fact, this error created one of the most memorable photographs in history.

Figure 4.5 shows Harry Truman, the winner of the election, holding a copy of the *Chicago Daily Tribune*, whose headline declared that "Dewey Defeats Truman." On election night the press deadline for the paper required the issue of the *Tribune* to go to press before all the states had reported most of the results. The paper's veteran political analyst Arthur Sears Henning and the political polls were almost unanimous that Dewey would win by a landslide, and therefore the *Tribune* went to press with the now famous banner headline (Wendt 1979).

So once again, the polls were wrong, and in this collaborative activity you and your group will discuss the possible reasons for this error by considering doing a poll of your own.

When we discussed the presidential election of 1936, the problem with the survey methods used by the *Literary Digest* was the fact that the sample of voters was taken from a population that was different from those who voted in the election.

The three survey organizations had all learned the lesson from the 1936 election and attempted to insure that for this election they would sample from the correct population. The method used by all three survey organizations was based on quota sampling. Each interviewer was given very specific demographic information about the people they should interview. This demographic information included categories such as gender, race, and economic status. For example, an interviewer in a certain city might be told to interview seven men. Of these men, three should be under 40 years old, while the remaining four should be over 40 years old. Moreover, one of the men should be African American while the remaining six should be white (Freedman et al. 2007; Stephan and MaCarthy 1958). If the individuals the interviewer spoke with followed the demographic requirements set by the polling agency, the interviewer could speak with whomever they pleased.

1. Suppose you and the members of your group have been hired by the university to conduct a survey about the students' opinions of the meal choices on campus. You are to interview twenty students according to the following demographics:
    1. Five students should be first-year students, four should be second-year, four should be third-year, three should be fourth-year, and four should be graduate students.
    2. Four of the first-year students should live in the dorms, while the remaining first-year student should live off campus.
    3. Two of the second-year students should live in the dorms, while the remaining second-year students should live off campus.
    4. One of the third-year students should live in the dorms, while the remaining third-year students should live off campus.
    5. One of the fourth-year students should live in the dorms, while the remaining fourth-year students should live off campus.
2. Working with the people in your group, come up with a plan for how the group could find these students and interview them.
3. Considering the plan that your group came up with, are there some problems that you might encounter when implementing the plan? Would some of the students be difficult to find? How could you go about finding these students?

3. Are there some biases that might be inherent in your group plan to find students to interview? Are there parts of the student population that might be difficult to reach? What type of bias could this introduce into the data that you collect?
4. What possible problems with confounding could result from the method your group came up with to find students to interview? Would these confounding problems be important to the conclusions of the study?
5. Now think about the presidential election polls of 1948. Using what you learned from considering the problem of sampling students on campus according using quota sampling, what possible types of bias could have been introduced using quota sampling for the 1948 presidential election poll? Are there members of the voting population who might be difficult to reach? What important confounding problems are there that might have been important to the conclusions of the polls?
6. Keeping the context of this collaborative activity and your answers to the above questions in mind, what is your general opinion of quota sampling? Do you believe that it is a reliable method for conducting polling? What types of problems would one need to consider when implementing this type of strategy?

This page titled 4.10: Collaborative Activity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

4.11: References

Addington, Lynn A. 2006. “Using National Incident-Based Reporting System Murder Data to Evaluate Clearance Predictors: A Research Note.” *Homicide Studies* 10, no. 2, 140–152. https://doi.org/10.1177/1088767905285439

Adibe, Chiche, P. Craigmile, Nathaniel Onnen, Elli Schwartz, and M. Roberts. 2019. “The Relationship Between Tobacco Retailer Density and Neighborhood Demographics in Ohio.” *Ohio Journal of Public Health* 2, no. 1, 12–18. https://doi.org/10.18061/ojph.v2i1.9036

Agnafors, S., M. Barmark, and G. Sydsjö. 2021. “Mental Health and Academic Performance: A Study on Selection and Causation Effects from Childhood to Early Adulthood.” *Social Psychiatry and Psychiatric Epidemiology* 56, 857–866. https://doi.org/10.1007/s00127-020-01934-5

Askey, P. J., J. R. Post, E. A. Parkinson, E. Rivot, A. J. Paul, and P. A. Biro. 2007. “Estimation of Gillnet Efficiency and Selectivity Across Multiple Sampling Units: A Hierarchical Bayesian Analysis Using Mark-Recapture Data.” *Fisheries Research* 83, 162–174. https://doi.org/10.1016/j.fishres.2006.09.009

Berzofsky, M. E., T. Duffy, and K. Johnson. 2019. *Ohio Medicaid Assessment Survey: Methodology Report*. https://medicaid.ohio.gov/resources-.../methods-specs

Biro, P. A. 2013. “Are Most Samples of Animals Systematically Biased? Consistent Individual Trait Differences Bias Samples Despite Random Sampling.” *Oecologia* 171, 339–345. https://doi.org/10.1007/s00442-012-2426-5

Biro, P. A., and N. J. Dingemanse. 2009. “Sampling Bias Resulting from Animal Personality.” *Trends in Ecology & Evolution* 24, no. 2, 66–68. https://doi.org/10.1016/j.tree.2008.11.001

Biro, P. A., and J. A. Stamps. 2008. “Are Animal Personality Traits Linked to Life-History Productivity?” *Trends in Ecology & Evolution* 23, no. 7, 361–368. https://doi.org/10.1016/j.tree.2008.04.003

Blakemore, S. J. 2019. “Adolescence and Mental Health.” *Lancet* 393, no. 10185, 2030–2031. https://doi.org/10.1016/S0140-6736(19)31013-X

Boardman, Anthony E. 1979. “Another Analysis of the EEOC ‘Four-Fifths’ Rule.” *Management Science* 25, no. 8, 770–776. https://doi.org/10.1287/mnsc.25.8.770

Bobko, Philip and Philip L. Roth. 2004. “The Four-Fifths Rule for Assessing Adverse Impact: An Arithmetic, Intuitive, and Logical Analysis of the Rule and Implications for Future Research and Practice.” In *Research in Personnel and Human Resources Management*. Emerald Group Publishing Limited. https://doi.org/10.1016/S0742-7301(04)23004-3

Boon, A., D. Reale, and S. Boutin. 2008. “Personality, Habitat Use, and Their Consequences for Survival in North American Red Squirrels Tamiasciurius hudsonicus.” *Oikos* 117, no. 9, 1321–1328. https://doi.org/10.1111/j.0030-1299.2008.16567.x

Borden, Kevin A., and Susan L. Cutter. 2008. “Spatial Patterns of Natural Hazards Mortality in the United States.” *International Journal of Health Geographics* 7, no. 1, 1–13. https://doi.org/10.1186/1476-072X-7-64

Burnham, K. P., and W. S. Overton. 1978. “Estimation of the Size of a Closed Population When Capture Probabilities Vary Among Animals.” *Biometrika* 65, no. 3, 625–633. https://doi.org/10.1093/biomet/65.3.625

Cantrell, Jennifer, Andrew Anesetti-Rothermel, Jennifer L. Pearson, Haijun Xiao, Donna Vallone, and Thomas R. Kirchner. 2015. “The Impact of the Tobacco Retail Outlet Environment on Adult Cessation and Differences by Neighborhood Poverty.” *Addiction* 110, no. 1, 152–161. https://doi.org/10.1111/add.12718

Carter, Alecia J., Robert Heinsohn, Anne W. Goldizen, and Peter A. Biro. 2012. “Boldness, Trappability and Sampling Bias in Wild Lizards.” *Animal Behaviour* 83, no. 4, 1051–1058. https://doi.org/10.1016/j.anbehav.2012.01.033

Carter, Ben, Najma Ahmed, Olivia Cassidy, Oliver Pearson, Marilia Calcia, Clare Mackie, and Nicola Jayne Kalk. 2024a. “There’s More to Life Than Staring at a Small Screen: A Mixed Methods Cohort Study of Problematic Smartphone Use and the Relationship to Anxiety, Depression and Sleep in Students Aged 13–16 Years Old in the UK.” *BMJ Mental Health* 27, e301115. https://doi.org/10.1136/bmjment-2024-301115

Carter, B., M. Payne, P. Rees, S. Y. Sohn, J. Brown, N. J. Kalk. 2024b. “A Multi-School Study in England, to Assess Problematic Smartphone Usage and Anxiety and Depression.” *Acta Paediatrica, International Journal of Paediatrics* 113, no. 10, 2240–2248. https://doi.org/10.1111/apa.17317

Centers for Disease Control and Prevention. 2017. *Smoking & Tobacco Use: Cigarette Smoking and Tobacco Use Among People of Low Socioeconomic Status*. https://www.cdc.gov/tobacco-health-e...th-burden.html

Chaiton, Michael O., Graham C. Mecredy, Joanna E. Cohen, and Melodie L. Tilson. 2013. “Tobacco Retail Outlets and Vulnerable Populations in Ontario, Canada.” *International Journal of Environmental Research and Public Health* 10, no. 12, 7299–7309. https://doi.org/10.3390/ijerph10127299

Colizzi, M., A. Lasalvia, and M. Ruggeri. 2020. “Prevention and Early Intervention in Youth Mental Health: Is it Time for a Multidisciplinary and Trans-Diagnostic Model for Care?” *International Journal of Mental Health Systems* 14, no. 23. https://doi.org/10.1186/s13033-020-00356-9

Cox, K., S. Neyrinck, and J. Mergeay. 2024. “Dealing with Assumptions and Sampling Bias in the Estimation of Effective Population Size: A Case Study in an Amphibian Population.” *Evolutionary Applications* 17, no. 9, e70015. https://doi.org/10.1111/eva.70015

Creamer, MeLisa R., Teresa W. Wang, Stephen Babb, et al. Karen A. Cullen, Hannah Day, Gordon Willis, Ahmed Jamal, and Linda Neff. 2019. “Tobacco Product Use and Cessation Indicators Among Adults—United States, 2018.” *Morbidity and Mortality Weekly Report* 68, no. 45, 1013. https://doi.org/10.15585/mmwr.mm6845a2

Crowcroft, P., and J. N. R. Jeffers. 1961. “Variability in the Behaviour of Wild House Mice (Mus Musculus L.) Towards Live Traps.” *Proceedings of the Zoological Society of London* 137, no. 4, 573–582. https://doi.org/10.1111/j.1469-7998.1961.tb06090.x

Dadvand, Payam, Jordi Sunyer, Xavier Basagana, Ferran Ballester, Aitana Lertxundi, Ana Fernandez-Somoano, Marisa Estarlich, Raquel Garcia-Esteban, Michelle A. Mendez, and Mark J. Nieuwenhuijsen. 2012. “Surrounding Greenness and Pregnancy Outcomes in Four Spanish Birth Cohorts.” *Environmental Health Perspectives* 120, no. 10, 1481–1487. https://doi.org/10.1289/ehp.1205244

Ellis, F. P., F. Nelson, and L. Pincus. 1975. “Mortality During Heat Waves in New York City July, 1972 and August and September, 1973.” *Environmental Research* 10, no. 1, 1–13. https://doi.org/10.1016/0013-9351(75)90069-9

Farley, Shannon M., Andrew R. Maroko, Shakira F. Suglia, and Lorna E. Thorpe. 2019. “The Influence of Tobacco Retailer Density and Poverty on Tobacco Use in a Densely Populated Urban Environment.” *Public Health Reports* 134, no. 2, 164–171. https://doi.org/10.1177/0033354918824330

Finkelhor, David, and Richard Ormrod. 2000. *Characteristics of Crimes Against Juveniles*. U.S. Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency. https://www.ojp.gov/pdffiles1/ojjdp/179034.pdf

Finning, K., O. C. Ukoumunne, T. Ford, E. Danielsson-Waters, L. Shaw, I. R. De Jager, L. Stentiford, and D. A. Moore. 2019. “The Association Between Child and Adolescent Depression and Poor Attendance at School: A Systematic Review and Meta-Analysis.” *Journal of Affective Disorders* 245, 928–938. https://doi.org/10.1016/j.jad.2018.11.055

Freedman, David, Robert Pisani, and Roger Purves. 2007. *Statistics*. Norton. https://wwnorton.com/books/Statistics

Gabriel, P. O., and J. M. Black. 2010. “Behavioural Syndromes in Steller’s Jays: The Role of Time Frames in the Assessment of Behavioural Traits.” *Animal Behaviour* 80, no. 4, 689–697. https://doi.org/10.1016/j.anbehav.2010.07.004

Garamszegi, L. Z., M. Eens, and J. Török. 2009. “Behavioural Syndromes and Trappability in Free-Living Collared Flycatchers, Ficedula Albicollis.” *Animal Behaviour* 77, no. 4, 803–812. https://doi.org/10.1016/j.anbehav.2008.12.012

Gastwirth, Joseph L., and Weiwen Miao. 2009. “Formal Statistical Analysis of the Data in Disparate Impact Cases Provides Sounder Inferences than the Us Government’s ‘Four-Fifths’ Rule: An Examination of the Statistical Evidence in Ricci v. DeStefano.” *Law, Probability and Risk* 8, no. 2, 171–191. https://doi.org/10.1093/lpr/mgp017

Gérard, D., V. Bauchau, and S. Smets. 1994. “Reduced Trappability in Wild Mice, Mus Musculus Domesticus, Heterozygous for Robertsonian Translocations.” *Animal Behaviour* 47, no. 4, 877–883. https://doi.org/10.1006/anbe.1994.1119

Givoni, Baruch. 1991. “Impact of Planted Areas on Urban Environmental Quality: A Review.” *Atmospheric Environment. Part B. Urban Atmosphere* 25, no. 3, 289–299. https://doi.org/10.1016/0957-1272(91)90001-U

Glasser, Allison M., Nathaniel Onnen, Peter F. Craigmile, Elli Schwartz, and Megan E. Roberts. 2022. “Associations Between Disparities in Tobacco Retailer Density and Disparities in Tobacco Use.” *Preventive Medicine* 154, 106910. https://doi.org/10.1016/j.ypmed.2021.106910

Golden, Shelley D., Tzy-Mey Kuo, Amanda Y. Kong, Christopher D. Baggett, Lisa Henriksen, and Kurt M. Ribisl. 2020. “County-Level Associations Between Tobacco Retailer Density and Smoking Prevalence in the USA, 2012.” *Preventive Medicine Reports* 17, 101005. https://doi.org/10.1016/j.pmedr.2019.101005

Goodyear, Victoria A., Amie Randhawa, Péymane Adab, Hareth Al-Janabi, Sally Fenton, Kirsty Jones, Maria Michail, Breanna Morrison, Paul Patterson, Jonathan Quinlan, Alice Sitch, Rebecca Twardochleb, Matthew Wade, and Miranda Pallan. 2025. "School Phone Policies and Their Association with Mental Wellbeing, Phone Use, and Social Media Use (SMART Schools): A Cross-Sectional Observational Study." *The Lancet Regional Health–Europe* 51, 101211. https://doi.org/10.1016/j.lanepe.2025.101211

Greenberg, Jerome H., Jill Bromberg, Christine M. Reed, Tracy L. Gustafson, and Richard A. Beauchamp. 1983. “The Epidemiology of Heat-Related Deaths, Texas—1950, 1970–79, and 1980.” *American Journal of Public Health* 73, no. 7, 805–807. https://doi.org/10.2105/ajph.73.7.805

Guenther, Anja. 2018. “Life-History Trade-Offs: Are They Linked to Personality in a Precocial Mammal (Cavia Aperea)?” *Biology Letters* 14, 20180086. https://doi.org/10.1098/rsbl.2018.0086

Guillette, L. M., A. R. Reddon, P. L. Hurd, and C. B. Sturdy. 2009. “Exploration of a Novel Space is Associated with Individual Differences in Learning Speed in Black-Capped Chickadees, Poecile Atricapillus.” *Behavioural Process* 82, no. 3, 265–270. https://doi.org/10.1016/j.beproc.2009.07.005

Harrison, Natasha D. 2025. “The Bold and the Biased: The Influence of Behaviour on Capture Probability in Woylies (Bettongia Penicillata Ogilbyi).” *Australian Mammalogy* 47, no. 3, AM25028. https://doi.org/10.1071/AM25028

Hart, Melissa A., and David J. Sailor. 2009. “Quantifying the Influence of Land-Use and Surface Characteristics on Spatial Variability in the Urban Heat Island.” *Theoretical and Applied Climatology* 95, no. 3, 397–406. https://doi.org/10.1007/s00704-008-0017-5

Hauenstein, Neil M. A., Jaron T. Holmes, and Emilee B. Tison. 2013. “Detecting Adverse Impact: The Four-Fifths Rule Versus Significance Testing.” *Public Personnel Management* 42, no. 3, 403–420. https://doi.org/10.1177/0091026013495762

Heynen, Nik. 2006. “Green Urban Political Ecologies: Toward a Better Understanding of Inner-City Environmental Change.” *Environment and Planning A* 38, no. 3, 499–516. https://doi.org/10.1068/a37365

Homer, Collin, Chengquan Huang, Limin Yang, Bruce Wylie, and Michael Coan. 2004. “Development of a 2001 National Land-Cover Database for the United States.” *Photogrammetric Engineering & Remote Sensing* 70, no. 7, 829–840. https://doi.org/10.14358/PERS.70.7.829

Hwang, Hee-Jae, Se-Jin Yook, and Kang-Ho Ahn. 2011. “Experimental Investigation of Submicron and Ultrafine Soot Particle Removal by Tree Leaves.” *Atmospheric Environment* 45, no. 38, 6987–6994. https://doi.org/10.1016/j.atmosenv.2011.09.019

Ismail, Rashid and Brian H. Kleiner. 2001. “The Use of the Four-Fifths Rule in Discrimination Cases.” *Managerial Law* 43, no. 1–2, 57–61. https://doi.org/10.1108/03090550110770327

Jesdale, Bill M., Rachel Morello-Frosch, and Lara Cushing. 2013. “The Racial/Ethnic Distribution of Heat Risk–Related Land Cover in Relation to Residential Segregation.” *Environmental Health Perspectives* 121, no. 7, 811–817. https://doi.org/10.1289/ehp.1205919

Johnstone, Kyla Chloe, Clare McArthur, and Peter Bruce Banks. 2021. “Catch Me if You Can: Personality Drives Technique-Specific Biases During Live-Capture Trapping.” *Wildlife Research* 48, no. 8, 713–721. https://doi.org/10.1071/WR20121

Jones, T. Stephen, Arthur P. Liang, Edwin M. Kilbourne, Marie R. Griffin, Peter A. Patriarca, Steven G. Fite Wassilak, Robert J. Mullan, Robert F. Herrick, H. Denny Donnell, Keewhan Choi, and Stephen B. Thacker. 1982. “Morbidity and Mortality Associated

with the July 1980 Heat Wave in St. Louis and Kansas City, MO." *Journal of the American Medical Association* 247, no. 24, 3327–3331. https://doi.org/10.1001/jama.1982.03320490025030

Kahneman, Daniel, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press. https://www.google.com/books/edition...sec=frontcover

Kaiser, Reinhard, Alain Le Tertre, Joel Schwartz, Carol A. Gotway, W. Randolph Daley, and Carol H. Rubin. 2007. "The Effect of the 1995 Heat Wave in Chicago on All-Cause and Cause-Specific Mortality." *American Journal of Public Health* 97, Supplement 1, S158–S162. https://doi.org/10.2105/AJPH.2006.100081

Keim, R. F., A. E. Skaugset, and M. Weiler. 2006. "Storage of Water on Vegetation Under Simulated Rainfall of Varying Intensity." *Advances in Water Resources* 29, no. 7, 974–986. https://doi.org/10.1016/j.advwatres.2005.07.017

Kong, Amanda Y., Allison E. Myers, Lisa F. Isgett, and Kurt M. Ribisl. 2020. "Neighborhood Racial, Ethnic, and Income Disparities in Accessibility to Multiple Tobacco Retailers: Mecklenburg County, North Carolina, 2015." *Preventive Medicine Reports* 17, 101031. https://doi.org/10.1016/j.pmedr.2019.101031

Landry, Shawn M., and Jayajit Chakraborty. 2009. "Street Trees and Equity: Evaluating the Spatial Distribution of an Urban Amenity." *Environment and Planning A* 41, no. 11, 2651–2670. https://doi.org/10.1068/a41236

Lowry, John H., Matthew E. Baker, and R. Douglas Ramsey. 2012. "Determinants of Urban Tree Canopy in Residential Neighborhoods: Household Characteristics, Urban Form, and the Geophysical Landscape." *Urban Ecosystems* 15, no. 1, 247–266. https://doi.org/10.1007/s11252-011-0185-4

MacKenzie, D. I., and W. L. Kendall. 2002. "How Should Detection Probability be Incorporated into Estimates of Relative Abundance?" *Ecology* 83, 2387–2393. https://doi.org/10.1890/0012-9658(20...SDPBI]2.0.CO;2

Magid, Hoda S. Abdel, Bonnie Halpern-Felsher, Pamela M. Ling, Patrick T. Bradshaw, Mahasin S. Mujahid, and Lisa Henriksen. 2020. "Tobacco Retail Density and Initiation of Alternative Tobacco Product Use Among Teens." *Journal of Adolescent Health* 66, no. 4, 423–430. https://doi.org/10.1016/j.jadohealth.2019.09.004

Maxfield, Michael G. 1999. "The National Incident-Based Reporting System: Research and Policy Applications." *Journal of Quantitative Criminology* 15, no. 2, 119–149. https://doi.org/10.1023/A:1007518620521

McCormack, Philip D., April Pattavina, and Paul E. Tracy. 2017. "Assessing the Coverage and Representativeness of the National Incident-Based Reporting System." *Crime and Delinquency* 63, no. 4, 493–516. https://doi.org/10.1177/0011128717694595

Mitchell, Richard, Thomas Astell-Burt, and Elizabeth A. Richardson. 2011. "A Comparison of Green Space Indicators for Epidemiological Research." *Journal of Epidemiology and Community Health* 65, no. 10, 853–858. https://doi.org/10.1136/jech.2010.119172

Mosteller, Frederick, Herbert Hyman, Phillip J. McCarthy, Eli S. Marks, David B. Truman, Leonard W. Doob, Duncan MacRae, Jr., Frederick F. Stephan, Samuel A. Stouffer, and S. S. Wilks. 1949. *The Pre-Election Polls of 1948, Report to the Committee on Analysis of Pre- Election Polls and Forecasts*. Social Science Research Council. https://www.google.com/books/edition...HGQePsoC?hl=en

Myers, Randy. 2017. "`I Accept That I Have Nobody': Young Women, Youth Justice, and Expectations of Responsibility During Reentry." *Social Justice* 44, no. 1, 62–82. https://socialjusticejournal.org/pro...e-of-security/

Nikolaou, Dimitrios. 2022. "Same-Sex Marriage Laws, LGBT Hate Crimes, and Employment Discrimination Charges." *Southern Economic Journal* 88, 869–905. https://doi.org/10.1002/soej.12548

Nowak, David J., and Eric J. Greenfield. 2010. "Evaluating the National Land Cover Database Tree Canopy and Impervious Cover Estimates Across the Conterminous United States: A Comparison with Photo-Interpreted Estimates." *Environmental Management* 46, no. 3, 378–390. https://doi.org/10.1007/s00267-010-9536-9

O'Neill, Marie S., Antonella Zanobetti, and Joel Schwartz. 2005. "Disparities by Race in Heat-Related Mortality in Four US Cities: The Role of Air Conditioning Prevalence." *Journal of Urban Health* 82, no. 2, 191–197. https://doi.org/10.1093/jurban/jti043

Ogneva-Himmelberger, Yelena, Hamil Pearsall, and Rahul Rakshit. 2009. "Concrete Evidence & Geographically Weighted Regression: A Regional Analysis of Wealth and the Land Cover in Massachusetts." *Applied Geography* 29, no. 4, 478–487.

https://doi.org/10.1016/j.apgeog.2009.03.001

Ohio Department of Health and Wellness. 2015. *Ohio 2013 BRFSS Annual Report, 2015*. https://odh.ohio.gov/wps/wcm/connect...0a7249-mthHqXb

Oke, Timothy R. 1982. "The Energetic Basis of the Urban Heat Island." *Quarterly Journal of the Royal Meteorological Society* 108, no. 455, 1–24. https://doi.org/10.1002/qj.49710845502

Oke, Timothy R. 1989. "The Micrometeorology of the Urban Forest." *Philosophical Transactions of the Royal Society of London B, Biological Sciences* 324, no. 1223, 335–349. https://doi.org/10.1098/rstb.1989.0051

Orben, A. 2020. "Teenagers, Screens and Social Media: A Narrative Review of Reviews and Key Studies." *Social Psychiatry and Psychiatric Epidemiology* 55, no. 4, 407–414. https://doi.org/10.1007/s00127-019-01825-4

Orben, A., A. K. Przybylski, S. J. Blakemore, and R. A. Kievit. 2022. "Windows of Developmental Sensitivity to Social Media." *Nature Communications* 13, no. 1, 1649. https://doi.org/10.1038/s41467-022-29296-3

Orchard, F., A. M. Gregory, M. Gradisar, and S. Reynolds. 2020. "Self-Reported Sleep Patterns and Quality Amongst Adolescents: Cross-Sectional and Prospective Associations with Anxiety and Depression." *Journal of Child Psychology and Psychiatry* 61, no. 10, 1126–1137. https://doi.org/10.1111/jcpp.13288

Palencia, P., J. M. Rowcliffe, J. Vicente, and P. Acevedo. 2021. "Assessing the Camera Trap Methodologies Used to Estimate Density of Unmarked Populations." *Journal of Applied Ecology* 58, no. 8, 1583–1592. https://doi.org/10.1111/1365-2664.13913

Patton, G. C., S. M. Sawyer, J. S. Santelli, David A. Ross, Rima Afifi, Nicholas B. Allen, Monika Arora, Peter Azzopardi, Wendy Baldwin, Christopher Bonell, Ritsuko Kakuma, Elissa Kennedy, Jaqueline Mahon, Terry McGovern, Ali H Mokdad, Vikram Patel, Suzanne Petroni, Nicola Reavley, Kikelomo Taiwoaa, Jane Waldfogel, Dakshitha Wickremarathne, Carmen Barroso, Zulfiqar Bhutta, Adesegun O. Fatusi, Amitabh Mattoo, Judith Diers, Jing Fang, Jane Ferguson, Frederick Ssewamala, Russell M. Viner. 2016. "Our Future: A Lancet Commission on Adolescent Health and Wellbeing." *Lancet* 387, no. 10036, 2423–2478. https://doi.org/10.1016/S0140-6736(16)00579-1

Pellegrini, A., B. Wisenden, and P. Sorensen. 2010. "Bold Minnows Consistently Approach Danger in the Field and Lab in Response to Either Chemical or Visual Indicators of Predation Risk." *Behavioral Ecology and Sociobiology* 64, 381–387. https://doi.org/10.1007/s00265-009-0854-y

Perkins, Harold A., Nik Heynen, and Joe Wilson. 2004. "Inequitable Access to Urban Reforestation: The Impact of Urban Political Economy on Housing Tenure and Urban Forests." *Cities* 21, no. 4, 291–299. https://doi.org/10.1016/j.cities.2004.04.002

Polansky, Alan M. 2011. *Introduction to Statistical Limit Theory*. Chapman and Hall/CRC. https://www.routledge.com/Introducti...erKFD-3iTGYYj_

Przybylski, A. K., A. Orben, and N. Weinstein. 2020. "How Much Is Too Much? Examining the Relationship Between Digital Screen Engagement and Psychosocial Functioning in a Confirmatory Cohort Study." *Journal of the American Academy of Child & Adolescent Psychiatry* 59, no. 9, 1080–1088. https://doi.org/10.1016/j.jaac.2019.06.017

Ramlow, Jonathan M. and Lewis H. Kuller. 1990. "Effects of the Summer Heat Wave of 1988 on Daily Mortality in Allegheny County, PA." *Public Health Reports* 105, no. 3, 283–289. https://pmc.ncbi.nlm.nih.gov/articles/PMC1579995/

Randhawa, Amie, Miranda Pallan, Rebecca Twardochleb, Peymane Adab, Hareth Al-Janabi, Sally Fenton, Kirsty Jones, Maria Michail, Paul Patterson, Alice Sitch, Matthew Wade, and Victoria A. Goodyear. 2024. "Secondary School Smartphone Policies in England: A Descriptive Analysis of How Schools Rationalize, Design, and Implement Restrictive and Permissive Phone Policies." *Journal of Research on Technology in Education*, 1–20. https://doi.org/10.1080/15391523.2024.2363204

Reale, D., S. M. Reader, D. Sol, P. T. McDougall, and N. J. Dingemanse. 2007. "Integrating Animal Temperament Within Ecology and Evolution." *Biological Reviews* 82, No. 2, 291–318. https://doi.org/10.1111/j.1469-185X.2007.00010.x

Rodriguez, Daniel, Heather A. Carlos, Anna M. Adachi-Mejia, Ethan M. Berke, and James Sargent. 2014. "Retail Tobacco Exposure: Using Geographic Analysis to Identify Areas with Excessively High Retail Density." *Nicotine & Tobacco Research* 16, no. 2, 155–165. https://doi.org/10.1093/ntr/ntt126

Roser, K., A. Schoeni, M. Foerster, M. Röösli. 2016. “Problematic Mobile Phone Use of Swiss Adolescents: Is it Linked with Mental Health or Behaviour?” *International Journal of Public Health* 61, no. 3, 307–315. https://doi.org/10.1007/s00038-015-0751-2

Samara, Theano and Thekla Tsitsoni. 2011. “The Effects of Vegetation on Reducing Traffic Noise from a City Ring Road.” *Noise Control Engineering Journal* 59, no. 1, 68–74. https://doi.org/10.3397/1.3528970

Sanders, T., M. Noetel, P. Parker, Borja Del Pozo Cruz, Stuart Biddle, Rimante Ronto, Ryan Hulteen, Rhiannon Parker, George Thomas, Katrien De Cocker, Jo Salmon, Kylie Hesketh, Nicole Weeks, Hugh Arnott, Emma Devine, Roberta Vasconcellos, Rebecca Pagano, Jamie Sherson, James Conigrave, and Chris Lonsdale. 2024. “An Umbrella Review of the Benefits and Risks Associated with Youths’ Interactions with Electronic Screens.” *Nature Human Behavior* 8, no. 1, 82–99. https://doi.org/10.1038/s41562-023-01712-8

Schwartz, Joel. 2005. “Who is Sensitive to Extremes of Temperature?: A Case-Only Analysis.” *Epidemiology* 16, no. 1, 67–72. https://doi.org/10.1097/01.ede.0000147114.25957.71

Sih, A. and A. M. Bell. 2008. “Insights for Behavioral Ecology from Behavioral Syndromes.” *Advances in the Study of Behavior* 38, 227–281. https://doi.org/10.1016/S0065-3454(08)00005-3

Sih, A., A. M. Bell, J. C. Johnson, and R. E. Ziemba. 2004. “Behavioural Syndromes: An Integrative Overview.” *The Quarterly Review of Biology* 79, no. 3, 241–277. https://doi.org/10.1086/422893

Shoben, Elaine. 1986. “Employment Discrimination Cases.” In *Statistical Methods in Discrimination Litigation*, D.H. Kaye, Mickel Aickin, editors. CRC Press. https://www.routledge.com/Statistica...yxNcDTd5WSJBnE

Scott, Klaus I., James R. Simpson, and E. Gregory McPherson. 1999. “Effects of Tree Cover on Parking Lot Microclimate and Vehicle Emissions.” *Journal of Arboriculture* 25, no. 3, 129–142. https://doi.org/10.48044/jauf.1999.019

Sobol, Marion Gross, and Charles J. Ellard. 1988. “Measures of Employment Discrimination: A Statistical Alternative to the Four-Fifths Rule.” *Industrial Relations Law Journal* 10, no. 3, 381–399. https://www.jstor.org/stable/24050044

Stamps, J. A. 2007. “Growth-Mortality Tradeoffs and ‘Personality’ Traits in Animals.” *Ecology Letters* 10, no. 5, 355–363. https://doi.org/10.1111/j.1461-0248.2007.01034.x

Stamps, J. A. and T. G. G. Groothuis. 2010. “The Development of Animal Personality: Relevance, Concepts and Perspectives.” *Biological Reviews* 85, no. 2, 301–325. https://doi.org/10.1111/j.1469-185X.2009.00103.x

Stephan, F. F., and P. J. McCarthy. 1958. *Sampling Opinions: An Analysis of Survey Procedure*. John Wiley. https://www.google.com/books/edition...wAEACAAJ?hl=en

Stiglic, N. and R. M. Viner. 2019. “Effects of Screentime on the Health and Well-Being of Children and Adolescents: A Systematic Review of Reviews.” *BMJ Open* 9, no. 1, e023191. https://doi.org/10.1136/bmjopen-2018-023191

Strom, Kevin J., and Erica L. Smith. 2017. “The Future of Crime Data: The Case for the National Incident-Based Reporting System (NIBRS) as a Primary Data Source for Policy Evaluation and Crime Analysis.” *Criminology and Public Policy* 16, no. 4, 1027–1048. https://doi.org/10.1111/1745-9133.12336

Uejio, Christopher K., Olga V. Wilhelmi, Jay S. Golden, David M. Mills, Sam P. Gulino, and Jason P. Samenow. 2011. “Intra-Urban Societal Vulnerability to Extreme Heat: The Role of Heat Exposure and the Built Environment, Socioeconomics, and Neighborhood Stability.” *Health & Place* 17, no. 2, 498–507. https://doi.org/10.1016/j.healthplace.2010.12.005

UNESCO. 2023. *Global Education Monitoring Report 2023: Technology in Education– a Tool on Whose Terms*? https://doi.org/10.54676/UZQV8501

Valiente, Roberto, Francisco Escobar, María Urtasun, Manuel Franco, Niamh K. Shortt, and Xisca Sureda. 2021. “Tobacco Retail Environment and Smoking: A Systematic Review of Geographic Exposure Measures and Implications for Future Studies.” *Nicotine and Tobacco Research* 23, no. 8, 1263–1273. https://doi.org/10.1093/ntr/ntaa223

Viner, Russell M., Aswathikutty Gireesh, Neza Stiglic, et al. Lee D. Hudson, Anne-Lise Goddings, Joseph L. Ward, Dasha E. Nicholls. 2019. “Roles of Cyberbullying, Sleep, and Physical Activity in Mediating the Effects of Social Media Use on Mental

Health and Wellbeing Among Young People in England: A Secondary Analysis of Longitudinal Data." *The Lancet Child & Adolescent Health* 3, no. 10, 685–696. https://doi.org/10.1016/S2352-4642(19)30186-5

Waddell, Elizabeth Needham, Rachel Sacks, Shannon M. Farley, and Michael Johns. 2016. "Point-of-Sale Tobacco Marketing to Youth in New York State." *Journal of Adolescent Health* 59, no. 3, 365–367. https://doi.org/10.1016/j.jadohealth.2016.05.013

Waite, M., and C. Atkinson. 2021. "A Review of Mental Health and Wellbeing Measures for Universal Secondary School Screening." *Emotional and Behavioural Difficulties* 26, no. 2, 206–222. https://doi.org/10.1080/13632752.2021.1939997

Watson, Kimberly A., Doris G. Gammon, Brett R. Loomis, Harlan R. Juster, and Elizabeth Anker. 2018. "Trends in Cigarette Advertising, Price-Reducing Promotions, and Policy Compliance in New York State Licensed Tobacco Retailers, 2004 to 2015." *American Journal of Health Promotion* 32, no. 8, 1679–1687. https://doi.org/10.1177/0890117118764852

Weisskopf, Marc G., Henry A. Anderson, Seth Foldy, Lawrence P. Hanrahan, Kathleen Blair, Thomas J. Török, and Peter D. Rumm. 2002. "Heat Wave Morbidity and Mortality, Milwaukee, Wis, 1999 vs 1995: An Improved Response?" *American Journal of Public Health* 92, no. 5, 830–833. https://doi.org/10.2105/ajph.92.5.830

Wendt, Lloyd. 1979. *Chicago Tribune: The Rise of a Great American Newspaper*. Rand McNally. https://archive.org/details/chicagotribuneri00wend

Wilson, D. S., K. Coleman, A. B. Clark, and L. Biederman. 1993. "Shy–Bold Continuum in Pumpkinseed Sunfish (Lepomis Gibbosus): An Ecological Study of a Psychological Trait." *Journal of Comparative Psychology* 107, 250–260. https://doi.org/10.1037/0735-7036.107.3.250

Wilson, A. D. M., T. R. Binder, K. P. McGrath, S. J. Cooke, and J.-G. J. Godin. 2011. "Capture Technique and Fish Personality: Angling Targets Timid Bluegill Sunfish, Lepomis Macrochirus." *Canadian Journal of Fisheries and Aquatic Sciences* 68, no. 5, 749–757. https://doi.org/10.1139/f2011-019

Wood, Grace, Victoria Goodyear, Peymane Adab, Hareth Al-Janabi, Sally Fenton, Kirsty Jones, Maria Michail, Breanna Morrison, Paul Patterson, Alice J. Sitch, Matthew Wade, and Miranda Pallan. 2023. "Smartphones, Social Media and Adolescent Mental Well-Being: The Impact of School Policies Restricting Daytime Use-Protocol for a Natural Experimental Observational Study using Mixed Methods at Secondary Schools in England (SMART Schools Study)." *BMJ Open* 13, no. 7, e075832. https://doi.org/10.1136/bmjopen-2023-075832

Xiao, Y., Y. Meng, T. T. Brown, K. M. Keyes, and J. J. Mann. 2025. "Addictive Screen Use Trajectories and Suicidal Behaviors, Suicidal Ideation, and Mental Health in US Youths." *Journal of the American Medical Association* 334, no. 3, 219–228. DOI: 10.1001/jama.2025.7829

Yan, Hanyi, Rui Zhang, Theresa M. Oniffrey, Guoxun Chen, Yueqiao Wang, Yingru Wu, Xinge Zhang, Quan Wang, Lu Ma, Rui Li, and Justin B. Moore. 2017. "Associations Among Screen Time and Unhealthy Behaviors, Academic Performance, and Well-Being in Chinese Adolescents." *International Journal of Environmental Research and Public Health* 14, no. 6, 596. https://doi.org/10.3390/ijerph14060596

Zhang, Yangjian, Michael A. Tarrant, and Gary T. Green. 2008. "The Importance of Differentiating Urban and Rural Phenomena in Examining the Unequal Distribution of Locally Desirable Land." *Journal of Environmental Management* 88, no. 4, 1314–1319. https://doi.org/10.1016/j.jenvman.2007.07.008

Zimmermann-Sloutskis, D., M. Wanner, E. Zimmermann, and B. W. Martin. 2010. "Physical Activity Levels and Determinants of Change in Young Adults: A Longitudinal Panel Study." *International Journal of Behavioral Nutritional and Physical Activity* 7, no. 2, 1–13. https://doi.org/10.1186/1479-5868-7-2

This page titled 4.11: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

CHAPTER OVERVIEW

5: The Nature of Data

“Any time scientists disagree, it's because we have insufficient data. Then we can agree on what kind of data to get; we get the data; and the data solves the problem. Either I'm right, or you're right, or we're both wrong. And we move on.”

–Neil deGrasse Tyson

5.1: What to Expect
5.2: Quantitative and Qualitative Data
5.3: The Structure of Data
5.4: Critical Analysis
5.4.1: Inequities in Employment During COVID-19
5.4.2: Attitudes About Marital Rape
5.4.3: Satisfaction with Sport Participation
5.4.4: The Supreme Court and Personal Attitudes
5.4.5: Product Placement in Movies
5.4.6: The Racial Wealth Gap
5.5: Summary
5.6: About the Quotation
5.7: Practice and Exploration
5.8: Collaborative Activity
5.9: References

This page titled 5: The Nature of Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.1: What to Expect

Considering the use of empiricism in the modern scientific method, data from observed variables is one of the most important parts of any study. Data comes in many different forms, including numerical data of several types, nonnumerical data that can be categorized, and nonnumerical data that has a more abstract form. In this chapter, we will define and offer examples of several types of numerical data and categorical data. By the end of this chapter, you will be able to:

1. Classify data as either quantitative or qualitative.
2. Determine if qualitative data can be represented as categorical data.
3. Identify what types of mathematical calculations are valid for each measurement scale.
4. Determine if a variable has been measured using a nominal, ordinal, interval, or ratio measurement scale.

This page titled 5.1: What to Expect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.2: Quantitative and Qualitative Data

When a population has been identified, and members or items from the population have been observed, there are characteristics or measurements on these items that researchers are interested in for the purpose of studying an issue. This is the data from which we hope to learn valuable information about the issue. Data can come in many different forms, including oral histories, photographs, demographics, scientific measurements, and visual graphs. A study of the conditions of African American families in the post-Civil War era may collect observations in the form of oral or written family histories, photographs, and demographic data. A study of economic conditions, race, and gender may consider demographic data such as income, race, consumer debt, loan default rates, and credit ratings. A study of health access and gender identity may use demographics, measures of health, and data on the frequency of doctor visits. In the pharmaceutical industry, studies of the efficacy of new drug treatments often include observations on patients' health histories, demographic information, health indicators, measures of the effectiveness of the treatment, and observed side-effects. A study of a new fertilizer may include information about soil nutrients, growth rates, and crop yields.

It is important to note that a single study may have all types of data, and it is up to the researchers to understand how to bring the information contained in the data together in a way that provides reliable answers to the important questions about the issues they hope to study. The forms of the data have a great bearing on what methods can be used to extract this information. Therefore, it is very important that researchers correctly classify the type of data that has been observed. It is also important for those who read and critically evaluate this research to be able to think about the form of the data that was observed and to be able to critically evaluate what decisions the researchers made about how information was extracted from the data.

The first major way to classify data into different types concerns whether the data can be represented in a natural numeric form, or whether the form of the data has a more abstract form that does not have a natural numeric form. Data that have a natural numeric form, that is, data that can be represented easily and uniquely as numbers or values, is known as quantitative data.

Definition: Quantitative Data

Data observed from a variable that have a natural and unique numeric value are called *quantitative data*.

Some types of data are based on measurements. These types of data are almost always quantitative. They include variables corresponding to scientific measurements such as height, weight, temperature, length, and time. For each of these variables there is a natural system of measurements that provides unique information about the observed values. For example, if someone states that a lecture lasted 62 minutes, this measurement means the same thing to each person who might observe the variable. Similarly, if a length is reported as 1.77 meters, this measurement means the same to everyone without any additional information.

Some data may be represented numerically, but the representation may not be unique. For example, consider a study where a variable corresponding to relationship status is observed. The actual observed values may be categorized as single, cohabitation, married, divorced, or separated. For simplicity, a researcher may represent this data numerically (see example in Table 5.1). Now suppose you are walking down the street, and someone asks you your relationship status. If you respond that your relationship status is 3, will the other person automatically know how to interpret that data? Unless they have access to the additional information contained in Table 5.1, they will not know what the data means. In fact, another researcher doing a study may represent the same categories in a completely different way, as shown in Table 5.2. Note that one of these representations is not better than the other. One is not more correct than the other. Both representations are equally valid, but the actual data values are completely different.

Table 5.1. A possible numerical representation of relationship status.

Category	Representation
Single	1
Cohabitation	2
Married	3
Divorced	4
Separated	5

Table 5.2. A different possible numerical representation of relationship status.

Category	Representation
Single	0
Cohabitation	1
Married	2
Divorced	3
Separated	4

The example based on marital status demonstrates why the qualifier *unique* is given in the definition of *quantitative data*. In the example of the length of a lecture, an observation of 62 minutes means the same to everyone. Someone may report the length alternatively as 3,720 seconds, or 1.03 hours, but the meaning is the same as there is a mathematical method for converting between the units of measurement. All these measurements represent the same idea. For the relationship status data, there is not a unique representation. You must have access to Tables 5.1 and 5.2 to decode the numerical data representations of the relationship statuses.

These examples highlight the difference between data that can be naturally and uniquely represented numerically, that is quantitative data, and data that do not have a natural unique numerical representation, which are called qualitative data.

Definition: Qualitative Data

Data observed from a variable that does not have a natural and unique numeric value are called *qualitative data*.

The names of the two types of data are very close to one another which can cause some confusion. Looking at the leading parts of each word can be helpful in remembering the two types of data. The word quantitative starts with *quant* which is close to the word quantity, which can be thought of as a numerical value. Similarly, the word qualitative starts with *qual* which is close to the word quality, which can be thought of as a non-numerical quality of an individual or item.

Relationship status is one example of a qualitative variable. Others include ethnicity, gender, housing type, and employment status. Data of this type can often be represented as categories. It is important to realize that it is sometimes difficult to determine whether data is quantitative or qualitative. Further, how the data is represented may change how the variable is categorized. For example, suppose that a study observes the type of vehicles that pass through a toll booth on a highway. It would be quite natural to categorize data using categories like motorcycle, car, truck, van, mini-van, and SUV. In this case the data are qualitative as there is no natural numbering system. However, the researchers might be really interested in the toll collected from each vehicle, which is a function of the vehicle type. If the researchers instead track the collected toll, then the variable would be quantitative.

Conversely, a variable like age might seem to be automatically quantitative since it is a length of time. But some studies might categorize age using broad categories such as Young (0-12), Young Adult (13-21), Adult (22-64), and Senior (65 and over). Even though the categories have some numerical structure to them in that we know the people in the “Young” category have an age

smaller than those in the “Adult” category, there are no natural numbers that would be associated with the categories. If we observed someone in the study and we just know that they are “Young”, what number do we assign to their age? There is not a logical answer unless we have more information.

Some information and observations are even more abstract and have less formal framework than those examples given above. A study of gender attitudes in the early part of the 20th century may have data that consist solely of diary entries of young women during the time. This type of quantitative research may rely on experts reading the diaries and offering opinions on gender attitudes based on these readings. However, it is also quite plausible for this type of data to be transformed into numerical data as well. For example, one could count the number of times certain key words such as “subservient” or “duty” appear per page in the diaries. In this case the original data would be qualitative, and quite abstract, but the transformed variables would be quantitative.

An example of such a qualitative study was based on interviews and observations at two juvenile justice centers for girls. The purpose of the study was to explore how the state failed to meet the material needs of young women on their release from the system, their future expectations about state support, and their sense of personal responsibility (Myers, 2017). The data gathered for the study was based on interviews with inmates in the system, and the conclusions from the study were based on expert sociological opinions of the content of the interviews.

This page titled 5.2: Quantitative and Qualitative Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.3: The Structure of Data

Numerical data, whether it originates from quantitative observations, or is being used to represent qualitative observations, is further classified by the amount of inherent structure there is in the data. Some data, such as physical measurements like length and mass, have a great deal of structure. For example, salaries have a great deal of structure. If we compare the salaries of two individuals, we can not only determine who has a higher salary, but we can also compute by how much one salary exceeds the other. Furthermore, we can also make comparisons based on fractions; for example, that one salary is twice the other.

Other measurements may have a little less structure. Temperature is a good example. We can certainly compare two temperatures and determine which temperature is hotter or colder, and we can determine how much one temperature is greater or less than or another. But we cannot compare temperatures using fractions. For example, it is not correct to say that 100 degrees Fahrenheit is twice the temperature of 50 degrees Fahrenheit. To see why this is true, convert both temperatures to degrees Celsius: 100 degrees Fahrenheit is about 38 degrees Celsius while 50 degrees Fahrenheit is about 10 degrees Celsius. But $100 \div 50 = 2$ and $38 \div 10 = 3.8$. It follows that when temperature is measured in Celsius, the same higher temperature is 3.8 times that of the same lower temperature. Hence, it does not make sense to say that "today is twice as hot as yesterday" because the comparison would be different if someone measured the temperature using different units.

Still, other numerical data has even less structure. We might decide to code data about age to be based on age ranges such as is given in Table 5.3. With this coding we can compare ages in that we can determine if one person is older than another. However, it does not make sense to say that an individual whose age is coded as 3 is one older than an individual whose age is coded as 2. Further, if two individuals have the same coded age then they could be the same age, one could be older than another, or vice-versa.

Table 5.3. An example of coding data for age ranges in years.

Range	Code
0-18	1
19-21	2
22-65	3
65+	4

We get numerical data with even less structure if we consider the codes for relationship status given in Table 5.1. In this case we can only determine if two individuals have the same relationship status. It does not make sense to say that cohabitation is greater than single because the corresponding code for cohabitation is greater than the code for being single. We could have just as easily reversed these codes.

The amount of structure in numerical data is classified by what type of numerical calculations make sense for the data. The data with the least amount of structure are called nominal data.

Definition: Nominal Measurement Scale

A variable has a *nominal measurement scale* if the only possible valid mathematical comparison of two data points is whether they are equal.

Nominal data usually arises from numerically coding qualitative data that is categorized, classified, named, or labeled. For nominal data the underlying categories do not have any mathematical structure, as with the example given by the coding in Table 5.1. The key idea is that the categories cannot even be rank ordered by which is greater. Hence, variables with numeric coding for marital status, gender identity, hair color, political affiliation, or religious affiliation are all examples of variables with a nominal scale of measurement. Nominal variables may also refer to non-numeric data represented by categories or names that cannot be ranked or sorted.

If data can be at most ranked or sorted, then the data has an ordinal measurement scale.

Definition: Ordinal Measurment Scale

A variable has an *ordinal measurement scale* if the only possible valid mathematical comparisons of two data points is whether they are equal and whether one is greater than another.

While observations from a variable with an ordinal measurement scale can be ranked or sorted, it is not valid to look at the differences between the numerical coding for the categories. The age range coding given in Table 5.3 are an example of a variable with an ordinal measurement scale. From the coding given in this table we can decide if two individuals are in the same age range or if one individual is in an older age range than the other. We cannot, however, determine how many years older one person may be than another by looking at the coded data. Social class is another example of a variable that can have an ordinal measurement scale. For example, social class may be coded as shown in Table 5.4. The ranking of these classes may reflect the perceived prestige of the corresponding classes by individuals in a population. Looking at the data we can conclude if two individuals are in the same social class, or if one is in a social class that is perceived as higher than another. But it is not valid to take the numerical difference between the class coding. As with nominal data, ordinal variables may also refer to non-numeric data represented by categories or names that can be ranked and sorted.

Table 5.4 An example of coding social classes.

Class	Code
Working Class	1
Middle Class	2
Upper Class	3

Ordinal data is very common in surveys. For example, a study that sought to determine if racial micro aggressions were influencing the health of Asian Americans asked everyone in the study: “How would you rate your overall health in the past-year?” The responses were reported using the categories excellent, very good, good, fair, and poor. This is qualitative categorical data on an ordinal scale (Nicholson and Mei, 2020). Another very common type of survey question in surveys gives a statement and asks the respondent to gauge how much they agree with the statement, usually on a scale with categories “Strongly Agree”, “Agree”, “Neither Agree nor Disagree”, “Disagree”, and “Strongly Disagree”, which is qualitative categorical data on an ordinal scale. This latter example is often called a Likert Scale (Joshi et al., 2015).

Ordinal data is often coded using a numerical system. For example, assigning the value 5 to “Strongly Agree”, 4 to “Agree”, 3 to “Neither Agree nor Disagree”, 2 to “Disagree”, and 1 to “Strongly Disagree”. These values are then averaged over the questions on a survey to get an average level of agreement with the statements in the survey. An example of this type of method will be studied later in this chapter. It should be noted that one should be cautious with studies that do these types of calculations with ordinal data in that, while this may allow researchers to compare the level of agreement between two individuals or the average level of agreement over two groups of individuals, the actual values themselves have little meaning. Further, using these types of calculations must be justified by the researchers before the results can be trusted. These types of justifications are addressed later in the book.

If data can be sorted or ranked, and the differences between the data points provide a valid mathematical comparison, then the data is said to have an interval measurement scale.

Definition: Interval Measurement Scale

A variable has an *interval measurement scale* if the possible valid mathematical comparisons of two data points are whether they are equal, or whether one is greater than another. Additionally, a difference between two datum is a valid comparison of how much one datum exceeds another.

While observations from a variable with an interval measurement scale can be ranked or sorted, and the differences between the values provide a valid measurement, ratios are still not valid. Measurements that can be both positive and negative are commonly on the interval measurement scale, with the temperature measurement discussed in Section 5.2 providing an example. Another

example of an interval measurement is the FICO credit score used in the United States which has a range between 300 and 850. Credit worthiness, as measured by this score, can of course be ranked. That is, one individual is considered to have better credit if their score is higher than another individual's score. It is valid to look at differences between credit scores. For example, it is very common for so called credit repair companies to advertise that they can increase your credit score by 100 points. However, it is not valid to state that a credit score can be doubled. For example, doubling a credit score of 500 would give a score of 1000 which is not even possible.

If data can be sorted or ranked, the differences between the data points provide a valid mathematical comparison, and ratios are also valid mathematical comparisons, then the data is said to have a ratio measurement scale.

Definition: Ratio Measurement Scale

A variable has a *ratio measurement scale* if the possible valid mathematical comparisons of two data points are whether they are equal or whether one is greater than the other. Additionally, differences and ratios between two data are a valid comparison.

While observations from a variable with an interval measurement scale can be ranked or sorted, and the differences between the values provide a valid measurement, ratios are still not valid. Hence, any of the usual arithmetic calculations that are valid for two numbers are also valid for ratio data. Many measurements are on the ratio scale, particularly if they are non-negative. For example, age is a ratio variable. Assuming that it is measured in years, then a person who is 30 years old is twice the age of someone who is 15 years old. Other examples of variables that have a ratio measurement scale include height, weight, time to graduation, and income.

As can be observed from the definitions given above, each successive definition adds more structure to the data over the last. This is visualized in Figure 5.3.1. As the definitions indicate, nominal data can only be compared to determine if two values are equal. Ordinal data has the structure of nominal data, but with ordinal data one can also determine if one value is larger than another. Interval data has the structure of ordinal data, but additionally allows one to determine how much one value exceeds another. Finally, ratio data has the structure of interval data, but additionally allows one to compare values through ratios.

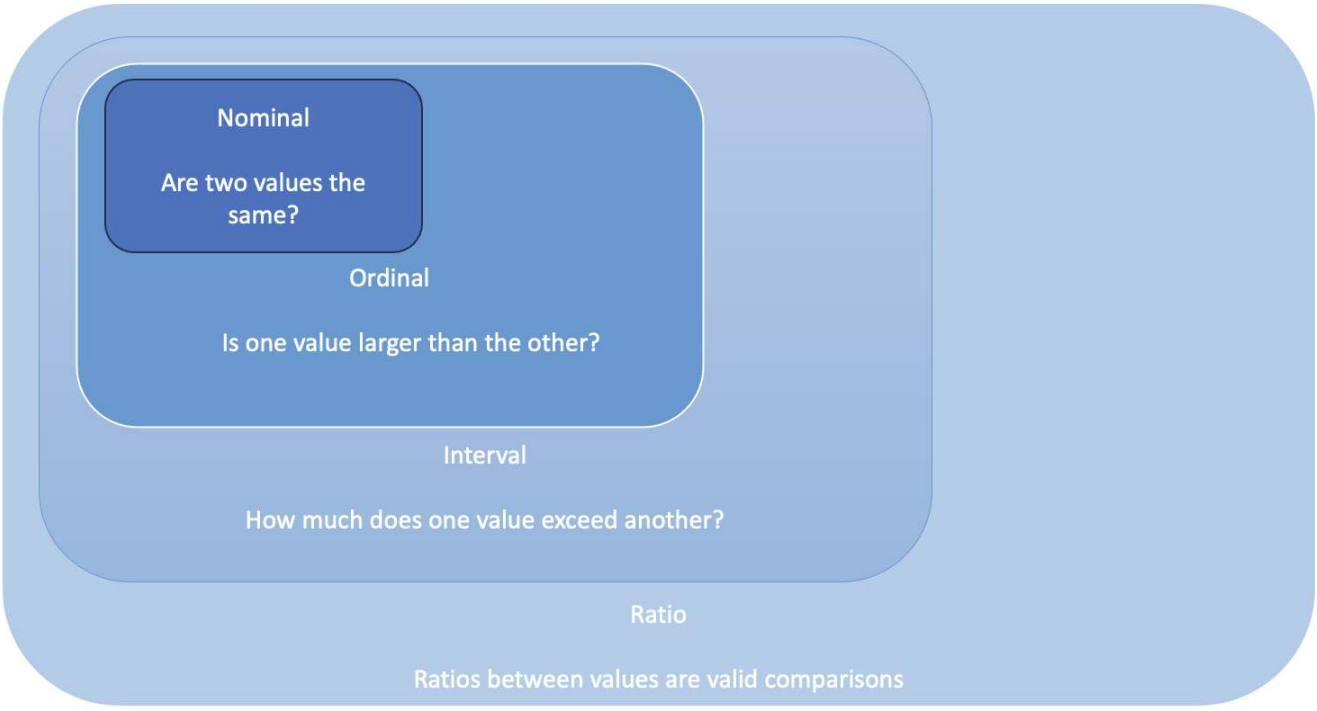


Figure 5.3.1: The relationship between the different types of measurement scales. (Public domian image created by Alan M. Polansky)

This page titled 5.3: The Structure of Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.4: Critical Analysis

When critically analyzing a research study, a look at the type and structure of the data is an important indicator of what type of analysis will follow. Data that has a special structure often needs to be analyzed using specialized statistical techniques. The important issue is whether a statistical technique assumes that the data has more structure than it really does. For example, many statistical procedures must assume that the underlying data has at least as much structure as a variable with an interval measurement scale. If the data is ordinal, or worse yet, nominal, then the corresponding statistical analysis will not be valid. Deciding whether a particular data summary, or statistical analysis, is valid for a particular variable can be complicated, but this section will provide you with some basic issues to look for when reading a study.

Many statistical techniques will be unfamiliar to you when you read a study, and you may not have enough knowledge to specifically know when a particular statistical technique is valid for certain variables. The idea at this phase of a critical analysis is not to possibly refute or confirm the validity of a study on this basis but make note of some possible issues that may need to be investigated further.

In summary, the key questions in critically analyzing the structure of the data in a study are:

1. What are the variables used in the study and what is the measurement scale of each variable?
2. What types of calculations are involved with each of the variables?
3. Are the calculations appropriate for each variable?
4. Are the calculations validated by previous studies?

This page titled 5.4: Critical Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.4.1: Inequities in Employment During COVID-19

It has long been known that people of color have historically experienced discrimination in hiring and employment and that these same populations have different access to fair salaries, paid sick leave, and unemployment benefits (Yearby and Mohapatra 2020; Quillian et al. 2017; McMurtry et al. 2019; Tessler et al. 2020; Bleich et al. 2019; Findling et al. 2019). A recent study considered these inequalities in the context of the COVID-19 pandemic in the United States (Gemelas et al. 2022). Of particular interest was whether people of color experienced higher levels of employment loss in frontline and non-frontline occupations during the onset of the pandemic.

The data for the study was based on the Current Population Survey of the Bureau of Labor Statistics. This survey is based on a representative sample of 60,000 households in the United States. The data used for the study were based on two time points, January and June of 2020 (Gemelas et al. 2022). The observed variables included the occupation of the individual, their racial or ethnic identity, and whether the individual was employed or not. The occupation variable reflects the job position of the individual, grouped using broad categories such as *Healthcare Practitioners and Support* and *Protective Service*. The categories were then grouped as either being *frontline*; that is, an occupation that is considered critical and requires in person interaction, or *non-frontline*. This is a qualitative variable with a nominal measurement scale. The racial or ethnic identity used the categories Black or African American, Asian American, White, Hispanic or Latinx, and Non-Hispanic or Latinx. This is also a qualitative variable with a nominal measurement scale. Finally, the question as to whether someone was employed or not is also a qualitative variable on a nominal measurement scale.

The analysis of the data collected in this study is mostly based on calculations with the number of working employed in several employment sectors. The percentages of individuals unemployed within each race category and whether the occupation is considered frontline or non-frontline was computed. Table 5.5 is adapted from a small portion of Table 1 in the research article showing the first two frontline occupation categories (Gemelas et al. 2022). The calculations are typical of those contained in the original research article. The data in this table are based on counting the number of observations in each category of a nominal variable. This is a permissible calculation with this type of data because we only need to identify whether each observation of the nominal variable is in the category of interest. Once these counts are obtained then it is permissible to compute the difference between the counts and the corresponding change in percent. All the remaining calculations shown in this paper are based on percent changes in the number of workers within specified categories of race, year, and whether the job was a frontline job or not. All these types of calculations are permissible with these types of counts.

Table 5.5. The number of healthcare practitioners and protective service workers employed in the first two quarters of 2020, adapted from Table 1 (Gemelas et al. 2022).

Occupation	First Quarter Employed	Second Quarter Employed	Change	Percent Change
Healthcare Practitioners	23,258,585	21,754,243	-1,504,342	-6.47
Protective Service	2,033,623	2,024,401	-9,222	-0.45

This page titled 5.4.1: Inequities in Employment During COVID-19 is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.4.2: Attitudes About Marital Rape

Previously we considered a study that used a survey of undergraduate students and alumni of a small liberal arts college. The goal of the survey was to learn what these groups thought about marital rape (Ferro et al. 2008). Recall that the survey consisted of short stories that described a rape scene that varied only the victim–offender relationship where some randomly selected participants had a vignette that described the offender as a neighbor, while the remaining participants had a story that described the offender as their spouse. The short narratives were rated by the participants using the Illinois Rape Myth Acceptance Scale which is designed to assess the participants' acceptance of rape myths.

The concept of the rape myth is a complex set of cultural beliefs that support and perpetuate male sexual violence against women and deliberately obscure the true nature of rape (Payne et al. 1999). These myths include the idea that rape is impossible without the consent of the victim, that women "ask for rape," and that rape is a result of uncontrollable male passions brought on by actions of the victim (Schwendinger and Schwendinger 1974; Brownmiller 1993).

The Illinois Rape Myth Acceptance Scale is a measure that assesses how much an individual accepts rape myths. The assessment has three subcategories which assesses what conditions must be met for participants to refer to the situation as rape, the level of blame that can be attributed to the offender, and beliefs about the deviant nature of rape and rapists. The measurements are taken using a questionnaire where each question uses scale ranging from 1 (not at all agree) to 7 (very much agree) with hypothetical situations. When the questionnaire is finished, the final score can be computed by taking the sum or the average of the responses, either across all of the questions, or of the questions corresponding to each of the subcategories.

Because these scores are based on a Likert-type scale, the collected data and the associated sums or averages have an ordinal measurement scale. As we saw earlier in the chapter, one should be wary of studies that do these types of calculations without further justification. In this case the data is based on a well-established methodology that has been specifically designed to measure how much an individual accepts rape myths. That is, the validity of using this survey was specifically studied by the researchers who developed the method (Payne et al. 1999). We will study the issues that are relevant to this type of research later, but for the moment we will trust that this methodology has been thoroughly vetted and is acceptable as standard practice by the research community.

This page titled 5.4.2: Attitudes About Marital Rape is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.4.3: Satisfaction with Sport Participation

The satisfaction that one gets when participating in sporting activities, whether as an individual or as part of a team, is an important part of good mental and physical well-being. However, some studies suggest that the rate of disorders such as depression and anxiety might be similar or even higher in athletes compared to nonathletes (Gouttebarge et al. 2019; Gorczynski et al. 2017; Moesch et al. 2018, Moore et al. 2025). One study sought to investigate how success, personality, emotional competence, and sport type (team or individual) are associated with sports satisfaction (Lepir and Lakić 2025).

The study was conducted in Bosnia and Herzegovina. Athletes were contacted and surveyed through various organizations such as clubs, associations, and official competitions. Athletes were also contacted through high school and collegiate educational institutions. The survey was implemented with informed consent and the participants in the study were guaranteed anonymity. A total of 277 athletes participated in the study, of which 41% were female. The participants were an average age of 21.6 years old, with the youngest and oldest athletes being 15 and 39 years old, respectively. The participating athletes represented 16 different sports, including six team sports and ten individual sports. Participants were required to have a minimum of two years of active involvement in their sport, and regular participation in competitions at least at the regional level. The athletes were categorized into three success levels based on their competition level over the past year.

No gender differences were observed in terms of sports satisfaction. The researchers indicated that this conclusion is consistent with findings from previous studies from the 20 years preceding their study and gave several examples of previous research (Baudin et al. 2011; Bebetsos and Theodorakis 2003; Polman et al. 2007). One interesting result from the study was that athletes competing at the lowest and highest

levels expressed greater satisfaction than those competing at the middle level. The researchers hypothesized that athletes at the lowest level were either at the beginning of their careers or were those who participated in the sport as a recreational activity. In either case, these participants could have been under less pressure to perform. In contrast, those at the highest level have established themselves as successful athletes. Participants at the middle level may have found it difficult to reach that level and are finding it difficult to transition to the highest level, which may contribute to lower levels of satisfaction.

The type of sport played is also associated with differences in satisfaction, though the forms of the associations were complicated by the fact that they interacted with some of the other factors studied. Because of this complexity, we will not discuss these findings here. Regarding individual personality traits, the researchers found that conscientiousness and agreeableness were strongly associated with satisfaction, while extraversion (outgoingness associated with individuals who enjoy social interaction and external stimulation) and neuroticism (anxiety and self-doubt) had weaker associations with satisfaction. The researchers reported surprising results concerning openness. For individual athletes, there was a positive association between openness and satisfaction, meaning that more open individuals tended to have higher levels of satisfaction while for team athletes, the association was the opposite. Researchers also found that the ability of an individual to perceive and understand emotions was very weakly associated with sport satisfaction, while the ability to express and label emotions had a stronger association with sports satisfaction. The associations differed for the ability of an individual to manage and regulate emotions, with a strong association observed for individual athletes and a weak association for team athletes.

To measure the abstract ideas contained in this study such as satisfaction, personality traits, and emotional skills and competencies, the researchers used established measures developed by previous research. To measure an individual's satisfaction with their participation in their sport, the Athletic Satisfaction Questionnaire was used (Riemer and Chelladurai 1998). This questionnaire measures sport satisfaction using fifteen ideas about what constitutes satisfaction in sports. The researchers found that some of these ideas were not relevant for the individuals in the study, and therefore the researchers used eight of these ideas, including such concepts as strategy, personal dedication, and individual performance. Each of these concepts was associated with specific questions in the survey. Additionally, the researchers developed two new concepts that measured support from family and friends and support from sport club management. Combining the relevant questions created a survey of 34 questions that measured 10 concepts. Because this survey changed an existing method, the authors used standard statistical techniques to conclude that the results were still statistically valid.

To measure personality traits, the authors used another standard questionnaire known as the "Big Five Inventory" or BFI (John et al. 1991). The BFI is a widely used self-reported personality trait questionnaire based on what is known as the Big Five model or

the Five Factor Model (McCrae and John 1992). The questionnaire consists of 44 items covering five personality traits: neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. Using a five-point Likert scale ranging from 1 (disagree strongly) to 5 (agree strongly), the respondent evaluates how well a first-person-singular description applies to them. For example, a question might say, “I see myself as someone who worries a lot.” The aggregated score within each trait provides a measure of how strong that trait is in the individual. Once again, this is a well-known method for measuring personality traits that has been statistically validated and tested.

The “Emotional Skills and Competence Questionnaire” or ESCQ45 (Takšić et al. 2009) is a questionnaire based on the model that assesses three emotional competences: perceiving and understanding, expressing and labelling, and managing and regulating emotions (Mayer and Salovey 1997). Respondents use a five-point Likert scale (rated from *never*, scored as 1, to *always*, scored as 5) to assesses the extent to which each of the 45 statements corresponds to their personality. For example, one statement says, “When I see how someone feels, I usually know what has happened to him.” Once again, the original data is ordinal, but this particular questionnaire has been statistically studied and the calculations used by the researchers appear to be valid.

This page titled 5.4.3: Satisfaction with Sport Participation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.4.4: The Supreme Court and Personal Attitudes

Earlier we considered a study of how the potential Supreme Court ruling on same-sex marriage would change individual's personal attitudes about the LGBTQIA+ community (Tankard and Paluck 2017). The research project of interest here tested these ideas in a controlled experimental setting where individuals read a brief article about the likely outcome of the upcoming Supreme Court ruling on same sex marriage. Each participant was randomly assigned to read either a version entitled “Supreme Court Likely to Rule in Favor of Gay Marriage” or a version entitled “Supreme Court Unlikely to Rule in Favor of Gay Marriage.” Participants then responded to a series of questions about demographics and other social views and behaviors (Tankard and Paluck 2017).

The researchers asked several questions, a few of which we will consider here in detail. These first questions dealt with what the individuals perceived as the feelings of society at the time about same-sex marriage. This score was computed by taking the average of the response to the questions, “To what extent do you think Americans oppose or support gay marriage?” and “To what extent do you think Americans oppose or support making gay marriage legal in the United States?” Each question used a scale from –4, which corresponded to being strongly opposed, to 4, which corresponded to being strongly supportive. Each of the questions are measured on an ordinal measurement scale and so the idea of averaging the two responses together should be questioned. For the researchers to average these two responses is must be permissible to add these values, a mathematical operation that is usually not necessarily allowed for ordinal data. To consider this question, we should determine what this value really means.

When a researcher is going to combine two measurements in this way, we need to first make sure that the two scales are comparable. In this case, both questions are measured using the same scale ranging from –4 to 4 with levels of agreement ranging from strongly opposed to strongly support. The next issue is whether the ratings mean the same thing for both questions. That is, does being strongly opposed in the first question have the same level of opposition as being strongly opposed in the other question? There is no way to theoretically know this for sure, and validating this idea would take further research. This is a problem that everyone would have to evaluate on their own, deciding how much trust they would like to put into these calculations.

One way to approach this issue is to imagine what such a combined score really means. Suppose that an individual reported a score of –4, indicating being strongly opposed for both questions, then the average would also be –4, which does indicate strong opposition. Similarly, a reported score of 4 for each question would have an average of 4, indicating strong support for both questions and strong support overall. But what if the response to one score is –4 and the other is 4? Then the average would be 0, indicating neither opposition nor support. Does this really mean the same as if an individual answered 0 to each question? That is, does strong opposition for one question and strong support for the other mean the same thing as neither opposition nor support on both questions? Again, this is an issue that everyone would have to evaluate on their own.

It should be noted that the issue of averaging over the two questions is an attempt to simplify the data used in the study and the corresponding analysis. Simplification will necessarily lose some information from the original data because the researchers are replacing the data from two questions with a single number. The question for the one evaluating the research is whether this simplifies the structure of the data to the point that the results cannot be trusted. In this case there are two relevant issues that might help us settle ourselves as to how we feel about this calculation. The first is that it is likely the responses to the two questions are strongly associated with one another. That is, an individual who gives a low score to the first question is probably also likely to give a low score to the second question as well. In this case one may reasonably assume that knowing how an individual answered one of the questions is almost as good as knowing how the individual answered both questions, and hence it may not be so bad if the two questions are summarized by an average. If this association holds, the case where an individual might answer –4 to the first question and 4 to the second question would be very rare and probably not worth worrying about.

The second relevant issue considers what will be done with the average score in the subsequent analysis. In this case the measurement is taken before, and then after, reading the randomly assigned article. Hence, in this case the researchers are interested in comparing the average score prior to the reading to the average score after the reading. Therefore, they are really interested in how the average might move because of the reading. Given the ordinal original measurement scale of the data, this type of comparison should be valid, as ordinal data does allow for size comparisons. Hence, if we are comfortable with the fact that the two measurements can be averaged, it should follow that the size comparison between the measurements should also be permissible.

This page titled 5.4.4: The Supreme Court and Personal Attitudes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.4.5: Product Placement in Movies

Earlier we considered a study that considered the reactions of individuals to movies containing product placements (d'Astous and Chartier 2000). This study was conducted in two separate phases. The first step of the study consisted of selecting 19 well-known movies subjectively grouped into three categories: action and adventure (6 movies), drama (6 movies) and comedy (7 movies). These movies were viewed by individuals with an interest in cinema, and all product placements were noted. Each placement was classified according to one of three product types: repeat purchase product, less frequently bought consumer good, and service. This process identified 191 product placements. The final sample of product placements was constructed by randomly selecting 5 placements per movie type and product type group resulting in 45 product placements to use in the study. Once the 45 product placements were extracted from the movies, they were watched by two judges who coded them on six concepts: visibility of the product or brand, mention of the brand, length, placement subtlety, principal actor uses the product, and principal actor is present.

Eleven individuals participated in the second phase of the study. Each individual watched a sample of the product placements, after which the individuals answered a series of questions on their thoughts about the placement, how well they recalled the product, and how the perceived the differences between the product placements. The actual design of the study is more complicated and precise, and we have summarized this part of the study design for the sake of brevity.

The participants answers were transcribed, and the content was analyzed. Six main categories were identified by the researchers: subtlety, length, integration within the scene, personal judgment, brand awareness, and specific mention of the brand. Ninety-six different terms were taken directly from the respondents transcripts, and twenty-four additional terms were added to represent other ideas based on the literature on product placements. This list of terms was reduced using a panel of 12 individuals who expressed their likelihood of using each term to describe or react to a product placement on a seven-point Likert scale. Terms whose mean rating was lower than a prescribed level were eliminated reducing the list to 31 terms. Finally, redundant terms were eliminated resulting in a final list of 27 terms. With this list of terms, the main part of the study was ready to proceed.

The main study recruited 103 students from a mid-size North American university. Each participant viewed 18 product placements that were presented in a random order. After viewing each placement, the respondents rated their agreement with the 27 terms developed in the preliminary study using a seven-point Likert scale. Data on the viewing habits and demographics of the participants were also gathered. Approximately one week later the participants were contacted by telephone and were asked what products or brands they recalled from the product placements viewed during the study. The participants were also given a list of products and were asked if they remembered having seen them in the product placements.

Before discussing the data used in this study, we will summarize the results. Some characteristics of the product placements were found to enhance consumer evaluations and memory of the products, while other product placement characteristics were found to have a negative effect. The data showed that the memory of a placement is enhanced when the principal actor is present and when the placement is positively evaluated. Similarly, consumer evaluations of the placements are more positive when the principal actor is present, when the placement is obvious and when it is well integrated in the movie scene. However, a placement that is judged as unacceptable and is well integrated in the movie scene has a negative influence on consumer memory. Researchers noted other effects in the study, but the results are much more complicated due to interactions between the variables. For example, the researchers noted that the prominence of product placement significantly enhances recognition memory but has a negative effect on recall.

The data collected for this study are based on the Likert scale, and the researchers performed an analysis that is statistically consistent with this type of data. The data that were merged for this study makes the corresponding analyses quite complicated. In addition to the data observed directly from the participants about their feelings and memory recall about the product placements, there was also demographic data on the participants along with data of the product placements themselves that were coded on the six concepts discussed earlier. Additionally, data was collected on the popularity of the movies used in the study as the researchers considered the possibility that the popularity of the movies could influence the evaluation and recall of the product placements.

Studies like this can be difficult to assess because of the many types and sources of data used and the complexity of the associated statistical analyses. The study is also particularly interesting because of the detailed development of the terms used with data collection prior to the main phase of the study. Overall, unless one is at least somewhat of an expert with the development of social

science measurement scales, a study like this can be difficult to assess. To some extent one must rely on the academic peer review process to ensure that the methods used with the observed data are appropriate.

This page titled 5.4.5: Product Placement in Movies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.4.6: The Racial Wealth Gap

The wealth gap between African American and white families in the United States has been a persistent problem, with little progress since the civil rights movement of the 1960s. Over fifty-five years after landmark civil rights legislation intended to equalize access to housing, education, and employment, the assets of African American families are less than 15 percent of the assets of white families (Bhutta et al. 2020). One research article considers how the income tax benefits for homeownership contribute to the racial wealth gap (Thomas 2022). As the researchers point out, the tax advantages in terms of wealth building associated with home ownership are significant, but the rate of home ownership for white families far exceeds that of African American families Bhutta et al. 2020). In other words, far more white families can benefit from these wealth-building tax breaks whereas African American families cannot.

Economists have traditionally used household income as a measure of financial well-being. However, recently economists have begun to focus on household wealth, the amount by which a family's assets exceed their debt and other liabilities, as a better measure. Although household income is associated with household wealth, it does not consider the resources that wealth represents, such as savings, home equity, and investments. Wealth is a measure of the resources available to improve the quality of life of the family and to provide stability and access to opportunities for well-being (Oliver and Shapiro 2013). Household wealth also accounts for how well a family would be able to survive unexpected events such as unemployment.

Data on household wealth for the study was provided by calculations from the *Urban Institute*, which used data from studies such as the *Survey of Financial Characteristics of Consumers*, *Survey of Changes in Family Finances*, *Survey of Consumer Finances*, and the *Current Population Survey*. All the surveys are administered by the Federal Government of the United States. The data collected from these surveys is based on family wealth, measured in dollars, which is measured on a ratio scale. That is, ratios can be taken with income data. For example, it makes sense to conclude that if one individual has wealth equal to $1000 and another has wealth equal to $500 then the first individual has twice as much wealth as the other because $1000 \div 500 = 21000 \div 500 = 2$. This means that nearly any mathematical operation is permissible with this type of data. In the research paper the median wealth is compared between African American, Hispanic, and non-white populations. If you are not familiar with the concept of a median, it is the point which divides the largest half of the data from the smallest half of the data. Therefore, to compute a median we need only be able to sort the data from smallest to largest which is permissible for data measured on a ratio scale. More about computing and interpreting the median appears later in this book.

Because the research article is considering the association between wealth and home ownership, rates at which families owned homes were also computed within each racial category. These data have a nominal measurement scale, as in each case a family either owns a home or not. Calculating the percentage of families in each category that own homes is permissible for nominal data.

This page titled 5.4.6: The Racial Wealth Gap is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.5: Summary

Data are central to the use of empiricism. Depending on what is being measured, and how data is being stored, the structure of data can vary from nominal data, which has no mathematical structure, to ratio data, which has a great deal of mathematical structure. The amount of structure that data has is referred to as the measurement scale of the data. It is extremely important that the calculations performed on the data in any statistical analysis be permitted by the measurement scale of the data. Performing mathematical calculations on data without the proper measurement scale can lead to conclusions that do not have any real meaning and can seriously weaken the impact and integrity of a research study. When critically evaluating a research study the measurement scale of the data should be considered along with whether the types of calculations performed on the data are valid.

This page titled 5.5: Summary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.6: About the Quotation

Neil deGrasse Tyson (1958) is an American astrophysicist and author whose research focuses on observations in cosmology, stellar evolution, galactic astronomy, and stellar formation. He has held numerous positions including at the University of Maryland, Princeton University, the American Museum of Natural History, and the Hayden Planetarium. He is also a prolific author of books on science and the universe. The quote at the beginning of the chapter is taken from in interview with *Parade* magazine.

This page titled 5.6: About the Quotation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.7: Practice and Exploration

1. The following question is used in a survey on the financial resources of individuals in a large city:

   Which of the following types of financial accounts do you have (mark all that apply)?

   - Savings Account
   - Checking Account
   - Retirement Account
   - Investment Account
   - Other

   The survey software is programmed to count the number of boxes that have been checked as the response to this question. What is the measurement scale for this data?

2. In a survey that tracks the online habits of residents in a city, a question asks for the age of the individual. Instead of asking for a numerical age, the question asks the individual to select one of the categories: Youth (17 years or younger), Young Adult (18 to 20 years), Adult (21 to 64 years), and Senior (65 years or older). What is the measurement scale for this data?
3. A research article on barriers to graduation for college students from lower-income backgrounds includes data on whether a student has unpaid fees, and tracks what type of unpaid fee the student has. The data are coded on using the table below:

Code	Meaning
0	No unpaid fee
1	Unpaid tuition fee
2	Unpaid laboratory fee
3	Unpaid non-academic fee
4	Unpaid parking fee

   In the research article the authors average the coded values over all the students. The resulting average is 2.3. The authors then state that on average students are stuck between unpaid laboratory and non-academic fees. What is the measurement scale of the data? Do you feel that the conclusions are valid?

4. One of the variables in a survey of undergraduate students asks what year they intend to graduate. For example, a student will enter the number 2027 if they intend to graduate in the year 2027. What is the measurement scale for this data? Would the measurement scale change if the question was re-worded to ask, "How many years until you graduate?" Why or why not?
5. A study of automobile affordability has data with variables corresponding to gas mileage and the retail price of each automobile. The data also contains a variable that identifies whether the vehicle is a hybrid or not. What are the measurement scales for each of these variables?
6. An employment study asks individuals in which sector of the economy they are employed. The question uses the categories "*education*", "*business*", "*service*", "*skilled labor*", and "*other*.". When studying this data, it is it valid to compute the percentage of observations in each sector? Knowing that this is nominal data, explain why it is valid to compute these percentages.
7. A study of the levels of education attained by individuals codes the data as follows:

Education Level	Code
Attained grade level k where k can be 1 to 12	k
Completed k years of undergraduate education	$12+k$
Completed k years of graduate education	$16+k$

Hence, for example, if an individual attained an 8th grade education, then the response would be coded as 8, and an individual who has completed 3 years of undergraduate education would be coded as in 15. What is the measurement scale of this data? When deciding on the measurement scale think carefully about how college education is counted.

8. In a large health database one of the variables is the weight of the individual in kilograms. What is the measurement scale for this data?
9. The following question is used in a survey on the home life of college students:

   Which of the following types of pets do you have at your place of residence (mark all that apply)?

   - Cat
   - Dog
   - Rabbit
   - Fish
   - Reptile
   - Other

   The survey software is programmed to record the list of which boxes have been marked. What is the measurement scale for this data?
10. In a survey to measure support for restarting the push for the Equal Rights Amendment to the United State Constitution, individuals are asked how much they support the amendment based on the measurement scale –1 (oppose), 0 (no feeling), and 1 (support). The individuals are then asked to read an essay by a famous scientist that supports the amendment. The individuals are then asked how much they support the amendment based on the measurement scale –2 (strongly opposed), –1 (oppose), 0 (no feeling), 1 (support), and 2 (strongly support). The data manager then computes how much the support of the individual has changed by taking the result from the second question and subtracting the result of the first question. That is, if an individual stated that they had no feeling about the amendment prior to reading the essay, and the stated that they strongly supported the amendment after reading the essay, then the final score would be 2−0=2. Do you feel that this calculation is valid for the data from these two questions? Why or why not? Can you suggest some improvements?
11. In a study of family wealth, individuals are asked about their family structure. Specifically, the study asks everyone to identify their marital status using the choices in the first column of the table shown below.

Status	Data Code
Single	0
Partner	1
Married	2
Divorced	3
Separated	4

   The status *partner* refers to cohabitation with another individual to whom they are not legally married. What is the measurement scale for this data? Given this measurement scale, do you think it would be permissible to compute the average response for the individuals in the study? Why or why not?
12. A historical researcher is interested in the racial views of working-class individuals who lived in a racially integrated city at the turn of the 20th century. To study these views the researcher has obtained eyewitness statements contained in historical county and city judicial records from the period 1895 to 1905. The researcher intends to summarize the views of these records by quoting representative passages that show either racial bias or racial tolerance. Are the data in this study qualitative or quantitative? If the data are quantitative, what is the measurement scale?
13. Two researchers are developing a project that will study the association between attitudes about civil rights, race, and age. The researchers have decided to use a simplified form of the questions used on census forms in the United States that asks

individuals to identify their race and ethnicity. This will allow the researchers to directly connect their data with census data available from the United States Census Bureau. The questions have the following form:

Are you of Hispanic, Latino, or Spanish origin?

- Yes
- No

What race do you consider yourself?

- African American
- American Indian or Alaska Native
- Asian
- Native Hawaiian or Pacific Islander
- White
- None of these categories describe my race

During the discussions about developing the survey, the question of how the race and ethnicity data were to be stored electronically became important. One researcher wanted to store the data using characters. For the first question each response would be stored as with Y or N, and for the second question the codes AF, AI, AS, NH, WH, and NO would be used. The second researcher objected to this scheme and suggested numbering the categories as 0 and 1 for the first question and 0–5 for the second question. Does the representation of the data as either characters or numbers affect the measurement scales for the data? Why? Suppose that the second researcher states that the second method is better because they could then compute the average race of the respondents. How would you respond to that statement?

14. A researcher interested in economic justice is studying programs that provide tax benefits for individuals who donate items to a charitable organization. The researcher is interested in the fact that for many the tax benefits will never be realized because their income is insufficient for such tax benefits to have any effect. In these cases, the researcher would like to argue that other types of benefits should be offered to those individuals. As part of the study, the researcher obtains data from several such programs. The data have been blinded so the researcher cannot identify the individuals involved, but data on income, demographics, and value of donation are available. Based on the income and demographics, the researcher can determine if the individual would likely be able to benefit from the tax benefit. The researcher creates a variable with the following coded values:

Code	Meaning
0	Not likely to get tax benefit
1	Even chance of getting tax benefit
3	Likely to get tax benefit

Identify the measurement scale for this variable. Would it make sense for the researcher to compute the percentage of individuals in each category? Would it make sense for the researcher to compute the average of this variable?

15. A historical researcher is interested in the racial views of working-class individuals who lived in a racially integrated city at the turn of the 20th century. To study these views, the researcher has obtained eyewitness statements contained in historical county and city judicial records from the period 1895 to 1905. For each statement the researcher will use specially designed computer software that will compute how many words are contained in the statement, and how many of those words are on the list of derogatory racial terms. The researcher will then compute the derogatory term rate by dividing the number of derogatory racial terms in the statement by the total number of words in the statement. Are the data in this study qualitative or quantitative? If the data are quantitative, what is the measurement scale?
16. Using an online news service, find a news article about an issue that is important to you personally that is based on an empirical analysis. Identify, as clearly as you can, the population that the authors of the news article have used to observe their data and gather information. What are the variables in the data they obtain? What are the measurement scales of each of the variables?

Do the authors do any calculations that seem questionable to you based on the measurement scales of the variables? Why or why not?

17. Using an online source or your library find an academic research article about an issue that is important to you personally that is based on an empirical analysis. Identify, as clearly as you can, the population that the authors of the news article have used to observe their data and gather information. What are the variables in the data they obtain? What are the measurement scales of each of the variables? Do the authors do any calculations that seem questionable to you based on the measurement scales of the variables? Why or why not?

This page titled 5.7: Practice and Exploration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.8: Collaborative Activity

This collaborative activity will consider some simple analyses and experience using a Likert scale. While the activity is a physical activity based on creating a “house” using standard playing cards, the physical abilities of the members of your group will not be judged, but rather members will share their experiences on the difficulty of the physical tasks. The members of the group should not feel the need to achieve the tasks if they do not feel comfortable in doing so.

This activity consists of the members of your group each attempting to build three different “houses” using standard paying cards. The activity of creating structures by stacking playing cards on top of one another has probably been around as long as playing cards have existed. What makes the structures interesting is that they hold together completely by friction. No glues or other types of adhesives are used in their construction. Additionally, the playing cards cannot be physically altered. For example, they may not be bent or have notches cut into them.

For this activity each member of the group will attempt to build three houses of cards. The members do not have to be successful in completing each task. After the three tasks are attempted, each group member will then take a short survey on their experience with the tasks, and the group will discuss how the data could be used to learn about the difficulty of the tasks.

The first task creates a simple house of cards using five cards. To begin building this card house, place one of the five cards on a flat surface with one of the long edges down. While holding this card, lean a second card up against it, also placing one of the long edges on the flat surface. The cards should form a T shape as shown in Figure **5.8.1** below.



Figure **5.8.1**: The T-shape formed by two cards. (Public domain photograph by Alan M. Polansky).

Now, on the end of one of the top parts of the T, carefully form another T shape using a third card as shown in Figure **5.8.2** below.

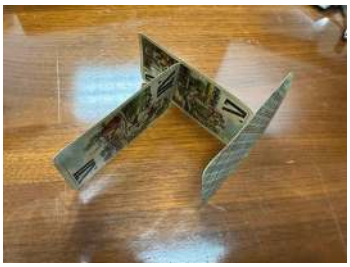


Figure **5.8.2**: The cards after the third card is added. (Public domain photograph by Alan M. Polansky).

Finally, at the end of this T, form a square by taking a fourth card, forming another T whose edge leans against the first card, as shown below.

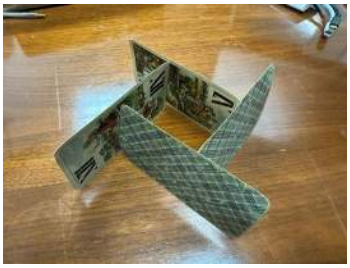


Figure **5.8.3**: The cards after the fourth card is added. (Public domain photograph by Alan M. Polansky)

Now carefully place a roof on the house by lightly placing a card on top as shown in Figure **5.8.4** below.



Figure 5.8.4: The completed house of cards. (Public domain photograph by Alan M. Polansky)

The second task creates a house of cards also using five cards. To begin take two cards and place them on the flat surface leaning into one another so that the top edges touch as shown below. See Figure 5.8.5.

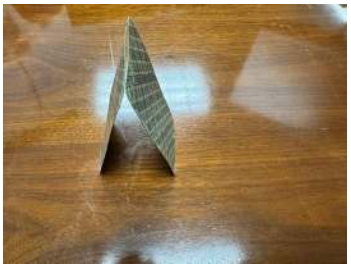


Figure 5.8.5: The first step in the second card house. (Public domain photograph by Alan M. Polansky)

As shown in the figure, the short edges should be on the flat surface and touching at the top. If the cards are balanced carefully enough, the friction on the flat surface and where the cards touch at the top should hold them in place. Be sure to have the cards vertical enough so that the bottom edges do not slide outward. Now, using two more cards, repeat this process right next to the two cards that have already been placed to that you get a structure that looks like an M (with a space in the middle) as shown in Figure 5.8.6 figure below.

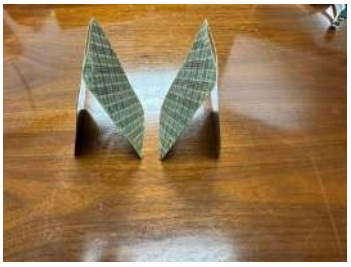


Figure 5.8.6: The second step in the second card house. (Public domain photograph by Alan M. Polansky)

To finish the house of cards, very carefully place the fifth card flat on top spanning the two peaks as shown in Figiure 5.8.7 below.



Figure 5.8.7: The final step of the second card house. (Public domain photograph by Alan M. Polansky)

The second type of house of cards is now complete!

The third type of house of cards is an extension of the second type of house of cards using two additional cards for a total of seven cards. For this house of cards, begin by assembling the second type of house of cards. Once that is complete, take the two last cards and add them on top of the top horizontal card by leaning them against each other in the same way that the two pairs of cards were constructed on the bottom, as shown in Figure 5.8.8 below.



Figure 5.8.8: Completed third house of cards. (Public domain photograph by Alan M. Polansky)

The third type of house of cards is now complete!

Once each member of your group has attempted to build each of the three types of houses of cards, you will each take the survey below, which is scored on a Lickert scale coded as follows:

Score = 0 if the task was very easy

Score = 1 if the task was easy

Score = 2 if the task was neither easy nor difficult

Score = 3 if the task was difficult

Score = 4 if the task was very difficult

The House of Cards Survey

Score each answer as a whole number between 0 and 4 using the scale shown above.

1. How difficult did you find building the first card house?

2. How difficult did you find building the second card house?

3. How difficult did you find building the third card house?

4. If you could practice building the first card house ten times, and then try again, how difficult do you imagine it would be?

5. If you could practice building the second card house ten times, and then try again, how difficult do you imagine it would be?

6. If you could practice building the third card house ten times, and then try again, how difficult do you imagine it would be?

7. Consider the card house shown in the Figure 5.8.9 below. How difficult do you think it would be to build?



Figure 5.8.9: House of cards for survey question 7. (This work has been released into the public domain by its author, Merzperson at English Wikipedia. This applies worldwide.)

The results from the survey should be recorded in a data table like the one shown below:

Group Member	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6	Question 7
1							
2							
3							
4							
5							

If there are more than five people in your group, additional lines can be added to accommodate their responses.

Questions:

1. Consider the data that was observed for building the first card house. Was there agreement between the participants as to how easy it was for them to build or attempt to build the card house? If you were asked to summarize in a single sentence how difficult each group member found building the first card house, what would that sentence be?
2. Consider the data that was observed for building the second card house. Was there agreement between the participants as to how easy it was for them to build or attempt to build the card house? If you were asked to summarize in a single sentence how difficult each group member found building the second card house, what would that sentence be?
3. Consider the data that was observed for building the third card house. Was there agreement between the participants as to how easy it was for them to build or attempt to build the card house? If you were asked to summarize in a single sentence how difficult each group member found building the first card house, what would that sentence be?
4. Compare the responses from building the first card house with the responses from building the second card house. Does it appear that one card house is more difficult to construct? Provide details on how you reached your conclusion.
5. Now compare the responses from building the second card house with the responses from building the third card house. Does it appear that one card house is more difficult to construct? Provide details on how you reached your conclusion. Was it easier to reach this conclusion that the one you found in the previous question?
6. Consider the answers given to questions 4 through 6 and compare them to the answers given to questions 1 through 3. Do the data indicate that your group members feel that practicing building the card houses will decrease the difficulty level of building the card houses?
7. Consider the responses given to question 7. How much agreement is there between your group members on how difficult the pictured card house would be to construct?
8. Suppose someone asks you to redesign the survey so that the data collected uses a ratio scale instead of a Likert scale. What type of data could you observe?

This page titled 5.8: Collaborative Activity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

5.9: References

Baudin, N., A. Aluja, J.-P. Rolland, and A. Blanch. 2011. “The Role of Personality in Satisfaction with Life and Sport.” *Behavioral Psychology* 19, no. 2, 333–345. https://doi.org/10.3389/fpsyg.2021.820045

Bebetsos, E., and N. Theodorakis. 2003. “Athletes’ Satisfaction Among Team Handball Players in Greece.” *Perceptual and Motor Skills* 97, no. 3 suppl, 1203–1208. https://doi.org/10.2466/ pms.2003.97.3f.1203

Bhutta, Neil, Andrew C. Chang, Lisa J. Dettling, Joanne W. Hsu, and Julia Hewitt. 2020. “Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances.” *FEDS Notes*. https://www.federalreserve.gov/econr...-20200928.html

Bleich, Sara N., Mary G. Findling, Logan S. Casey, Robert J. Blendon, John M. Benson, Gillian K. SteelFisher, Justin M. Sayde, and Carolyn Miller et al. 2019. “Discrimination in the United States: Experiences of Black Americans.” Health Services Research 54, 1399–1408. https://doi.org/10.1111/1475-6773.13220

Brownmiller, Susan. 1993. Against our Will: Men, Women, and Rape. Ballantine Books. https://frauenkultur.co.uk/wp-conten...t-Our-Will.pdf

D’Astous, Alain, and Francis Chartier. 2000. “A Study of Factors Affecting Consumer Evaluations and Memory of Product Placements in Movies.” *Journal of Current Issues and Research in Advertising* 22, no. 2, 31–40. https://doi.org/10.1080/10641734.2000.10505106

Ferro, Christine, Jill Cermele, and Ann Saltzman. 2008. “Current Perceptions of Marital Rape: Some Good and Not-So-Good News.” *Journal of Interpersonal Violence* 23, no. 6, 764–779. https://doi.org/10.1177/0886260507313947

Findling, Mary G., et al. 2019. “Discrimination in the United States: Experiences of Native Americans.” *Health Services Research* 54, 1431–1441. https://doi.org/10.1111/1475-6773.13224

Gemelas, Jordan, Jenna Davison, Case Keltner, and Samantha Ing. 2022. “Inequities in Employment by Race, Ethnicity, and Sector During COVID-19.” *Journal of Racial and Ethnic Health Disparities* 9, no. 1, 350–355. https://doi.org/10.1007/s40615-021-00963-3

Gorczynski, P. F., M. Coyle, and K. Gibson. 2017. “Depressive Symptoms in High-Performance Athletes and Non-Athletes: A Comparative Meta-Analysis.” *British Journal of Sports Medicine* 51, no. 18, 1348–1354. https://doi.org/10.1136/bjsports-2016-096455

Gouttebarge, V., et al. 2019. “Occurrence of Mental Health Symptoms and Disorders in Current and Former Elite Athletes: A Systematic Review and Meta-Analysis.” *British Journal of Sports Medicine* 53, no. 11, 700–706. https://doi.org/10.1136/bjsports-2019-100671

John, O. P., E. M. Donahue, and L. R. Kentle. 1991. *The Big Five Inventory—Versions 4a and 54*. University of California, Berkeley, Institute of Personality and Social Research. https://doi.org/10.1037/t07550-000

Joshi, Ankur, Saket Kale, Satish Chandel, and D. Kumar Pal. 2015. “Likert Scale: Explored and Explained.” *British Journal of Applied Science and Technology* 7, no. 4, 396–403. https://doi.org/10.9734/BJAST/2015/14975.

Lepir, D., and S. Lakić. 2025. “Correlates of Sport Satisfaction: The Role of Success Level, Personality Traits, and Emotional Competence in Team and Individual Sports.” *Psihologija*, 3 (online issue). https://doi.org/10.2298/PSI240723003L

Mayer, J. D., and P. Salovey. 1997. “What is Emotional Intelligence?” In *Emotional Development and Emotional Intelligence: Educational Implications*, edited by P. Salovey and D. J. Sluyter. Basic Books.

McCrae, R. R., and O. P. John. 1992. “An Introduction to the Five-Factor Model and Its Applications.” *Journal of Personality* 60, no. 2, 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McMurtry, Caitlin L., et al. 2019. “Discrimination in the United States: Experiences of Asian Americans.” *Health Services Research* 54, 1419–1430. https://doi.org/10.1111/1475-6773.13225

Moesch, K., G. Kenttä, J. Kleinert, C. Quignon-Fleuret, S. Cecil, and M. Bertollo. 2018. “FEPSAC Position Statement: Mental Health Disorders in Elite Athletes and Models of Service Provision.” *Psychology of Sport and Exercise* 38, 61–71.

https://doi.org/10.1016/j.psychsport.2018.05.013

Moore, E. M., M. A. Simmons, and J. Threatt. 2025. “Mental Health Challenges and Substance Use in Student-Athletes: Prevalence and Impact.” *Frontiers in Sports and Active Living* 7, 1527793. https://doi.org/10.3389/fspor.2025.1527793

Myers, Randy. 2017. “‘I Accept That I Have Nobody’: Young Women, Youth Justice, and Expectations of Responsibility During Reentry.” *Social Justice* 44, no. 1, 62–82. https://www.jstor.org/stable/26405738?seq=1

Nicholson, Harvey L., and Di Mei. 2020. “Racial Microaggressions and Self-Rated Health Among Asians and Asian Americans.” *Race and Social Problems* 12, no. 3, 209–218. https://doi.org/10.1007/s12552-020-09293-1

Oliver, Melvin, and Thomas Shapiro. 2013. *Black Wealth/White Wealth: A New Perspective on Racial Inequality*. Routledge.

Payne, Diana L., Kimberly A. Lonsway, and Louise F. Fitzgerald. 1999. “Rape Myth Acceptance: Exploration of Its Structure and Its Measurement Using the Illinois Rape Myth Acceptance Scale.” *Journal of Research in Personality* 33, no. 1, 27–68. https://doi.org/10.1006/jrpe.1998.2238

Polman, R., N. Rowcliffe, E. Borkoles, and A. Levy. 2007. “Precompetitive State Anxiety, Objective and Subjective Performance, and Causal Attributions in Competitive Swimmers.” *Pediatric Exercise Science* 19, no. 1, 39–50. https://doi.org/10.1123/pes.19.1.39

Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen. 2017. “Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring Over Time.” *Proceedings of the National Academy of Sciences* 114, no. 41, 10870–10875. https://doi.org/10.1073/pnas.1706255114

Riemer, H. A., and P. Chellandurai. 1998. “Development of the Athlete Satisfaction Questionnaire (ASQ).” *Journal of Sport & Exercise Psychology* 20, no. 2, 127–156. https://doi.org/10.1123/jsep.20.2.127

Schwendinger, Julia R., and Herman Schwendinger. “Rape Myths: In Legal, Theoretical, and Everyday Practice.” *Crime and Social Justice* 1, 18–26, 1974. https://www.jstor.org/stable/29765884

Takšić, V., T. Mohorić, and M. Duran. 2009. “Emotional Skills and Competence Questionnaire (ESCQ) as a Self−Report Measure of Emotional Intelligence.” *Psihološka Obzorja / Horizons of Psychology* 18, no. 3, 7–21. http://psiholoska-obzorja.si/arhiv_c...oric_duran.pdf

Tankard, Margaret E., and Elizabeth Levy Paluck. 2017. “The Effect of a Supreme Court Decision Regarding Gay Marriage on Social Norms and Personal Attitudes.” *Psychological Science* 28, no. 9, 1334–1344. https://doi.org/10.1177/0956797617709594

Tessler, Hannah, Meera Choi, and Grace Kao. 2020. “The Anxiety of Being Asian American: Hate Crimes and Negative Biases During the Covid-19 Pandemic.” *American Journal of Criminal Justice* 45, no. 4, 636–646. https://doi.org/10.1007/s12103-020-09541-5

Thomas, Ann F. 2022. “The Racial Wealth Gap and the Tax Benefits of Homeownership.” *NYLS Law Review* 66, no. 2, 247–274. https://digitalcommons.nyls.edu/cgi/...yls_law_review

Yearby, Ruqaiijah, and Seema Mohapatra. 2020. “Law, Structural Racism, and the Covid-19 Pandemic.” *Journal of Law and the Biosciences* 7, no. 1, lsaa036. https://doi.org/10.1093/jlb/lsaa036

This page titled 5.9: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

CHAPTER OVERVIEW

6: Measuring Abstract Information

“I am constantly trying to communicate something incommunicable, to explain something inexplicable, to tell about something I only feel in my bones and which can only be experienced in those bones.”

– Franz Kafka in a letter to Milena Jesenská

6.1: What to Expect
6.2: Abstract Concepts
6.3: Conceptualization and Operationalization
6.4: Validity and Reliability
6.5: Established Measurement Systems
6.6: Critical Analysis
6.6.1: Attitudes About Marital Rape
6.6.2: Gender and Race-Related Stress
6.6.3: Measuring Satisfaction in Sports
6.7: Summary
6.8: About the Quotation
6.9: Practice and Exploration
6.10: Collaborative Activity
6.11: References

This page titled 6: Measuring Abstract Information is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.1: What to Expect

There are many abstract concepts in research studies of vital interest that cannot be measured directly. In this chapter we will investigate the methods that researchers used to measure these abstract concepts indirectly using measurements that can be observed. This process is based on a system of conceptualization, deciding what is to be measured, and operationalization, defining how the concept is going to be measured. To ensure that the measurement system is sound, the reliability and validity must be carefully studied. We will discuss how researchers decide if a measurement system is both reliable and valid.

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

1. Identify when a measurement system is being used to measure an abstract concept that cannot be measured directly.
2. Determine if a measurement is directly observable, indirectly observable, or a construct.
3. Define a concept and be able to identify concepts.
4. Define and discuss the process of conceptualization.
5. Determine the dimensions and indicators of a concept.
6. Define and discuss the process of operationalization.
7. Define and discuss validity and reliability.

This page titled 6.1: What to Expect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.2: Abstract Concepts

Many researchers have documented the detrimental health effects resulting from the recent wave of aggression and discrimination against Asian Americans during the recent COVID pandemic (Iwamoto 2022; Nicholson and Mei 2020; Yellow Horse 2021; Le et al. 2020; Wang et al. 2020; Wu et al. 2021; Chen et al. 2020). This type of research can provide important insights for health professionals when assessing and treating the mental and physical health of their patients.

One research study sought to determine if racial microaggressions were influencing the health of Asian Americans (Nicholson and Mei 2020). Microaggressions are statements, actions, or incidents of indirect or subtle discrimination often resulting in harmful consequences to members of marginalized groups (Sue, 2010). When considering undertaking such a study, an immediate problem becomes obvious. How do the researchers measure health and exposure to microaggressions?

A person's health is a rather complex and abstract concept. What contributes to a person being “healthy”? Certainly, there are many related components that contribute to the overall health of a person. Some numerical measures may be somewhat easy to determine if a member of the population can be examined by a medical professional. These include measures such as resting heart rate, weight, and cholesterol levels. There are literally hundreds of other measures that can be taken on an individual. But how does one compute a measure of a person's health based on these measurements? This is a daunting task particularly considering the many complex relationships that may exists between these measurements and the overall health of a person. For example, a healthy weight for one person may not be considered healthy for another. Beyond these measurements, mental health may also play an important role in determining the general health of an individual. How do we measure mental health?

Assuming we can collect enough measurements from everyone that we would consider to be important to measuring the overall health of a person, how do we use them to compare the health of individuals? This comparison is important because our research hypothesis states that microaggressions have a negative effect on health. Therefore, we need to be able to compare the health of two individuals and definitively state which individual has better health. Given the apparent large number of measurements that we have taken on everyone, how do we decide who is healthier? It seems that we should combine all these measurements into a single measure of overall health, but how should this be done? Is high blood pressure less healthy than a high cholesterol score? How does mental health compare to physical health?

It should become apparent that this has become an exceedingly difficulty concept to measure. Nonetheless, these questions highlight the major difficulties that researchers routinely encounter in statistical research studies.

If microaggressions are having a significant adverse effect one or more groups in a society, then that group is being denied not only their health but they are also potentially being penalized with the economic consequences of additional health care needs and the potential for the loss, or reduction, in employment opportunities. So, the question is, how can this important topic be studied in a way that is relatively simple, reliable, effective, and efficient?

Let us consider how the researchers in this study approached this problem (Nicholson and Mei 2020). There are two important things that must be measured for this study to be effective: exposure to microaggressions and overall health. The data used in the study was based on observations from the 2016 National Asian American Survey, a large and ethnically diverse source of nationally representative data on Asians living in the United States. In this study the health of the individuals was measured by their responses to the question, “How would you rate your overall health in the past-year?” The responses were reported using the categories excellent, very good, good, fair, and poor. This is qualitative categorical data measure using an ordinal measurement scale.

This method for measuring overall health represents an incredibly simple solution given the complexities discussed above. However, there is a very important question that this raises, namely do these responses measure the individual's overall health? Of course, this is not a perfect measure of overall health, particularly because individuals may suffer from conditions that they may not be aware of, or they may discount other relatively important conditions. Hence, someone who is in poor health may report that they are healthy, and correspondingly someone who is in good health might report that they are in poor health. Hence, the measure is not going to be a perfect measure of overall health. However, can this measure still give us an indication of health that may be useful for a study such as this one?

To address this issue, we need to do a little more research. Looking into past research it was found that these measures were strongly correlated with multi-item health measures (Ahmad et al. 2014), predicted risk of mortality with high reliability (Idler and Kasl 1991), and were also strongly correlated with standard well-being scales (McDowell 2009). So, what does all this mean?

Some of the terminology used above may be unfamiliar, and much more on these subjects will be covered later. For now, let us tackle these issues one-by-one. The first result states that these measures are strongly correlated with multi-item health measures; instead of asking everyone their overall health as a single question, the researcher could ask a series of questions that are all intended to assess some aspects of the individual's health. The idea that the single item measurement strongly correlates with the multi-item measurements means that, basically, the researcher will get similar information from the single-item measure as a multi-item measure. It is very important to note that the exchange is not perfect, and some information will be lost. But what this result states is that not too much information will be lost. In terms of the study, the researchers will hope that enough information is retained about overall health so that some interesting conclusions can be obtained from the data.

So, the single item measure gives us similar information as the multi-item measure. But does this mean that we get useful information about the true health of the individual, or is an individual's self-reported health not informative for physical health? That is, is it enough that we get similar information from the single item measure and is the multi-item measure itself a good indicator of health? The next result states that the single-item measure can predict the risk of mortality for individuals. In the context of this type of research, this means that the responses from the single-item measure are closely associated with mortality rates. That is, those who reported poorer health tended to have a higher mortality rate within a specified time than those who reported better health. Furthermore, this association was strong enough to conclude that useful information is being obtained by the prediction, which is a pretty strong endorsement in that an individual's self-reported health score usefully predicts something that will happen in the future that is associated with general health. Of course, individuals may die from non-health related causes such as fatal accidents or violence, but mortality rates should provide a good indication of general health.

The last result states that the self-reported health measurements have a high degree of reliability and are also strongly correlated with standard well-being scales. There are two results to consider here. The first refers to the reliability of the measure, a concept we will discuss in more detail later in this chapter. For now, what this roughly means is that the measure provides consistent results over time (Van de Sande and Byvelds 2015). That is, the same individual with the same health profile will give the same measurement with repeated observations. The fact that the health measure strongly correlates with standard well-being scales provides further evidence of the validity of the measure. What is being stated here is that the measure seems to be measuring overall health in the same way as several other established health measures. By stating this conclusion, the researchers are borrowing from the strength of other research. The implication here is that these established health measures have been shown to be useful in other studies, and that if this measure is essentially giving very similar observations, then it should be useful as well.

In the material that follows we consider how researchers develop these measures along with how to develop your skills to critically analyze how abstract information is measured in research studies. Research studies begin with a process of conceptualization and operationalization where the researcher decides what and how items should be measured. As in the example above, measurements will be useful only after the measurement systems have been shown to be both valid and reliable. No studies are perfect, and all research studies must forge a compromise between practicality and usefulness. From a critical point of view, one must determine if the compromises are reasonable and if the conclusions of the study are warranted considering these compromises.

This page titled 6.2: Abstract Concepts is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.3: Conceptualization and Operationalization

At the beginning of any observational research study, the researcher must decide what exactly will be observed from the individuals or items in the population. This process is known as conceptualization. Conceptualization is an important part of any research study, and poor conceptualization can often result in weak results that can be easily contradicted by other studies (Blalock 1982). This results in general confusion in the research community about what the important issues really are, and in the broader community, the entire research field can seem unreliable and contradictory. This problem makes it exceedingly difficult for well planned research with important conclusions to be recognized.

Conceptualization can be particularly challenging because many of the variables of greatest interest can be extremely difficult to measure. In many cases research studies deal with abstract concepts for which there is not a specific method for measurement. For example, a researcher may wish to study the anxiety levels of children in a public school. Because there is no direct method for observing the anxiety level of a child, such a measurement must be accomplished indirectly. Other data may not be available in the historical record. For example, we might be able to design a method for measuring the anxiety level of a child, but there is no way to go back and make these measurements on children from previous years. Some types of data may be expensive to collect. In a large study of the effect of exercise on health, it would be cost prohibitive to have every study participant go through a complete medical examination for the sole purpose of the study. There may also be ethical or legal reasons that some data cannot be collected. Studies that consider health related variables, in addition to being expensive, may be hampered by privacy concerns. Still other variables may be difficult to measure because of societal reasons. An invasive health study may offer a small payment in exchange for participating in the study, but such an incentive is unlikely to motivate wealthier individuals who may form an important comparison group to participate in the study.

These problems have profound implications on the application of empirical studies. In some cases, researchers may simply decline to attempt to measure such variables, while in other cases researchers may simply side-step this issue and consider less complicated research questions based on data that is easier to measure. These are unfortunate approaches considering the relative importance of many of the related issues in empirical research. On the other hand, researchers may simply seek out convenient measurement methods, stating that the variable is difficult to measure and so they will choose an arbitrary method for measuring it, which is better than having no study at all. This type of approach is scientifically dangerous as many convenient methods of measurement may not really measure what the researcher is assuming that they measure after all. This can lead to erroneous results. Even in the case where the results may be correct, the perceived reliability of the study can be fundamentally damaged by the lack of care taken with selecting the measurements.

While the best approach for researchers is arguably to face the problem and do the best they can with the data they can observe, the researcher is obligated to defend and justify how abstract concepts are measured. The researcher is also obligated to admit to the shortcomings of the study, and to further explore ways that these shortcomings can be addressed.

In considering the development of measurement systems, all measurement methods, even the ones used regularly in the scientific community, are necessarily indirect measurements of physical phenomenon. For example, the conceptual idea of temperature can be measured using a mercury-based thermometer. Such a device does not measure temperature directly but is rather based on a physical model that describes the physical behavior of mercury in a glass capillary. By observing the physical reaction of the mercury in the vial, we are really measuring—indirectly—the volume of the liquid metal using a known physical model to infer the corresponding temperature. Similarly, the conceptual idea of mass can be measured by how much a spring is compressed when put under the weight of an object.

In the social sciences the physical model is replaced by one that is more abstract and more complicated, and about one for which much less may be known. Therefore, researchers should be cautious. At the most basic level the conceptual idea of a physical and social model is very similar. Consider a basic cause-and-effect type relationship in which there are stimuli acting on a system that evokes a response (Trimmer 1950). This simple idea is diagrammed in Figure 6.3.1. Within the system additional sub-systems could interact with one another to produce a response. In this case researchers may be able to observe the stimuli but may be unable to observe the interactions between the sub-systems. This causes the system to appear to be very complex. It is often the case in social science research that these sub-systems and their interactions may be unobservable, and in this case the system is seen as a type of black box (or in this case a blue box), where stimuli are interpreted by a very complex, invisible process to produce the response.

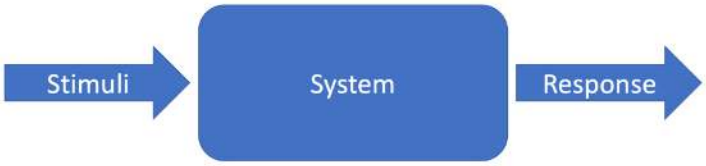


Figure **6.3.1**: Diagram of stimuli acting on a system to invoke a response (Public domain image created by Alan M. Polansky).

This type of black box system is fraught with difficulty for those who wish to make connections between the stimuli and responses. It is quite possible that some, or even all, of the stimuli we can observe do not have a direct bearing on the system but are interacting with unobserved stimuli that do. Similarly, even if we can observe all the stimuli that affect the system, the complex interactions within the unknown system may make it impossible to learn about how the stimuli affect the system. This is made even more difficult by the fact that researchers usually do not control the stimuli. That is, they observe the stimuli simultaneously with the response.

An example of a simple system can help emphasize the context of these issues. Consider a stimulus that can have one of two possible values: 1 or 2. From the black box we then observe whole numbers between 1 and 10. Some example observations from this process are given in Table 6.1. The relevant question is, what does the black box do with the stimulus to get the response?

Table 6.1 Values of a stimuli that can have one of two possible values (1 or 2) and the resulting response from a black box.

Stimulus (x)	2	1	2	2	1	1	1
Response	6	4	7	3	3	3	4

From Table 6.1 it is impossible to predict what will happen for each possible value of the stimulus. When the stimulus is equal to 1, we observe 4, 3, 3, 3, and 4, whereas when the stimulus is 2, we observe 6, 7, and 3. The only hint seen in these observations is that when the stimulus is equal to 2, the numbers tend to be larger. However, given the variation in the observations, most people would not have too much confidence in saying that this general trend would hold if we were to have access to further observations. It is simply impossible to say for sure what happens in the black box.

The back box is actually very simple in this case, but we have not been allowed to see all the stimuli. It turns out that there are two other stimuli, which we will call y and z. What happens in the black box is that if $x = 1$, then the response is simply equal to y. If $x = 2$ then the response is equal to $x + y$. This can be written in a mathematically succinct way as

$$response = y + (x - 1) \times z. \tag{6.3.1}$$

Table 6.2 shows all the stimuli and the related responses. Note that once all the stimuli are known, predicting the response if very simple. However, if these stimuli are unknown, the behavior seems very complex. That is, if only x is known, then the problem of predicting the response is intractable.

Table 6.2 Values of stimuli and the resulting response from a black box.

Stimulus (x)	2	1	2	2	1	1	1
Stimulus (y)	3	4	3	1	3	2	4
Stimulus (z)	3	2	4	2	5	2	5
Response	6	4	7	3	3	3	4

In practice we can think about how stimuli may operate in a larger complex system to produce the data we observe. For example, we may postulate that race and ethnicity influence income, but of course the relationship is not a direct relationship. That is, all the people with the same race do not have the exact same income. Hence, there must be other stimuli, in this case probably a large number of stimuli, that also affect income. These stimuli could also possibly interact with race. A researcher must carefully analyze what these stimuli may be, determine if they can be measured, and consider whether they may also be related or associated with the

levels of the other stimuli. In some cases, researchers may observe a variable that might not have any direct effect on the response but is merely associated with it.

To see how this can happen, consider a second simple black box example. In Table 6.3 we have a potential stimulus for a response. In this case the stimulus x that can be a whole number between 0 and 5, while the response is a whole number between 0 and 10. While larger values of the response tend to be associated with larger values of the stimulus, it is not possible to predict what the response will be based on the stimulus x. For example, the stimulus is equal to 3 for three of the observations, but the response is equal either to 8 or 7. The black box is again very simple, two unobserved stimuli y and z whose numbers are between 0 and 5 that are added to get the response. The potential stimulus x does not have any effect on the response but is itself also a function of y and z, as x is the smaller of these two numbers. See Table 6.4. In terms of prediction, if we knew the values of y and z, we could predict the response perfectly. If we only know x, then we cannot predict the response perfectly. In this case x is really a second response, not a stimulus, though in practice we may not be able to always distinguish between the two. More about stimuli that cause, or are simply associated with, responses will be considered later when we study association and causation.

Table 6.3 Values of a stimulus that can be a whole number between 0 and 5 and the resulting response from a black box.

Stimulus (x)	3	3	0	3	1	2	2
Response	8	7	3	8	4	5	5

Table 6.4 Values of the stimuli and the resulting response from a black box.

Stimulus (x)	3	3	0	3	1	2	2
Stimulus (y)	5	3	3	5	1	3	2
Stimulus (z)	3	4	0	3	3	2	3
Response	8	7	3	8	4	5	5

In the research sciences the process of hypothesizing what stimuli and processes are contained in the black box is called conceptualization. If a researcher wants to understand how a response behaves, they need to conceptualize what stimuli may be acting together to elicit the response from the system. This is a complex process that is never complete or perfect, but researchers attempt to do the best they can. Note that the researchers do not necessarily intend to measure all the stimuli, and indeed many stimuli will never be known. The purpose at this point is to understand the system as well as possible from a conceptual viewpoint.

To begin a more formalized development of conceptualization, we begin by classifying the types of observations that are often used in research studies. The factors or observations that researchers can measure are often classified into three categories (Kaplan 2017). The first classification specifies measurements that can be observed directly.

Definition: Direct Observable

A variable is a *direct observable* if the observation can be measured directly and easily.

Examples of variables that are direct observables include data that can be obtained by visual inspection. For example, the number of applicants who show up on time for a job interview. When meeting with an individual participating in a study, an interviewer may be able to determine gender identity, hair color, skin tone, and the height of the individual. Any data that must be obtained through additional processes, but that can still be observed by the researcher, are known as indirect observables.

Definition: Indirect Observable

A variable is an *indirect observable* if the observation can only be measured through an additional process.

Indirect observables are usually inferred from secondary measurements. Nearly everyone has witnessed a partial solar eclipse, but few have observed it directly. Unless one has specialized eye protection, such as a welder's helmet, the usual method of observing

solar eclipses is through a pinhole apparatus where one observes the light projected onto a flat surface (see Figure **6.3.2**). When using this method, the eclipse is not observed directly; rather, the effect of the eclipse is observed by the shape of the projected light.

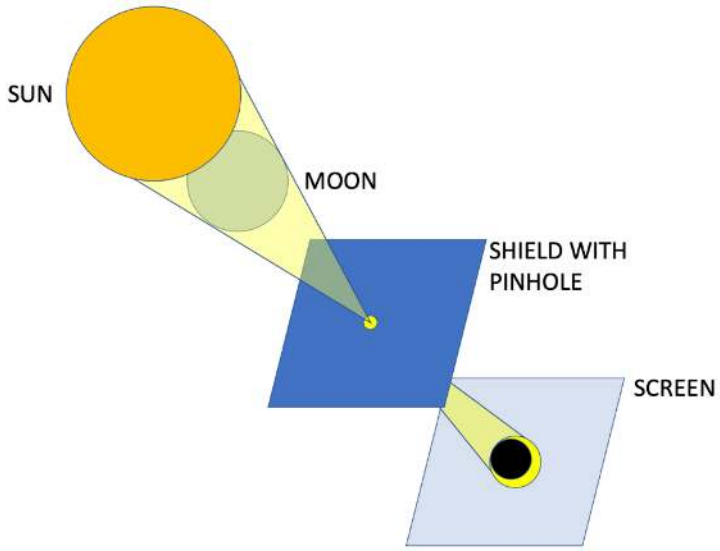


Figure **6.3.2**: Diagram of a simple device for observing a solar eclipse (Public domain image reated by Alan M. Polansky).

Radiocarbon dating provides another example. This method is used to approximate the age of an object that contains organic material. The actual age of the object is not observed but is inferred by the amount of carbon-14 remaining in the object.

In research studies indirect observables may include variables such as gender, ethnicity, and age. For example, gender may be self-identified by an individual, which is directly observed. In other cases, gender identification may be inferred by observing the manner of dress and social interactions of an individual, which is indirectly observed. Note that there is a chance of misidentification with this type of indirect observation. In the social sciences indirect observables refer to any observations taken from secondary material. For example, using birth records to determine the age of individuals in a survey.

Some types of data cannot be observed at all. These types of data are called constructs.

Definition: Construct

Constructs are variables that cannot be observed either directly or indirectly.

Constructs are often abstract ideas whose approximate values may be inferred from direct and indirect observables. We previously considered measuring microaggressions against Asian Americans, and the effect that these microaggressions can have on the victims' health. Both variables, microaggressions and health, cannot be observed. However, as we discussed, both variables can be inferred to some extent from other measurements.

Moving from a construct to more practical matters for research brings us to the idea of a concept. A concept is essentially a construct with the addition of a well-defined meaning. For a construct, this well-defined meaning must be based on direct or indirect observable measurements. The process of moving from the theoretical abstraction of a construct to the well-defined abstract concept is called conceptualization.

Definition: Concept

A *concept* is a construct that has been clearly defined in terms of ideas that can be observed either directly or indirectly. The process of developing a construct into a concept is called *conceptualization*.

For example, the construct of healthiness is meant to convey the theoretical abstract idea of the general health of an individual. The process of conceptualization develops this abstract idea into the concept of a self-evaluated measure of health. This is a concept that is observable. Initially, the concept of microaggressions can be conceptualized to a self-assessment of perceived microaggression encounters over a specified period.

Conceptualization can be a difficult and complex process, and the result can be complex as well. It is important to consider how a construct is conceptualized, and the researchers publishing the results of a study should provide some details about the conceptualization of a construct. When researchers consider complicated concepts, there may be many specific aspects of a concept that are considered. That is, there may be many parts of the concept that need to be considered separately. Each aspect of a concept is called a dimension.

Definition: Dimension

A particular aspect of a concept is called a *dimension*.

If we think about conceptualizing general health, we might consider the health of an individual to have two aspects: physical health and mental health. Each of these aspects is a dimension of health. A researcher would have to decide what indicators there may be for each of these dimensions. For example, physical health might include indicators for high blood pressure, body mass index, alcohol and drug use, and high cholesterol levels. Similarly, indicators of mental health might include depression, anxiety, and self-destructive behavior.

Definition: Indicator

A particular aspect of a dimension is called an *indicator*.

As another example, consider a study that looked for differences between Mexican American and Anglo-American families in their attitudes toward money (Medina et al. 1996). For this purpose, they used the modified money attitude scale (Yamauchi and Templer 1982). The first dimension measures the perceived power and prestige that an individual attaches to money. The second dimension measures how one feels about saving money and making and enacting financial plans. The third dimension measures the amount of anxiety and distrust that money causes is one's life. The fourth dimension measures whether the individual perceives that money is associated with quality.

Within each dimension of a concept, indications about the value of the dimension are likely to exist. This is essentially a list of what characteristics would contribute to a high or low measurement for the dimension. In some cases, indicators consist of yes or no evaluations, so that an individual either has the indicator or they do not. In other cases, indicators are often based on a Likert-type scale.

In the example considering the study that looked for differences between Mexican American and Anglo-American families in their attitudes toward money, four dimensions were identified to measure the concept of attitude toward money. The first dimension measures the perceived power and prestige that an individual attaches to money. This dimension is defined to have eight indicators as shown in Table 6.5.

Table 6.5 The eight indicators of how an individual views power and prestige with respect to money (Medina et al. 1996).

Indicator	Statement
1	I tend to judge people by their money rather than their deeds.
2	I behave as if money were the ultimate symbol of success.
3	I find that I show more respect to those people who possess more money than I do.
4	I own nice things in order to impress others.
5	I purchase things because I know they will impress others.
6	People tell me that I place too much emphasis on the amount of money I have.
7	I enjoy telling people about the money I make.
8	I try to find out if other people make more money than I do.

One can observe from Table 6.5 that the indicators are represented as questions to which a respondent can register an amount of agreement. Indicating more agreement to any of the indicators provides evidence that the individual considers money to be associated with high power and prestige. These statements include how the individual views their own prestige, as well as the prestige of others as associated with wealth. The more indicators that an individual agrees with is an indication that the individual considers those with more wealth to have more power and higher prestige.

The second dimension measures how one feels about saving money and making and enacting financial plans. This dimension is defined to have seven indicators, as shown in Table 6.6. Note that most of these indicators ask whether the individual tends to save money, plan on financial matters, and use a budget. The statements are designed so that the higher scores from the indicators mean the individual considers planning and careful wealth management to be important.

Table 6.6 The seven indicators of how an individual views saving and planning with respect to money (Medina et al. 1996).

Indicator	Statement
1	I put money aside on a regular basis for the future.
2	I do financial planning for the future.
3	I save now to prepare for my old age.
4	I have money available in the event of an economic depression.
5	I follow a careful financial budget.
6	I am prudent with the money I spend.
7	I keep track of my money.

The third dimension corresponds to the amount of anxiety and distrust that money causes is one's life. This dimension is defined to have eleven indicators, as shown in Table 6.7. These indicators tend to assess the individual's feelings about purchasing with an emphasis on the cost of items. The statements are designed so that the higher scores on the indicators mean that monetary transactions tend to cause distrust and anxiety for the individual.

Table 6.7 The eleven indicators of the amount of anxiety and distrust that money causes is one's life (Medina et al. 1996).

Indicator	Statement
1	It bothers me when I discover I could have gotten something for less elsewhere.
2	I complain about the cost of things I buy.
3	I show worrisome behavior when it comes to money.
4	I worry about not being financially secure.
5	When I make a major purchase, I have suspicion that I have been taken advantage of.
6	I show signs of anxiety when I don't have enough money.
7	After buying something I wonder if I could have gotten the same thing elsewhere for less.
8	I hesitate to spend money, even on necessities.
9	It is hard for me to pass up a bargain.
10	I automatically say “I cannot afford it.”
11	I am bothered when I have to pass up a sale.

The fourth dimension measures whether the individual perceives that money is associated with quality. This dimension is defined to have five indicators, as shown in Table 6.8. These statements measure the willingness of the individual to spend money to gain the associated perceived quality. Higher values of the indicators indicate that the individual considers high cost and good quality to be synonymous.

Table 6.8 The five indicators of whether the individual perceives that money is associated with quality (Medina et al. 1996).

Indicator	Statement
1	I am willing to spend more to get the very best.
2	I buy top of the line products.
3	I buy name brand products.
4	I pay more for some things because I know I have to in order to get the best.
5	I buy the most expensive items available.

The final phase in developing methods for gathering information about a construct or a concept is operationalization, where the researcher develops the procedures that will result in the empirical observations that will be used in the study.

Definition: Operationalization

Operationalization is the process of developing the procedures that will result in the empirical observations representing the concepts.

Through the process of operationalization, the researcher will decide how to score the responses from a survey, what the range of the observations will be, and what level of measurement will be used. For concepts with multiple dimensions and multiple indicators, the researcher must make important decisions on how a composite measure of the concept will be computed from the individual indicators and dimensions.

Returning to the study of attitudes toward money, the operationalization of the four dimensions and their associated indicators are taken together to form what is known as the Money Attitude Scale. The implementation of this schedule relies on everyone responding to each indicator statement using a Likert-type scale with a numerical range from 1 to 7, where 1 means the behavior "never applies" and 7 means it "always applies." The midpoint value of 4 indicates that the behavior sometimes applies to the individual. The responses from each of the dimensions are summed to get a total score for that dimension. These values indicate how strongly each dimension applies to the individual, and the total score from all the dimensions indicates the relative importance of money for the associated individual.

This page titled 6.3: Conceptualization and Operationalization is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.4: Validity and Reliability

Once a measurement system has been developed, it is crucial that it be evaluated for validity and reliability. This is important because researchers need to know if the measurement instrument works well for the concept and provides stable measurements. It is often helpful to think about abstract concepts in terms something familiar, and hence we will discuss these two concepts in terms of a measurement that most people have done themselves.

Consider the four types of rulers shown in Figure **6.4.1**. All four of the rulers have been designed to measure one foot, or twelve inches, but they have been made in different ways. The first ruler is very carefully made of stainless steel and has been marked so that the distance between the first and last marks is exactly one foot. The second ruler is also made of stainless steel but was carelessly marked, so the distance between the first and last marks is one foot and two inches—even though it is labeled as being exactly one foot. The third ruler is made of a stretchable material. When the ruler was in a relaxed state, it was marked carefully so that the distance between the first and last marks is exactly one foot. The problem is that the material is very stretchable, so when someone uses the ruler, they will sometimes expand it so that the distance between the first and last marks is greater than one foot, while at other times the ruler will get compressed, and the distance between the first and last marks shrinks to less than one foot. The fourth ruler is also made of a stretchable material. When the ruler is in a relaxed state, it is evident that it was not marked carefully because the distance between the first and last marks is one foot and two inches, even though it is labeled as exactly one foot.

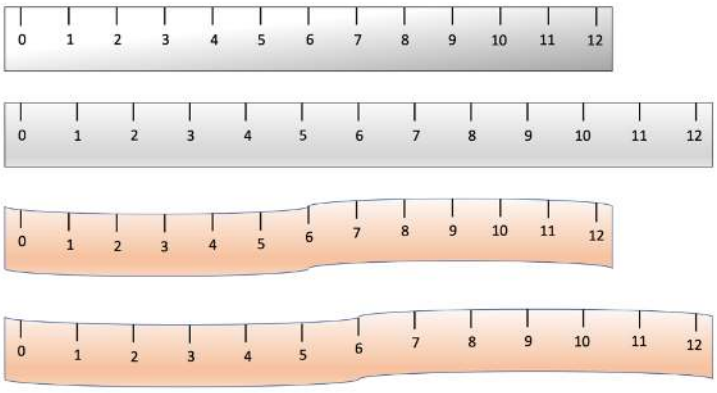


Figure **6.4.1**: Visualizations of the four rulers used in the discussion of validity and reliability (Public domain image created by Alan M. Polansky)

Now let us consider how measurements from these rulers would behave in practice. Suppose that twenty people are given a wooden block that is exactly 12 inches long. Four groups of five people will each be given one of the rulers. Everyone in each group will measure the block. This experiment is done in such a way that nobody measures the block more than once. Starting with Ruler 1, we have an instrument that is correctly marked and made of a material that is very stable. When the five people measure the block, we would expect that each of them would get measurements very close to twelve inches, and the measurements might look something like what is given in the first row of Table 6.9. It should be noted that even with a very stable and correctly marked ruler that different people would get slightly different measurements, even using the same ruler, and the same block of wood.

Table 6.9 Hypothetical Measurements of a twelve-inch block using the four rulers in Figure 6.3.

Ruler	Measurements				
1	12.0	12.1	12.2	12.0	12.1
2	10.1	10.4	10.0	10.2	10.1
3	13.0	11.9	11.3	12.2	11.5
4	10.1	10.1	10.1	10.6	10.3

Next consider Ruler 2, where we have an instrument that is incorrectly marked and made of a material that is very stable. When the five people measure the block, we would expect that they would all get measurements very close to 10.25 inches, which is the value on Ruler 2 corresponding to a true length that is incorrectly marked as twelve inches on Ruler 2. These measurements might look something like what is given in the second row of Table 6.9. Considering Ruler 3, we have an instrument that is correctly marked and made of a material that is very unstable. When that group measures the block, we would expect that each person would get measurements that vary around twelve inches, but they might be farther away from twelve inches compared to the Ruler 1 group's measurements, as some people might stretch the ruler out or compress it when they measure the block. These measurements might look something like what is given in the third row of Table 6.9. Finally, Ruler 4, where we have an instrument that is incorrectly marked and made of a material that is very unstable: when the five people measure the block, we would expect them to get measurements that vary around 10.25 inches, but they might be farther away from 10.25 inches than the second ruler. These measurements might look something like what is given in the last row of Table 6.9.

In terms validity and reliability, whether the ruler is marked correctly determines the validity of the measurements from the ruler. That is, the ruler is marked correctly so that the twelve-inch mark corresponds to twelve inches. The reliability of the rule refers to whether the measurements are stable. The first two rulers, which are made of the stable material, are reliable. That is, repeated measurements with these rulers will give very similar results. Therefore, the first ruler is both valid and reliable, the second ruler is not valid but is reliable, the third ruler is valid but not reliable, and the fourth ruler is nether valid nor reliable. The best measurements will come from the ruler that is both valid and reliable, which corresponds to Ruler 1. That is, the ruler is marked correctly, and repeated measurements are very stable.

Researchers are often interested in measuring an abstract concept such as anxiety, security, satisfaction, and emotional stability. For these concepts the validity of a measurement instrument, as with the rulers, refers to the idea that the concept that the system is measuring matches the concept that they are attempting to measure.

Definition: Valid Measurement System

A measurement system is *valid* if the system correctly measures the concept without any systematic bias.

What makes thinking about validity difficult for abstract concepts such as the ones that are common in research is that we usually do not have a good concrete understanding of what is being measured. From our viewpoint it suffices to imagine the construct and that a valid measurement system would act much like Ruler 1 in the example. That is, while not all the measurements would exactly give us the correct measurement, the measurements would tend to cluster around the correct measurement.

In practice the validation process is quite complicated and technical; we will cover some of the more pertinent issues below without the technical details. Validity has many facets, and we will discuss several here. The first is face validity, which essentially refers to whether the dimensions and the indicators are logical and are consistent with known research about the subject.

Definition: Face Validity

A measurement system has *face validity* if the dimensions and the indicators make sense and are consistent with known research about the subject.

For example, it would be widely accepted to include blood pressure as an indicator of physical health. This is consistent with the known research that links conditions like hypertension to increased mortality rates. However, a measure like whether an individual prefers to listen to podcasts or music may have little to do with physical health, and there would probably be better indicators, in any case.

The second type of validity is called content validity, which is the degree to which the measurement system covers the range of meanings within the concept.

Definition: Content Validity

A measurement system has *content validity* if the dimensions and indicators cover the range of meanings within the concept.

As with face validity, content validity is largely based on the collective judgement and experience in the research community. In the previous discussion of general health, it was pointed out that such a measure should include components corresponding to both physical health and mental health.

A stronger type of validity is called *criterion validity*, where the measurement system is statistically compared to another indicator of measure of the concept.

Definition: Criterion Validity

A measurement system has *criterion validity* if it coincides statistically with other measures of the same concept.

For the self-reported health measure considered earlier, the researchers reported that the measure was compared to other measures of health, including mortality. The fact that the self-reported health measure was associated with mortality rates provides evidence of criterion validity for the measure.

There are two types of criterion validity that can be considered. The first is *predictive criterion validity*, which implies that the measurement system does a good job of predicting future values of an established measure of the concept.

Definition: Criterion Validity

A measurement system has *predictive criterion validity* if it does a good job of predicting future values of an established measure of a concept.

Note the time precedence inherent in the definition. The value of the measurement system must be observed first. The measurement system then predicts the value of the established measure that is observed later in time. In the case of the self-reported health assessment, this would mean that the self-reported measures of health would be observed, and then the mortality rate of those same individuals would be observed later. If the self-reported health assessment does a good job of predicting the mortality rate, then the measure would have predictive criterion validity.

Predictive criterion validity is a very powerful concept but is not always possible to implement in practice. For example, in the case of validating the self-reported health measure with mortality rates, several years may pass after individuals report their health until enough mortality data is available to verify that the self-reported health predictions are accurate. Researchers may not have sufficient time to wait for these results before publishing their results. The predictive nature of the criterion also implies that the same individuals need to be tracked during this period, which may require substantial resources. An alternative method of validation is based on measurements taken at the same time.

Definition: Concurrent Criterion Validity

A measurement system has *concurrent criterion validity* if it is strongly associated with a value of an established measure of a concept that is observed at the same time.

In the case of the self-reported health measure, one could validate this measure by having individuals come to clinic and ask them about their health, and then during the same visit assess their general health medically.

A slightly more complex type of validity is *construct validity*, which assesses whether the measure behaves the way researchers would expect with respect to measures of other concepts.

Definition: Construct Validity

A measurement system has *construct validity* if it behaves the way researchers would expect in how it is related to measures of other concepts.

For example, it may be known that individuals with lower overall general health have more doctor visits per year. If the self-reported measure of general health is valid, then lower measures of that measurement system should also be associated with more doctor visits per year. A lack of a relationship could also be used for validation purposes. If it is known the general health is not associated with the amount of time one spends online, then one should expect that the self-reported measure of general health would also not be associated with the amount of time an individual spends online.

The final type of validity is called *factor validity* and relies on a technical statistical technique called *factor analysis*. Factor validity determines whether the dimensions and indicators are consistent in practice with what the researcher intended. That is, the proposed measurement system is observed on a group of individuals. The factor analysis technique can then be used to determine whether the number of dimensions appears to be correct along with whether the indicators are grouped under the correct dimensions.

Definition: Factor Construct Validity

A measurement system has *factor construct validity* if a statistical factor analysis indicates that the number of dimensions and the grouping of the indicators within the dimensions is correct.

If a researcher contends that four dimensions should be used to measure general health where each dimension has a specified number of indicators, the factor analysis should give results consistent with the proposed structure.

Definition: Reliable Measurement System

A measurement system is *reliable* if repeated measurements on the same individual are nearly identical.

Essentially, this is the difference between the stainless-steel ruler, which was very reliable as repeated measurements on the same object would be nearly the same, and the flexible ruler, which was not as reliable as repeated measurements on the same object would not be nearly the same.

At first it may seem that a measurement system should always give the same result when it is applied to the same object. But imagine what might happen if a simple measurement scale like the self-reported health scale is given to the same person on two consecutive days. On one day the individual might be feeling well with a good outlook on life, and they rate their health as being good. On the next day they might not be feeling well and, in that case, they might rate their health as fair or even bad. Reliability then is an indication of how much a researcher can trust the measurement taken at a certain time. Reliability provides the researcher with the assurance that if the measure were taken again, a very similar result would be obtained.

Researchers have several methods that can be used to assess the reliability of a measurement system. In the simplest case the spirit of the definition of reliability can be used directly and the same measurement can be taken two or more times on the same group of individuals to assess reliability. This is known as test-retest reliability.

Definition: Test-Retest Reliability

The reliability of a measurement system is assessed using *test-retest reliability* if the same measurement is taken two or more times on the same group of individuals.

For some measurements, such as medical measurements like weight and blood pressure, this method can be a simple, direct method for assessing reliability. For measurement systems based on surveys or written tests, this method can give biased results and can often indicate that a measurement system is more reliable than it is in reality. For example, individuals probably remember most of their answers on a survey when they retake it a second time, and to save time might repeat those answers instead of indicating their true feelings.

A method for assessing the reliability of a measurement system that does not require more than one measurement on a group of individuals is called internal consistency reliability.

Definition: Internal Consistency Reliability

The reliability of a measurement system is assessed using *internal consistency reliability* if the association between the dimensions and the indicators within the dimension are all strongly associated with one another.

It should be apparent from the definition that the internal consistency of a measure can only be assessed if there are at least two dimensions or indicators. The idea behind internal consistency is that if all the indicators tend to provide very similar information, then the final measurement should be very stable. For example, when one indicator acts very differently from another, it is an indication that a small change in the indicators could result in a large change in the measurement. If all the indicators tend to act together, it would be very rare for a large change in one of the indicators to occur because this would imply that many indicators would also have to change. Internal consistency reliability is assessed using statistical techniques based on correlations, which will be discussed later in this book.

This page titled 6.4: Validity and Reliability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.5: Established Measurement Systems

The development of valid and reliable measurement systems is a nontrivial task. In fact, many researchers may solely publish research on developing and evaluating measurement systems for certain situations. These measurement systems can then be used by the research community for research into specific problems. For both the researcher and those evaluating the research, the use of established measurement systems can be convenient. From the perspective of a researcher, it means they can use an established measurement system that has already gone through the conceptualization and operationalization processes and has been evaluated for reliability and validity. From the perspective of someone who is critically evaluating a research study it means the measurement system has been critically evaluated within the research community.

The main concern with evaluating research based on an established measurement system is that the researchers need to be using the measurement system correctly. Any established measurement system was developed and validated to measure a very specific concept under very specific conditions. This means one should check either with the original research source or other established sources to ensure that the measurement system is being implemented, evaluated, and interpreted correctly. For example, a measure of mental health should not be used to measure general health or physical health.

There are many standard sources for the measurement systems. For example, the American Psychiatric Association publishes the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association 2013). This manual contains the standard measurement systems that are used by psychiatrists for diagnosing mental disorders, including depression and anxiety-related disorders. While researchers may use these measurement systems on their own, these types of conditions should only be professional assessed by a clinician. Hence, if a researcher is using questions related to assessing depression, the results of a survey may indicate symptoms of depression but cannot on their own be used to diagnose the individuals participating in a study. Similarly, the American Psychological Association provides online access to psychological measurement systems, though membership in the association is required.

Some measurement systems have been developed and are available from government organizations. For example, the United States Centers for Disease Control and Prevention published *Measuring Violence-Related Attitudes, Behaviors, and Influences Among Youths: A Compendium of Assessment Tools* (Dahlberg et al. 2005). This compendium provides researchers with a set of 170 measurement systems to assess violence-related beliefs, behaviors, and influences in individuals. All these measures have been validated and are considered reliable.

This page titled 6.5: Established Measurement Systems is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.6: Critical Analysis

Many research studies measure abstract concepts using the process of conceptualization and operationalization. Some studies develop their own measurement methods and include conceptualization and operationalization as part of the research project. For these studies one should carefully examine the dimensions and the indicators to obtain a clear understanding of how the construct was conceptualized and operationalized. Familiarity with how the measurement system was developed will aid in understanding how the concept is measured and how the corresponding results can be interpreted. Furthermore, when assessing a measurement system, one should critically think about whether the dimensions and indicators make sense for what the researchers are attempting to measure. As someone critically evaluating research, you should consider whether the dimensions and indicators are consistent with what you would consider to be facets of the concept and determine if you feel that some important issues are left out of the dimensions or indicators. The authors should provide some justification for their conceptualization of the construct, either based on expert opinion or previous research.

Once conceptualization and operationalization have been considered, the measurement system should be evaluated in terms of validity and reliability. Note that you may not be familiar with the technical details of how a particular measurement system is evaluated in terms of these two issues, but there should be some statement in the research that states that this process has been done, and what limitations were revealed. You can also consider what method was used to assess both the validity and reliability of the measurement system. For example, construct validity and criterion validity are usually more powerful than face and content validity. The best measurement systems have been assessed using more than one method. In any case it should always be remembered that researchers must assess both validity and reliability to argue that the measurement system can be used effectively. The results of the study should be evaluated with these limitations in mind.

A study that used an established measurement scale has added legitimacy to the research results. However, one must ensure that the authors of the research employed the measurement system as it was intended, and that the same construct was measured.

In summary, the key questions in critically analyzing measuring abstract information in a study are:

1. What dimensions are used to conceptualize the construct? Do these dimensions make sense in terms of what the researchers are attempting to measure?
2. What are the indicators within each dimension? Do these indicators make sense in terms of what the researchers are attempting to measure with each dimension?
3. How is the measure operationalized? What is the measurement scale? What is the range of the measurements?
4. Has the validity of the measurement system been evaluated? What methods were used to validate the measurement system? Were any limitations revealed?
5. Has the reliability of the measurement system been evaluated? What methods were used to study the reliability of the measurement system? Were any limitations revealed?
6. If an established measurement system is being used, are the authors of the research using the measurement system in the way it was intended? Is the measurement system measuring the same construct as the authors of the researchers intend to measure?

This page titled 6.6: Critical Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.6.1: Attitudes About Marital Rape

Previously we considered an internet survey that was conducted on 85 undergraduate students and 44 alumni from a small liberal arts college (Ferro et al. 2008). The purpose of the study was to determine if the relationship between the victim and the offender influenced how the participants rated a scenario described in the narrative in terms of commons myths about rape.

A rape myth is defined as “prejudicial, stereotyped, or false beliefs about rape, rape victims, and rapists” (Burt 1980). Most rape myths relate to denying the culpability of the perpetrator and undermining the experiences of the victim (Barber 1974; Whatley 1996). It has also been noted that traditional rape myths exist to justify sexual aggression toward women (Lonsway and Fitzgerald 1994).

The narratives were rated by the participants using three measurement systems. The first measurement system consisted of parts of the Illinois Rape Myth Acceptance Scale, the second measurement system was the Rape-Supportive Attributions Scale, and the third measurement system was the Sex-Role Stereotypical Victim Attributions Scale.

The Illinois Rape Myth Acceptance Scale was developed to improve on an existing rape myth measurement system through a reconceptualization of rape myths using research from psychology, sociology, anthropology, and philosophy (Payne et al. 1999; Lonsway and Fitzgerald 1994, 1995).

This research led to the definition that rape myths are “attitudes and beliefs that are generally false but are widely and persistently held, and that serve to deny and justify male sexual aggression against women” (Lonsway and Fitzgerald 1994). Using this conceptualization, a measurement system based on 20 indicators was developed to assess how much an individual believes in rape myths. The measurement system has been categorized into seven dimensions or sub-scales. In the study the overall score from the 20 indicators was used along with three of the dimensions. The three dimensions used were: It Wasn’t Really Rape, which assesses the conditions the individual feels need to be met for a situation to be considered a rape; He Didn't Mean To, which assesses the level of blame the individual attributes to the offender; and Rape is a Deviant Event, which assesses how the individual feels about the deviant nature of rape. The dimensions and indicators are given in Table 6.10. The answers to each of the questions were reported using a seven-point Likert scale ranging from 1 (not at all agree) to 7 (very much agree). The measurement system also includes several “filler questions,” which are meant to reduce bias in the responses from the participating individuals. The Illinois Rape Myth Acceptance Scale has gone through extensive validation, which included six studies using empirical observations using real individuals. These studies established construct validity, factor validity, and internal consistency reliability.

Table 6.10 The seven dimensions and indicators used in the Illinois Rape Myth Acceptance Scale (Payne et al. 1999).

Dimension 1: She asked for it	
1.	If a woman is raped while she is drunk, she is at least somewhat responsible for letting things get out of control.
2.	A woman who dresses in skimpy clothes should not be surprised if a man tries to force her into having sex.
3.	A woman who “teases” men deserves anything that might happen.
4.	When a woman is raped, it's often because the way they said no was ambiguous.
Dimension 2: He didn’t mean to	
1.	Men usually don't intend to force sex on a woman, but sometimes they get too sexually carried away.
2.	Rape happens when a man’s sex drive goes out of control.
Dimension 3: It wasn’t really rape	
1.	If a woman is raped and doesn't physically fight back, you can't really say that it was rape.
2.	If the rapist doesn’t have a weapon, you really can’t call it a rape.
Dimension 4: She wanted it	
1.	Although most women wouldn't admit it, they generally find being physically forced in sex a real “turn on.”
2.	Many women secretly desire to be raped.
Dimension 5: Rape is a trivial event	
1.	If a woman is willing to “make out" with a guy, then it's no big deal if he goes a little further and has sex.
2.	Women tend to exaggerate how much rape affects them.
Dimension 6: She lied	
1.	Rape accusations are often used as a way for getting back at men.
2.	A lot of women lead a man on and the cry rape.
Dimension 7: Rape is a deviant event	
1.	It is usually only women who dress suggestively who are raped.
2.	Men from nice middle-class homes almost never rape.
3.	Rape is unlikely to happen in a woman's own neighborhood.
Filler Questions	
1.	Most rapists are not caught by the police.
2.	All women should have access to self-defense classes.
3.	It is preferable that a female police officer conduct the questioning when a woman reports rape.

The Rape-Supportive Attributions Scale is a measurement system consisting of four questions that assess an individual's false beliefs about rape (Monson et al. 2000; Monson et al. 1996; Bridges 1991). The measurement system was developed specifically for the situation described in the narratives used in this study. The indicators are given in Table 6.11, where it should be noted that in the original research the perpetrator was named “Kevin” and the victim was named “Jenny” instead of using the identifiers perpetrator and victim. This was done to reduce bias. The Sex-Role Stereotypical Victim Attributions Scale is a measurement system consisting of four questions that also assesses an individual's false beliefs about rape (Monson et al. 2000; Monson et al. 1996; Bridges 1991). The indicators are given in Table 6.12. This measurement system comes from the same research as the Rape-Supportive Attributions Scale. Individual responses were recorded for each of these scales using a 7-point Likert-type scale ranging from 1 (minimal) to 7 (extremely). The research mentions conceptualization and internal consistency reliability, but validity is not mentioned in these sources.

Table 6.11 The Rape-Supportive Attributions Scale (Monson 2000)

1.	How violent do you feel this situation was?
2.	How psychologically damaged do you feel “Jenny” will be from this experience?
3.	To what degree were “Kevin’s” actions a violation of “Jenny’s” rights?
4.	How certain are you that this incident would be considered rape?

Table 6.12 The Sex-Role Stereotypical Victim Attributions Scale (Monson 2000)

1.	How much control did “Jenny” have in this situation?
2.	How much did “Jenny” enjoy this situation?
3.	How obligated was “Jenny” to engage in sexual relations in this case?
4.	How interested was “Jenny” in having sexual relations?

The measurement systems used for this research project appear to have gone through many of the processes required to establish a well conceptualized, valid, and reliable measurement system. In particular, the Illinois Rape Myth Acceptance Scale has been extensively tested and has become a standard assessment tool in this area. The lack of validation for the remaining two measurement systems is less of an issue as these two systems are used less extensively in the research, and it is possible that validation is addressed in research not cited within the research project.

This page titled 6.6.1: Attitudes About Marital Rape is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.6.2: Gender and Race-Related Stress

Research on the impact of racism on the general health of African Americans has consistently shown that racial discrimination is associated with depression, substance abuse, anxiety, and distress (Jones et al. 2007; Karlsen and Nazroo 2002; Taylor and Turner 2002; Williams et al. 2003). Research studies have also considered possible factors that can moderate these issues (Fischer and Shaw 1999; Jones et al. 2007). That is, there may be other factors in the identity of an individual that may lessen the impact of race-related stress. One such research study considered the role that gender may have on such stress (Greer et al. 2009).

The study consisted of a survey of 183 individuals who were self-identified African American students at a mid-sized university. Of these 183 individuals, 121 identified themselves as women, and the remaining 62 identified themselves as men. Individual experiences of race-related stress were measured using the Index of Race Related Stress: Brief Version (Utsey 1999). This scale is based on a conceptualization of race-related stress based on individual racism, cultural racism, and institutional racism. Individual racism refers to racial experiences occurring on a personal level. Cultural racism refers to experiences of inferiority of African American culture. Institutional racism refers to experiences of racial inequalities in the form of practices and policies of societal institutions. Hence, this measurement instrument has three dimensions. Individual racism has six indicators, cultural racism has ten indicators, and institutional racism has six indicators. The statements associated with these indicators are given in Table 6.13. Everyone scored each statement using the five-point Likert scale detailed in Table 6.14.

Table 6.13 Index of Race-Related Stress: Brief Version (Utsey 1999).

Dimension 1: Individual Racism	
1.	Sales people and clerks did not say thank you or show other forms of courtesy and respect when you shopped at some White/non-Black owned businesses.
2.	While shopping at a store, or when attempting to make a purchase, you were ignored as if you were not a serious customer or didn't have any money.
3.	While shopping at a store the sales clerk assumed that you couldn't afford certain items.
4.	White people or other non-Blacks have treated you as if you were unintelligent and needed things explained to you slowly or numerous times.
5.	You were the victim of a crime and the police treated you as if you should just accept it as part of being Black.
6.	Whites/non-Blacks have stared at you as if you didn't belong in the same place with them; whether it was a restaurant, theater, or other place of business.
Dimension 2: Cultural Racism	
1.	You notice that crimes committed by White people tend to be romanticized, whereas the same crime committed by a Black person is portrayed as savagery, and the Black person who committed it, as an animal.
2.	You have observed situations where other Blacks were treated harshly or unfairly by Whites/non-Blacks due to their race.
3.	You have observed the police treat Whites/non-Blacks with more respect and dignity than they do Blacks.
4.	You notice that the media plays up those stories that cast Blacks in negative ways, usually accompanied by a large picture of a Black person looking angry or disturbed.
5.	You seldom hear or read anything positive about Black people on radio, TV, in newspapers, or history books.
6.	You have been given more work, or the most undesirable jobs at your place of employment while the White/non-Black of equal or less seniority and credentials is given less work, and more desirable tasks.
7.	You have observed that White kids who commit violent crimes are portrayed as “boys being boys,” while Black kids who commit similar crimes are wild animals.
8.	You have heard or seen other Black people express a desire to be White or to have White physical characteristics because they disliked being Black or thought it was ugly.
9.	You notice that when Black people are killed by the police, the media informs the public of the victims criminal record or negative

	information in their background, suggesting they got what they deserved.
10.	You have heard reports of White people/non-Blacks who have committed crimes, and in an effort to cover up their deeds falsely reported that a Black man was responsible for the crime.
Dimension 3: Institutional Racism	
1.	You have been threatened with physical violence by an individual or group of Whites/non-Blacks.
2.	You were treated with less respect and courtesy than Whites and other non-Blacks while in a store, restaurant, or other business establishment.
3.	You were passed over for an important project although you were more qualified and competent than the White/non-Black person given the task.
4.	You have been subjected to racist jokes by Whites/non-Blacks in positions of authority and you did not protest for fear they might have held it against you.
5.	You have heard racist remarks or comments about Black people spoke with impunity by White public officials or other influential White people.
6.	You were refused an apartment or other housing; you suspect it was because you're Black.

Table 6.14. The Likert scale used to score the responses for the Index of Race-Related Stress: Brief Version.

0	Event never happened to me.
1	Event happened but did not bother me.
2	Event happened and I was slightly upset.
3	Event happened and I was upset.
4	Event happened and I extremely upset.

The Index of Race Related Stress: Brief Version has been extensively conceptualized, validated, and studied in terms of reliability (Utsey 1999). This measurement system is based on a reanalysis of a longer measurement scale with respect to factor validity. The reanalysis revealed a smaller form of the system that coincided well with established conceptualizations of race-related stress. The new measurement system was also assessed based on criterion validity, and the reliability was verified using internal consistency reliability.

The mental health of the individuals in the study was assessed using the Hopkins Symptom Checklist (Derogatis et al. 1974). There are 58 indicators based on self-reporting that measure several dimensions of psychiatric symptoms. There are five dimensions that specifically measure anxiety (6 items), somatization (12 items), obsessive compulsiveness (8 items), interpersonal sensitivity (7 items), and depression (11 items). The indicators were answered using a four-point Likert scale with values 0 (not at all distressed), 1 (a little distressed), 2 (quite a bit distressed), 3 (extremely distressed). Larger scores for each of these dimensions indicate higher degrees of psychological distress. The measurement system is too extensive to describe here.

This measurement system is a standard instrument used in psychology and psychiatry. The original research paper reported that a series of studies were used to establish the factorial and criterion validity and the reliability of the measure. The internal consistency reliability was also established by the authors.

This page titled 6.6.2: Gender and Race-Related Stress is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.6.3: Measuring Satisfaction in Sports

We have previously considered a research study that attempted to determine what factors might contribute to an athlete having satisfaction in their sport (Lepir and Lakić 2025). This study used three separate measurement scales to measure athletic satisfaction, personality traits, and emotional skill and competence. We will consider the development of the measurement scale for athletic satisfaction.

To measure athletic satisfaction the researchers in the study used the Athletic Satisfaction Questionnaire (Riemer and Chelladurai 1998). The Athletic Satisfaction Questionnaire is a 15-dimension, 56-item questionnaire that is designed to measure the important aspects of athletic satisfaction based on previous research about the important aspects of athlete satisfaction (Chelladurai and Riemer 1997). The researchers define satisfaction as “a positive affective state resulting from a complex evaluation of the structures, processes, and outcomes associated with the athletic experience” (Chelladurai and Riemer 1997, p. 135).

Using psychological theories of satisfaction, the researchers identified 33 different facets in seven different categories as a starting place for developing the questionnaire. These facets included such concepts as personal growth, team goal attainment, team contributions, and family support. Through an initial phase of development some of the facets were combined, and an initial version of the questionnaire based on 29 facets measured using 197 questions was developed. This initial version of the questionnaire was then tested for readability and was judged by a panel of experts. The questionnaire was then given to 20 undergraduate sport psychology majors who were asked to match each question with a randomized list of the 29 facets to determine if the questions were capturing the correct ideas behind each of the facets. Finally, the preliminary version of the questionnaire was administered to 172 first-year undergraduate sport management and physical education majors. These data were subjected to several types of statistical analyses. These three analyses reduced the questionnaire to 121 questions on 19 facets.

The final phase of development was based on administering the edited questionnaire to a sample of 614 Canadian university athletes in basketball, hockey, and volleyball. Using this data the researchers were able to verify the construct validity and reliability of the measurement system, which is based on a 15-dimension, 56-item questionnaire (summarized in Table 6.15).

Table 6.15. The dimensions and items in the *Athletic Satisfaction Questionnaire* (Riemer and Chelladurai 1998).

Dimension 1: Individual Performance	
1.	The degree of which I have reached my performance goals during the season.
2.	The improvement of my performance over the previous season.
3.	The improvement in my skill level thus far.
Dimension 2: Team Performance	
1.	The team's win/loss record this season.
2.	The team's overall performance this season.
3.	The extent to which the team has met its goals for the season thus far.
Dimension 3: Ability Utilization	
1.	The degree to which my abilities are used.
2.	The level to which my talents are employed.
3.	The extent to which my role matches my potential.
4.	The amount of time I play during competitions.
5.	The degree to which my role on the team matches my preferred role.
Dimension 4: Strategy	
1.	The tactics used during games.
2.	Coach's choice of strategies during games.
3.	How the coach makes adjustments during competitions.
4.	Coach's game plans.
5.	The manner in which coach combines the available talent.
Dimension 5: Personal Treatment	
1.	The recognition I receive from my coach.
2.	The friendliness of the coach towards me.
3.	The level of appreciation my coach shows when I do well.
4.	My coach’s loyalty towards me.
5.	The extent to which my coach is behind me.
Dimension 6: Training and Instruction	
1.	The training I receive from the coach during the season.
2.	The instruction I have received from the coach this season.
3.	The coach's teaching of the tactics and techniques of my position.
Dimension 7: Team Task Contribution	
1.	The extent to which teammates provide me with instruction.

2.	The guidance I receive from my teammates.
3.	The constructive feedback I receive from my teammates.
Dimension 8: Team Social Contribution	
1.	My social status on the team.
2.	The role I play in the social life of the team.
3.	The degree to which my teammates accept me on a social level.
Dimension 9: Ethics	
1.	The extent to which all team members are ethical.
2.	My teammates' sense of fair play.
3.	My teammates' "sportsmanlike" behavior.
Dimension 10: Team Integration	
1.	How the team works to be the best.
2.	The degree to which teammates share the same goal.
3.	Team member's dedication to work together toward team goals.
4.	The extent to which teammates play as a team.
Dimension 11: Personal Dedication	
1.	The degree to which I do my best for the team.
2.	My dedication during practices.
3.	My enthusiasm during competitions.
4.	My commitment to the team.
Dimension 12: Budget	
1.	The funding provided to my team.
2.	The amount of money spent on my team.
3.	The fairness of the team's budget.
Dimension 13: Medical Personnel	
1.	The competence of the medical personnel.
2.	The fairness with which the medical personnel treat all players.
3.	The medical personnel's interest in the athletes.
4.	The promptness of medical attention.
Dimension 14: Academic Support Services	
1.	The tutoring I received.
2.	The academic services provided.
3.	The personnel of the academic support services (e.g., tutors, counselors).
Dimension 15: External Agents	

1.	The media’s support of our program.
2.	The support from the university community.
3.	The supportiveness of the fans.
4.	The local community’s support.

This page titled 6.6.3: Measuring Satisfaction in Sports is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.7: Summary

Research studies are often based on attempting to measure abstract types of information that cannot be directly observed. To measure these abstract types of information, researchers develop indirect methods for observing the information. This development follows a process of conceptualization where the researcher assesses, based on their own professional opinions and established research, what behaviors or indications are associated with the measure, developing broad categories known as dimensions and specific indications known as indicators. Once conceptualization has been completed, the measurement system is made practical through the process of operationalization within which the researcher decides exactly how the concept will be measured and what type of data will result. Once the measure has been operationalized, a researcher must verify that the measure is valid and measures the concept that it is supposed to measure, and that it is reliable—the measure does not vary to a great degree when repeatedly measuring the same subject or individual. Because of the difficulties in establishing a new measurement system, many researchers take advantage of standard ones that have been established by other researchers. This ensures that the measurement system has been thoroughly researched and, in many cases, has been used in many other studies by the research community.

This page titled 6.7: Summary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.8: About the Quotation

Franz Kafka (1883–1924) was a Jewish Czech writer born in Prague whose writing blended expressionism, realism, and surrealism to explores themes of alienation, loneliness, and the search for meaning in an indifferent world (Figure **6.8.1**). Kafka is best known novels are *The Trial* and *The Metamorphosis*. The quotation is taken from a book collecting some of Franz Kafka's letters to Milena Jesenská who was a was a Czech journalist, writer, editor and translator.



Figure **6.8.1**: The last known photograph of Franz Kafka (1883–1924) the author of *The Trial* and *The Metamorphosis* (public domain image).

This page titled 6.8: About the Quotation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.9: Practice and Exploration

This page titled 6.9: Practice and Exploration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.10: Collaborative Activity

This page titled 6.10: Collaborative Activity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

6.11: References

This page titled 6.11: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

CHAPTER OVERVIEW

7: Probability and Randomness

“So much of life, it seems to me, is determined by pure randomness.”

Sidney Poitier

7.1: What to Expect
7.2: Randomness
7.3: Defining Randomness
7.4: Probability
7.5: Odds
7.6: Critical Analysis
7.6.1: The Birthday Problem
7.6.2: Risk of Death from a Shark Attack
7.6.3: Bias in the Application of the Death Penalty
7.6.4: The Vietnam War Draft Lottery
7.6.5: Ovid and Gambling
7.6.6: Sports Activities and Lower Back Pain in Adolescence
7.7: Summary
7.8: About the Quotation
7.9: Practice and Exploration
7.10: Collaborative Activity
7.11: References

This page titled 7: Probability and Randomness is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.1: What to Expect

In this chapter, we consider the concepts of randomness and probability. At first, it may seem out of place to be learning about probability and randomness when we have been considering how the scientific method applies to research studies. Earlier we learned about defining populations in practical ways and about the data that are observed from these populations. However, we have not addressed what methodology is used to determine how data is observed from populations. As we have discussed before, observing an entire population is usually impractical due to time and budget constraints. Using a logical process of called *induction*, we usually only observe a part of a population and base our conclusions on the observed part. For this process to be effective, the method which determines which part of the population is to be observed is usually based on methods that rely on randomness. These random methods make it possible for researchers to calculate specific rates at which they make errors in their estimates and conclusions. Because these methods are based on randomness, an understanding of what randomness is and how probability is used to measure how often random events occur is important to an understanding of the practical application of the scientific method using statistics. This chapter will introduce the concept of randomness and the associated measure of the likelihood of random occurrences as a basis for considering these concepts later. By the end of this chapter, you will be able to:

1. Discuss randomness in terms of the concept of absolute certainty.
2. Interpret likelihoods that are reported in terms of probabilities, chances, or odds.
3. Compute probabilities in simple situations.
4. Compute and interpret odds and odds ratios.

This page titled 7.1: What to Expect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.2: Randomness

Randomness is usually defined as a lack of a definite pattern or predictability in the outcomes of an observable process. We will discuss a more mathematical and precise definition of randomness later in this chapter. For now, we will simply discuss how randomness is an important part of our lives. Randomness is encountered throughout our human experience. Even the simple games played by small children are usually based on some element of a random outcome to make the games more interesting. Certainly, no one would be interested in playing a game that always proceeded in the exact same way. Any game based on spinners, dice, or a shuffled deck of cards is based on the idea that the game is made more interesting by not being able to predict the outcome. Even in games like chess and checkers, where there is no obvious random device, the enjoyment of the game is enhanced by the fact that no one can always predict what move a player may make. Even professional chess players who have met the same opponent many times never know exactly how that opponent will react in a certain situation.

Consider the popular childhood (and adulthood) game rock-paper-scissors. In this game two participants simultaneously form one of three possible hand shapes, which are named rock, paper, and scissors. The winner of the game is then determined by a set of rules that determine which hand shapes beat other hand shapes: scissors beat paper, paper beats rock, and rock beats scissors. Due to the circular (also called “non-transitive”) nature of the relationship between the hand shapes, there is not one choice that has any advantage over the others. See Figure 7.2.1. Therefore, the essential nature of gameplay is based on attempting to guess what shape the opposing player is going to choose. This is where randomness and strategy enter the game. An important aspect of strategy is that the game is usually replicated three times, with the majority winner being declared winner of the set of games.

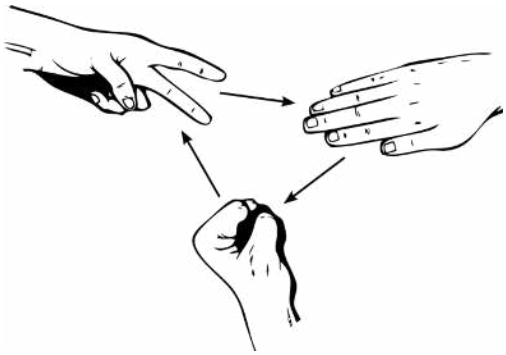


Figure 7.2.1: The non-transitive nature of how the hand shapes are related in the game rock-paper-scissors (public domain image).

When playing another individual in a game of rock-paper-scissors there are several strategies that can be used. You could, for example, always choose the same hand sign, such as rock. Overall, this is not a good strategy, particularly if known to the opponent, who could easily beat you every time by simply always choosing paper. Another option is to always choose the hand sign randomly. If the hand sign is chosen randomly so that each possible sign is equally likely to be chosen, then this strategy is decent. The advantage of this type of strategy is that your opponent cannot use any strategy to try to beat you because they cannot predict what you are going to do and, when using this strategy, you are just as likely to either win or lose. The last strategy is to use any behavioral information you may have about the tendencies of your opponent. This may be from previous games, or from what you observe in the first one or two plays of a set of games. If this strategy can be used effectively, your chance of winning can be increased—but this also depends on whether the other player is using a similar strategy. For example, an opponent may anticipate your strategy and make a counterintuitive play to nullify the information you have about them.

If your opponent is using either of the two latter strategies, you might not be able to predict what they will do on any play of the game. From your viewpoint, this is an example of how an element of randomness is introduced into the game. Because you cannot predict with certainty what your opponent will do, their play can be considered as random, and your strategy must take this into account. Further, the outcome of the game can be considered as random as any observer will not be able to predict the outcome of the game with certainty. It is this unpredictability the keeps the game interesting. In fact, theoretical investigations of this game have been shown to have important applications, for example, in biology (Sinervo and Lively 1996; Kerr et al. 2002; Reichenbach et al. 2007; Kirkup and Riley 2004; Menezes 2021; Lewin-Epstein and Hadany 2020; Schreiber and Killingback 2013). This same type of unpredictability is present in most games of strategy, including chess, checkers, and backgammon.

Just as randomness is present in some form or another in the games we play, randomness flows throughout our lives. There are few days where we can exactly predict what will happen to us. We may arrive late to work or school because of traffic or an accident on the roadways. We might accidentally step wrong on a flight of stairs and walk with a limp for the rest of the day. The person in front of you in the drive-through line at a fast-food restaurant may decide to pay for your meal as a sign of humanity. Humans thrive on and live with randomness, but we often have a difficult time describing exactly what it is. What makes something random?

This page titled 7.2: Randomness is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.3: Defining Randomness

Is a coin flip random? There are many who will state that a coin flip is random while others will state that if we knew all the relevant physical and dynamical characteristics of a coin flip, then certainly the knowledge that we have of physics would be able to tell us exactly how the coin would land. And in fact, research has shown that coin flips can be manipulated in some specialized cases to a certain extent to create situations where coins flips may not be random (Clark and Westerberg 2009; Vulovic and Prange 1986). Similar arguments occur for rolling a die and other physical random generating devices such as spinners. In fact, the nature of randomness can be seen to have deep philosophical implications (Ometto 2016; Beltrami 2020).

For our purposes we do not need to get very philosophical, and indeed while data analysts and scientists use the mathematical theory of randomness daily, a very simple definition suffices. We will start by recalling the definition of an experiment which we first encountered in Chapter 3. In that chapter we defined an experiment as any process that produces observations that provide evidence that can be used to determine if a hypothesis is true or not. The important part of this definition is that an experiment produces something that can be observed. The idea that we may use such an observation to help us support or refute a hypothesis is important for statistical studies and in studies based on data science, but for our discussion in this chapter we will focus on the idea of an observable process.

We usually think of experiments as things that go on in a lab, with beakers and chemicals or electronic equipment. Our definition of experiment is much more general. In fact, anything that we observe every day is an experiment. Experiments can be as simple as flipping a coin to observe whether it lands on heads or tails, to more complex endeavors such as applying pressure to a car windshield do determine how much pressure it can withstand before breaking. It is worth noting that we do not actually have to do anything physically to perform an experiment. We could simply wait until noon and note whether it is raining. We do not have to be manipulating a system to perform an experiment.

Now we need to address the concept of randomness. When we observe randomness, we are observing the result of an experiment, and therefore it is natural to start with the concept of an experiment and define a random experiment. In our first definition of randomness, we stated that there is a lack of definite pattern or predictability in what is observed. In the definition below we use a more precise definition of what we mean when there is a lack of predictability.

Definition: Random Experiment

A *random experiment* is any experiment whose outcome cannot be predicted with absolute certainty.

In some research articles you may also see the word *stochastic* used instead of *random*. The two words mean essentially the same thing. The key phrase used in the definition above is “absolute certainty.” From a practical point of view, an outcome can be predicted with absolute certainty if you can predict the outcome of an experiment every time without ever failing. For example, is friend is flipping a coin, you might be able to predict the outcome of the coin flip a few times just by luck. However, the definition above insists that you must be able to predict the outcome every single time the coin is flipped without ever making a mistake. There are very few situations where this could be done with a coin flip, and therefore we would conclude that a coin flip is random.

So where does this leave us on the philosophical side of things? Is a coin flip or a die roll random, or are they part of a mechanical universe that could be predicted every time without error? The key for our development is that we will never know enough about the experiments we observe for us to predict the result without error. Even if we could construct a laboratory and a coin flipping device that we could predict perfectly, the predictions would only work for that setup. We could not predict the outcome of any other coin flips in the universe. Thus, it does not matter whether a coin flip is truly random or not. What matters is that we cannot predict the outcome with absolute certainty; so, it acts randomly from our perspective whether it really is random or not.

To anyone who has played video games this is a perspective that may be well appreciated. Seemingly random events in video games, from the outcome of a slot machine app to the movements of that final boss in a roleplaying game that is attempting to end your (virtual) life, are all based on what are called pseudo-random sequence generators. A *pseudo-random sequence generator* is simply a numerical sequence that will look random to anyone looking at the sequence but is based on a completely deterministic, but complex, mathematical sequence. That is, if you knew the current position of the sequence, then you could predict with absolute certainty what the next number would be without error. From the player’s perspective, we never know where we are in the

sequence or even the exact mathematical specification that generates the sequence, and the corresponding illusion is good enough for us to be fooled into thinking that it is random. The behavior is so complex that it looks random to us.

An interesting application of randomness and pseudo-random sequence generators is in encryption technologies used on the internet. We will not go into the technical details, but secure online communications rely on the use of random number sequences. In 1995, computer scientist Phillip Hallam-Baker discovered that an implementation of a common secure communication method used on the internet used a pseudo-random sequence generator, and the starting place in the sequence was derived from the time of day, the process ID, and the parent process ID. The problem was that someone could potentially guess these values, which would predict the starting place in the sequence, and could then compromise a secure online session (Tashian 2017). To overcome this problem, computer scientists developed one of the most famous secure methods for determining the start of a pseudo-random sequence, which is based on taking a digital photograph of the current state of a wall of lava lamps (see Figure **7.3.1**). The chaotic swirls of the paraffin in the lamps are unpredictable, thereby providing a physical device for starting the sequence. This method was invented by Landon Curt Noll, Robert G. Mende, and Sanjeev Sisodiya in 1996 (Peterson 2001).



Figure **7.3.1**: A wall of lava lamps in the entrance area of the Cloudflare offices in San Francisco (the author HaeB has licensed this image under the Creative Commons Attribution-Share Alike 4.0 International license).

The concept of randomness discussed above is more than sufficient for us to have a basic understanding of how randomness is used in conjunction with the scientific method. The next step is to have some basic understanding of how statisticians describe how of certain outcomes are observed from random experiments. When statisticians talk about how rare or how frequent observations are they use a measure called probability, or chance.

This page titled 7.3: Defining Randomness is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.4: Probability

You and a friend want to see a movie but you both want to see different ones. To decide on the movie, you and your friend are going to flip a coin. This is not an uncommon method for making such a choice. This is because of the instinctive idea that using a coin flip is a fair way to decide between two choices. Otherwise, we would not consider such a manner of choosing to a be a fair way to make such a decision, or at the very least you would have a clear preference for which side of the coin will correspond to your movie and which to theirs. The idea of fairness, meaning that you and your friend are using the coin to evenly, but randomly, choose between your choices, implies that you implicitly believe that neither you nor your friend have an advantage in the outcome of the movie choice.

On the other hand, there are cases where we inherently feel that some outcomes are very likely while other outcomes are not. On a given day you might feel the need to take an umbrella with you to work or school, even though it is not raining when you leave. This may be because you either have listened to a weather report or looked at the skies and decided rain was likely enough that taking the umbrella was warranted. However, on most days you probably leave the umbrella at home, deciding it is likely you will not need it. On those days you have decided that the likelihood of raining is so low that you will take a chance by not taking your umbrella.

Consider rolling a standard six-sided die once. In most cases we would consider such a die to be fair if each side of the dice is not any more or less likely to be observed than any other side of the die. If we are comfortable with such an assertion, we can conclude that we would roll an even number (the faces with two, four or six dots) just as often as we would roll an odd number (the faces with one, three or five dots). Hence, rolling a die and observing whether an even or odd number is rolled is the same as flipping a coin and observing whether heads is flipped. However, if we consider just rolling a six, then this outcome happens less often than rolling one of the remaining faces.

Statisticians, mathematicians, and scientists use a mathematical measure called probability to talk about how likely it is for a random outcome to occur.

Definition: Probability

The *probability* that an outcome from a random experiment occurs is a number between 0 and 1, inclusive, that reflects the proportion of time the outcome would occur in a very large number of replications of the experiment under the same conditions for each replication.

You will also encounter another measure of how often random events occur, called *chance*. Chance is the same measure as probability but expressed as a percentage instead of a proportion.

Definition: Chance

The *chance* that an outcome from a random experiment occurs is a percentage between 0% and 100%, inclusive, that reflects the percentage of time the outcome would occur in a very large number of replications of the experiment under the same conditions for each replication.

Hence, if we have determined that an outcome has a probability equal to p, then the chance of the same outcome is equal to $p \times 100\%$.

This is an informal and intuitive definition of probability and chance that is based on a particular mathematical interpretation of probabilities known as *frequentism*. This is not the formal mathematical definition used by statisticians and mathematicians, but it is a convenient definition that allows for easy intuitive interpretation. Indeed, the mathematical definition of probability took a little over 250 years to develop from the initial investigations by Blaise Pascal (Figure **7.4.1** and Pierre Fermat (Figure 7.4) in the 1650s to the pioneering mathematical work of Andrey Kolmogorov (Figure **7.4.3**) in the early twentieth century.



Figure 7.4.1: French mathematician Blaise Pascal (1623–1662) who, along with Pierre Fermat, developed the first widely known mathematical approach to probability theory (public domain image).



Figure 7.4.2: French mathematician Pierre Fermat (1607–1665) who, along with Blaise Pascal, developed the first mathematical approach to probability theory (public domain image).



Figure 7.4.3: Russian mathematician Andrey Kolmogorov (1903–1987), who developed the first complete and coherent mathematical theory of probability in the early twentieth century (photograph by Konrad Jacobs. This image is licensed under the Creative Commons Attribution-Share Alike 2.0 Germany license).

The purpose here is not for you to become experts at computing probabilities for the outcomes of complex experiments. However, we will consider one simple case where probabilities are easy to compute for the sake of illustration and interpretation. Suppose an experiment has a set number of n possible outcomes. If we can assume that all the possible outcomes are equally likely to occur, then the probability that any one of the outcomes will occur is equal to $1/n$. More generally, the probability of one of m possible outcomes from a set of n equally likely outcomes is m/n with $m \leq n$. This is the method for computing probabilities that was used

by Pascal and Fermat in the mid-seventeenth century. Today this method is often called the classical method for computing probabilities.

Consider the simple example of a single coin flip. There are two possible outcomes for a coin flip: heads and tails. We will consider the simplified case where strange outcomes like the coin landing on edge are excluded. If we assume that the coin is fair and is flipped using a fair method, then we are equivalently assuming that heads and tails are equally likely to occur, and the classical method then states that the probability of flipping heads is $1/2$ and the probability of flipping tails is $1/2$. If you would rather use chance, then the chance that heads if flipped is 50% and the chance that tails if flipped is also 50%. The definition of probability then implies that if we flipped such a coin a very large number of times, then the proportion of times we would observe heads would be close to $1/2$ and the proportion of times we observed tails would also be very close to $1/2$. Probabilistic theory tells us that the proportion will probably not be exactly equal to $1/2$ but would tend to get closer as the number of times we flip the coin increases. From an intuitive viewpoint this would mean that when you see a probability near $1/2$, or a chance near 50%, the behavior is somewhat like observing heads in a coin flip.

Now consider a standard six-sided die (Figure 7.4.4), where there are six possible outcomes so that $n = 6$. If we assume that the die is fair, and that it is rolled in a fair way, then each outcome should be equally likely with a probability equal to $1/6$, about 0.1667 as a decimal, or about a 17% chance. The definition of probability then implies that if we rolled such a die a very large number of times, the proportion of times we would observe 1 would be close to 0.1667. Because you are probably very familiar with rolling a die, you can use this experience to help you think about probabilities. From an intuitive viewpoint this would mean that when you see a probability near 0.17, or a chance near 17%, the behavior is somewhat like rolling a 1 with a fair die.



Figure 7.4.4: A standard six-sided die (public domain photograph created by Alan M. Polansky).

If we were interested in other rolls, the classical method is equally effective in computing those probabilities as well. For example, if we want to compute the probability of rolling either a one or a two, then $m = 2$, $n = 6$, and the probability is $2/6 = 1/3$. The probability of rolling an even number, either a 2, 4, or 6, is $3/6 = 1/2$.

As anyone who has played various role-playing games knows, a die can have a different number of other faces than the usual six. Figure 7.4.5 shows a die with twenty sides. If such a die is fair and is rolled in a fair way, the probability of observing a 1 is $1/20 = 0.05$, or a chance of 5%. The definition of probability then implies that if we rolled such a die a very large number of times, the proportion of times we would observe 1 would be close to 0.05. From an intuitive viewpoint this would mean that when you see a probability near 0.05, or a chance near 5%, then the behavior is somewhat like rolling a 1 with a fair die that has twenty sides. As a last example we can consider a specialized die that has 100 sides (Figure 7.4.6). If such a die is fair and is rolled in a fair way, the probability of observing a 1 is $1/100 = 0.01$, or a chance of 1%. The definition of probability then implies that if we rolled such a die a very large number of times, the proportion of times we would observe 1 would be close to 0.01. From an intuitive viewpoint this would mean that when you see a probability near 0.01, or a chance near 1%, the behavior is somewhat like rolling a 1 with a fair die that has one hundred sides.



Figure 7.4.5: A twenty-sided die (public domain photograph created by Alan M. Polansky)



Figure 7.4.6: A one hundred-sided die (public domain photograph created by Alan M. Polansky)

Conceptually we can use these examples to try to intuitively understand how likely outcomes are in other applications. If a research study reports a probability equal to some number p, we can intuitively visualize a fair die with roughly $1/p$ sides and the outcome that the study is referring to would happen as often as rolling a die with $1/p$ sides and observing a 1. Note that $1/p$ will often not be a whole number, but you can usually round the number up or down to a whole number without affecting things too much. For example, a research study may report that the probability that no women would be hired from a pool of applications, assuming there was no gender discrimination in the application evaluation process, is 0.005. Setting $1/p = 1/0.005 = 200$ it follows that so that the outcome would occur about as often as rolling a fair die with 200 sides in a fair way and observing a 1. Similarly, the probability of winning the jackpot in a popular high-stakes lottery is 0.000000003422. Setting $p = 0.000000003422$, it follows that

$$\frac{1}{p} = \frac{1}{0.000000003422} = 292,226,767,$$

so that the outcome would occur about as often as rolling a fair die with 292,226,767 sides in a fair way and observing a 1.

Beyond games of chance, probability plays a major role in research studies. We will consider many of these types of studies, though we will usually not go into specific details about how the probabilities were computed. However, a few examples will help you understand the types of calculations that researchers must use to get the probabilities that they compute.

A typical, but greatly simplified, argument in an employment discrimination case can be considered using the following example. Suppose that seven equally qualified individuals apply for a job at a large local company. Three of the applicants are Hispanic, two are African American, and two are white. The two white individuals are hired from the pool of applicants. Was there discrimination in the hiring process?

One method for approaching this question is to consider it as a problem in probability theory. If the applicants are equally qualified and the hiring process is completely fair with respect to race, it can be argued that the two individuals selected for employment are selected at random with respect to race. That is, this problem can be thought of as putting all the candidates' races on slips of paper and drawing two of them at random, such as is done in a raffle, where the drawing is done in such a way that each slip of paper is equally likely to be drawn. All the possible draws, without regard to the order in which the papers are drawn, are listed in Table 7.1. In the table the three Hispanic candidates are labeled as H1, H2, and H3, the two African American candidates are labeled as A1 and A2, and the two white candidates are labeled as W1 and W2. The mathematical theory of probability can be used to show that if the selection of each slip of paper is fair on each draw, then each of these twenty-one possibilities will have equal probability of $1/21 = 0.0476$, or about 0.05 for simplicity.

Table 7.1 The 21 possible choices of two individuals from a hiring pool of seven individuals, three of whom are Hispanic (H1, H2, and H3), two of whom are African American (A1 and A2), and two of whom are white (W1 and W2).

H1, H2	H1, H3	H1, A1	H1, A2	H1,W1
H1, W2	H2, H3	H2, A1	H2, A2	H2, W1
H2, W2	H3, A1	H3, A2	H3, W1	H3, W2
	A1, A2	A1, W1	A1, W2	
	A2, W1	A2, W2	W1, W2	

In observing the possible selections listed in Table 7.1, there is only one which corresponds to both white candidates being selected. Therefore, we can argue that if the selection of the candidates is completely fair with respect to race, then the probability that both white candidates would be selected at random is about $1/21$. Now we can interpret the probability in terms of the methodology outlined above. If the selection process was fair, then the probability that two white candidates would be selected would happen about as often as rolling a 1 with a twenty one-sided fair die. We can either believe that the process was unfair, and the two white candidates were selected out of racial bias, or we can believe that a rare event occurred that would only happen about 1 in 21 times.

In the final analysis we do not know for sure which has happened, but can we try to weigh this outcome as evidence that discrimination has been proven beyond a reasonable doubt? That is, does the 5% chance that the selection could have happened create enough reasonable doubt to conclude that the corresponding selection was fair? This is something that everyone can only answer for themselves, but traditionally, in statistical calculations, this probability would be low enough to conclude that there must have been at least some bias in favor of the white candidates.

Suppose that there had been twenty-five candidates, of whom only two were white. We could construct a table like the one shown in Table 7.1. For that table there would be 300 possible selections, only one of which corresponds to both white candidates being selected. Therefore, we can argue that if the selection of the candidates is completely fair with respect to race, then the probability that both white candidates would be selected at random is about $1/300 \approx 0.0031$. Now we can interpret the probability in terms of the methodology outlined above. If the selection process was fair, then the probability that the two white candidates would be selected is as likely as rolling a 1 with a three hundred-sided fair die. Thus, we can either believe that the process was unfair, and the two white candidates were selected out of racial bias, or we can believe that a rare event occurred that would only happen in about 1 in 300 times.

This type of evidence is usually not the only type of evidence presented in discrimination litigation, but it can provide a powerful tool to aid in describing how unusual some hiring outcomes can be. Similar calculations can even be used to aid human resource departments in determining if there is bias in their advertising and recruitment techniques. For example, say it is known that 10% of a recruitment population is known to have a certain characteristic, and we have cultivated a pool of 100 applicants. We would expect about 10 individuals in the pool, which is 10% of the hiring pool, to have that characteristic. If only four people in the hiring pool have that characteristic, then we have some evidence that the cultivation process may be biased. We can get more concrete information by computing the probability that we would get four or fewer applicants with that characteristic. That probability turns out to be about 0.0237. Hence, the probability of us getting four or fewer individuals with this characteristic is about the same as rolling a 1 with a 42-sided die, based on the calculation $1/p \approx 42.171$. Once again, it is up to the individual to determine whether this probability is small enough to conclude that the recruitment process is biased.

This type of argument provides a preview of the process of formal statistical inference that will be covered later in the book. In that process a hypothesis is proposed, and then data are gathered to find evidence against the truth of the hypothesis. The evidence is evaluated by computing the probability that more evidence could be found against the hypothesis under the assumption that the hypothesis is true. If the probability is small, then there is an indication that the hypothesis is false. In the example above the hypothesis is that the recruitment process is not biased. The probability 0.0237 is the probability that even more evidence than would be found in the data could suggest the process is biased. Because this probability is relatively small, there is an indication of a considerable amount of evidence against the hypothesis that the recruitment process is not biased, and therefore we might be convinced that the process is biased.

This page titled 7.4: Probability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.5: Odds

In some research, probabilities are computed in terms of an equivalent measure called *odds*, which relates the probability that an outcome occurs to the probability that the outcome does not occur. Odds may be familiar to those who enjoy playing games of chance. For example, the probability against rolling a double six—that is, not rolling a double six—when rolling two fair six-sided dice is $35/36 = 0.9722$. In odds this same probability is said to be 35 to 1. What this says is that the probability of not rolling a double six is 35 times that of rolling a double six, which conveys the idea that a double six will not be rolled very frequently. Noting that the probability of rolling a double six is $1/36$ so that

$$\text{probability of not rolling a double 6} = \frac{35}{36} = 35 \times \frac{1}{36} = \text{probability of rolling a double 6.}$$

When there is a total of n outcomes that are equally likely to occur, and m of those outcomes correspond to an outcome we are interested in, then the odds of that outcome is said to be m to $n - m$ where $m \leq n$. In the case of rolling two six-sided dice, there are 36 possible outcomes that all have an equal chance of occurring so that $n = 36$ (Table 7.2). Of these, $m = 35$ of the outcomes correspond to not rolling a double six. Hence the odds of not rolling a double six are $m = 35$ to $n - m = 36 - 35 = 1$.

Table 7.2. The 36 possible outcomes from rolling two six-sided dice, in this case a blue die and a red die. For each pair (a, b) shown in the table, the value of a reflects the observed outcome of the blue die while the value of b reflects the observed outcome of the red die.

		Red Die					
		1	2	3	4	5	6
Blue Die	1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
	4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	5	(5,1)	(5,2)	(5,3)	(5,4)	(5,6)	(5,6)
	6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

If we just know the probability that an outcome from a random experiment is p, it follows that the probability that the outcome does not occur is $1 - p$. Therefore, if we know the probability of an outcome, we can compute the odds because we also know the probability that the outcome does not occur. Recall that if an outcome has probability p, then this is like rolling a one using a fair die with $1/p$ sides. Hence, we can take $m = 1$ and n to be approximately equal to $1/p$, and therefore $n - m$ is approximately $1/p - 1$.

Theorem 7.5.1

Suppose that an outcome from a random experiment has probability equal to p, then the *odds* of the outcome are 1 to $1/p - 1$.

Consider flipping a fair coin. We know that the probability of flipping heads is $p = 1/2$, so that $1/p - 1 = 2 - 1 = 1$. Therefore, the odds of flipping heads is 1 to 1. This means that the probability of flipping heads is equal to not flipping heads, that is, flipping tails. Because of this, the odds of flipping tails is also 1 to 1.

Earlier we considered a case where seven equally qualified individuals apply for a job. Three of the applicants are Hispanic, two are African American, and two are white, and the two white individuals are hired. In that example we calculated the probability the two white candidates would be selected is $p = 1/21$, and therefore the odds that the white candidates are selected is 1 to 20. In the second example, based on selecting two from a pool of twenty-five applicants, the probability was equal to $p = 1/300$, and therefore the odds that the two white candidates are selected is 1 to 299.

We can also convert odds back to a probability. If the odds of an outcome are m to $n-m$, then the probability that the outcome occurs is m/n. For the simpler cases where one of the numbers in the odds is 1, the formula for computing the probability is simplified. For example, if the odds of an outcome is 1 to $n-1$, then the probability of the outcome is $1/n$, whereas if the odds of an outcome is m to 1, then the probability of the outcome is $m/(m+1)$.

Theorem 7.5.2

If the odds of an outcome are m to $n-m$, the probability that the outcome occurs is m/n.

Once again consider flipping a fair coin. We know that the odds of flipping heads is 1 to 1 so that $m=1$ and $n-m=n-1=1$ then $n=2$, and the probability of flipping heads is $m/n=1/2$.

A roulette wheel is a gambling device that consists of a horizontal wheel that has numbered and colored slots around its outer edge (Figure 7.5.1). The wheel is spun, and a ball is dropped in the middle. Eventually the ball settles in one of the slots. If the wheel is fair, each slot should have an equal chance of being selected for each spin of the wheel. Roulette wheels may differ by the number of slots they contain. Our discussion here will focus on the type of roulette wheel picture in Figure 7.5.1, which has slots numbered 0–36. The slot numbered 0 is colored green, while the remaining slots are colored black and red; there are 18 red slots and 18 black slots. Players can bet on a single number, groups of numbers, or colors. The odds of winning when betting on 0 is 1 to 36. Therefore, $m=1$ and $n-m=36$, which means that $n=37$. Then it follows that the probability of winning when betting on 0 is $m/n=1/37$. Betting on a split refers to betting on any two numbers that are next to one another on the betting table (Figure 7.5.2). For example, you could bet on 1 and 2, or 7 and 10. The odds of winning when betting on a split is 2 to 35. Therefore, $m=2$ and $n-m=35$, which means that $n=37$. Then it follows that the probability of winning when betting on a split is $m/n=2/37$.



Figure 7.5.1: A roulette wheel (public domain photograph created by Alan M. Polansky).



Figure 7.5.2: A roulette betting table (public domain photograph created by Alan M. Polansky).

Odds are easiest to interpret when one of the numbers is a 1. For example, 2 to 1 odds means that the probability that the event occurs is twice that of it not occurring, or in terms or probabilities, the probability that the event occurs is $2/3$ and the probability that the event does not occur is $1/3$. Similarly, 1 to 2 odds means that the odds that the event does not occur is twice that it does occur, or in terms of probabilities, the probability that the event occurs is $1/3$ and the probability that the event does not occur is $2/3$. While odds are often reported in this way, there are some instances where one of the numbers is not a 1. The meaning of odds is still the same, but the interpretation may require some additional thought. For instance, we may have a situation where the probability of an event occurring is $2/5$ and thus, the probability it does not occur is $3/5$. This gives us the odds of 2 to 3. What this

means is that three times the probability that the event occurs is the same as two times the probability that the event does not occur. That is,

$$\begin{aligned} 3 \times \text{probability that event occurs} &= 3 \times \frac{2}{5} = \frac{6}{5} = 2 \times \frac{3}{5} \\ &= 2 \times \text{probability that event does not occur} \end{aligned} \tag{7.5.1}$$

While more complicated to interpret, there are two guidelines that can be helpful with these types of odds. The first is that the small number is associated with the smaller probability. In this example the smaller number is the first number, which is associated with the probability that the event occurs. Therefore, we know that the probability that the event occurs is smaller than the probability that is does not occur. The second guideline is that the closer two numbers are to one another, the closer the probabilities are to one another. Remember that 1 to 1 odds means that the probability that the event occurs is equal to the probability that it does not occur, which corresponds to a probability of $1/2$. Hence, for 2 to 3 odds, the probability that the event occurs is much closer to $1/2$ than if the odds were 2 to 10.

In medical studies, odds are often used to describe the probability that an individual develops a disease or condition when they have been exposed to certain circumstances. In these studies, it is often important to compare the odds of two sets of circumstances to show how risk changes under different conditions. However, comparing odds directly is difficult unless both odds start with a 1. For example, if one outcome has 1 to 3 odds and a different outcome has 1 to 10 odds, clearly the outcome with 1 to 10 odds happens less often than the outcome with 1 to 3 odds. In fact, we can use what we developed above to see that 1 to 3 odds corresponds to a probability of $1/4$, whereas 1 to 10 odds corresponds to a probability of $1/11$. If the odds are a bit different, say 3 to 5 odds compared to 4 to 7 odds, the comparison is much more difficult without converting both odds to probabilities. For this example, 3 to 5 odds corresponds to the probability of $3/8 = 0.375$, and 4 to 7 odds corresponds to the probability $4/11 \approx 0.3636$, so that the second outcome has a slightly smaller probability.

To make comparing odds easier, an odds ratio is used. Suppose that the odds of one outcome is m to $n-m$, and the odds of another outcome is k to $l-k$. For each set of odds, compute the ratio of the probability that the outcome occurs to the probability that it does not occur. For the odds above, these ratios are $m/(n-m)$ and $k/(l-k)$. The odds ratio is then the ratio of the first ratio to the second.

Definition: Odds Ratio

If the odds of one outcome is m to $n-m$ and the odds of another outcome is k to $l-k$, then the odds ratio of the first outcome compared to the second outcome is

$$\text{odds ratio} = \frac{m/(n-m)}{k/(l-k)} = \frac{m(l-k)}{k(n-m)}. \tag{7.5.2}$$

If the odds ratio is less than one, we can conclude that the odds of the first outcome are lower than the odds of the second outcome. Similarly, if the odds ratio is greater than 1, we can conclude the odds of the first ratio is greater than the odds of the second. If the odds ratio is equal to 1, the odds are the same. In the first example above, we compared 1 to 3 odds to 1 to 10 odds, where we have $m = 1$, $n = 4$, $k = 1$, and $l = 11$. The corresponding odds ratio is:

$$\text{odds ratio} = \frac{m(l-k)}{k(n-m)} = \frac{1 \times 10}{1 \times 3} = \frac{10}{3} \approx 3.33, \tag{7.5.3}$$

which indicates that the odds of the first outcome are greater than the odds of the second outcome, as we concluded above. Moreover, the odds ratio tells us that the odds of the first outcome are a little more than three times the odds of the second outcome.

Let us now consider a real-world application of odds ratios. In the early 2000s, the tobacco industry promoted the use of electronic cigarettes to replace traditional cigarettes as a safer delivery device for nicotine. The idea was that the aerosolized liquid used in such products does not involve the combustion of tobacco and contains lesser amounts of toxic substances than tobacco smoke (Sapru et al. 2020; Collins et al. 2019). The use of tobacco products is most often associated with lung cancer, and it has been of interest to study the effect of introducing electronic cigarettes to someone who still smokes traditional cigarettes. One case control study compared cigarette smoking and use of electronic cigarettes among 4,975 cases with lung cancer to 27,294 control subjects without cancer (Bittoni et al. 2024). Statistical techniques were employed to use the data from a case control study to compare the

odds of developing lung cancer for those who smoke compared to those who use electronic cigarettes and smoke. The odds ratios reported from this study have been adjusted for gender, age, and race. The researchers reported that the odds ratio for developing lung cancer for chronic smokers who also used vaping products compared to those who only smoked is 13.9. This indicates that the odds of developing lung cancer when combining vaping with chronic smoking is almost 14 times the odds of developing lung cancer when smoking alone. The researchers concluded that the results suggest the addition of vaping to smoking accelerates the risk of developing lung cancer.

As shown in the example above, odds ratios are often used in medical studies to compare the risks associated with two conditions, usually a treatment and some type of control. The data from these studies are often reported in a 2×2 table that includes the number of cases where an associated outcome such as a disease is present, and the number of cases where the outcome is not observed for the treatment group and the control group. An example of this type of table is shown in Table 7.3, where the number of cases is represented by a, b, c, and d)\.*InTable*7.3, \(a represents the number of cases in the treatment group that developed the disease, b represents the number of cases in the treatment group that did not develop the disease, c represents the number of cases in the control group that developed the disease, and d represents the number of cases in the control group that did not develop the disease. The odds ratio for this data can be computed as:

$$\text{odds ratio} = \frac{a \times d}{b \times c}. \tag{7.5.4}$$

Table 7.3 Data for comparing the development of a disease for a treatment and control group.

	Disease Developed	
	Yes	No
Treatment	a	b
Control	c	d

For example, consider the data given in Table 7.4, which considers a study of the use of taking vitamin C and its effect on the common cold (Anderson et al. 1972). In this study, 818 subjects were divided randomly into a group of 407 subjects who took vitamin C over for a period, and another group of 411 subjects who took a placebo over the same time period. Based on the data given in Table 7.3, the odds ratio is 0.6252. This indicates that the odds of developing a cold-related illness while taking vitamin C are less than the odds of developing a cold-related illness when not taking vitamin C, indicating evidence that vitamin C may reduce the chance of developing a cold-related illness.

Table 7.4 Data comparing the occurrence of illness for subjects taking Vitamin C against those who took a placebo (Anderson et al. 1972).

	Illness	
	Yes	No
Vitamin C	302	105
Placebo	335	76

This page titled 7.5: Odds is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.6: Critical Analysis

While probability is used extensively in many studies, we will need further development of the statistical framework used with the scientific method before most of those studies can be critically analyzed. We will highlight a couple of applications of probability where we will discuss how to interpret some probabilities that are reported in several situations. Some of these applications deal with such things as having the same birthday as someone else, while others are about more serious subjects.

When considering studies that use probability, keep in mind that the computation of specific probabilities may be very complicated and may use techniques that require specialized training to understand. Even the assumptions used to compute probabilities in some cases are very technical in nature. In these cases, an individual can still assess the conclusions that are based on the probabilities. For example, when a study concludes that something is rare, and includes a probability, you can assess for yourself whether the probability really means that the event is rare. Such a determination is often related to the context of the outcome. For example, a probability equal to 0.10 may seem rare if you are considering the probability that your favorite team might win a game, while the same probability may not seem rare if it is connected to contracting serious disease. These are conclusions that everyone makes for themselves when assessing a probability.

Many people do not have a great deal of experience thinking in terms of odds, particularly when they are not in the form of 1 in x, where x is some whole number. It is often convenient to convert odds into probabilities to make the results easier to understand, though it can be somewhat tedious for interpreting the results of large research projects.

This page titled 7.6: Critical Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.6.1: The Birthday Problem

Suppose that it is your birthday, and your friends take you out to a restaurant to celebrate. During the evening you are brought a small cake, and perhaps your server will sing the birthday song to you. You are enjoying the evening, and then from another part of the restaurant, you hear your server singing the same song to another person. You have now realized that you share the same birth date with someone else in that restaurant!

Sharing a birthday with another person may seem surprising at first. After all, birthdays feel like very special and personal days that you would not want to share with someone else. But what is the probability that, if you gather a group of individuals together, there is a least one pair of people who have the same birthday? As we shall see below, the probability that this can occur might be higher than you think.

Before we begin thinking about computing this probability, we need to make some assumptions. To begin, we will exclude leap years for our discussion. The problem when leap years are included is that birthdays on February 29th are noticeably less common than other birthdays, and this will violate our second assumption. Next, we will assume that if we randomly select a person from the general population, then each of the possible 365 birthdays under consideration are equally likely. This assumption is not true as birth dates follow seasonal patterns (Nunnikhoven 1992). However, these assumptions will make it easier to compute the associated probability and demonstrate the method without additional complications. We will discuss the results of a more comprehensive and realistic approach at the end of this section.

To begin understanding how to compute this probability, first consider selecting two people at random. The first person selected will have a birthday that we can represent by a number b_1, which is some integer between 1 and 365. Hence, 1 will represent the outcome that the person's birthday is on January 1st, 2 represents the outcome that the person's birthday is on January 2nd. This continues until we get to 365, which represents the outcome that the person's birthday is on December 31st. When the second person is selected, their birthday will be represented by a number b_2, with the same representation as shown above. Using the classical method for computing probabilities, the probability that b_1 and b_2 are the same is $1/365$, since there are 365 equally likely possible birthdays for the second person and only one of these matches the birthday for the first person. Hence, the probability that two people having a matching birthday is $1/365$.

Now consider three people, whose birthdays will be represented by the numbers b_1, b_2, and b_3. In this case there are many ways in which at least a two of the individuals share a birthday. Either $b_1 = b_2$, $b_2 = b_3$, $b_1 = b_3$, or $b_1 = b_2 = b_3$. To compute the probability that one of these outcomes occurs, we would need to list out all possible birthday combinations for the three people, which corresponds to $365 \times 365 \times 365 = 48,627,125$ possibilities, so that $n = 48,627,125$. Next, we would need to count how many of these possibilities correspond to one of the conditions shown above. Because of the four possible ways that this outcome can occur, this calculation can be rather complicated. Alternatively, we could compute the probability that all three birthdays are different, which is the remaining possibilities. We can this subtract this from n to get the number m that we need to compute the probability.

Suppose that the first sample person's birthday is b_1. There are 365 possibilities for their birthday. When we sample the second person, there are 364 possibilities for their birthday to be different than the first person, that is, $b_1 \neq b_2$, for each of the 365 possibilities for b_1. Therefore, there are a total of $365 \times 364 = 132,860$ outcomes where $b_1 \neq b_2$. When we sample the third person, there are 363 possibilities for their birthday to be different than the first two people, that is, $b_3 \neq b_1$ and $b_3 \neq b_2$ for each of the 132,860 possibilities where $b_1 \neq b_2$. Therefore, there are a total of $132,860 \times 363 = 48,228,180$ outcomes where all three birthdays are different, or $b_1 \neq b_2 \neq b_3$. In all the remaining outcomes, at least two people share a birthday. Therefore, there are $m = 48,627,125 - 48,228,180 = 398,945$ outcomes where at least two people share a birthday. Hence, the probability that at least two of the three people share a birthday is

$$p = \frac{398,945}{48,627,125} = \frac{1093}{132132} \approx 0.008204,$$

or about an 0.8% chance. Hence there is a little less that a 1% chance that at least two people will share a birthday for three randomly selected people. This is unlikely but is slightly larger than that for two people, which has a chance equal to $1/365 \times 100$.

This argument can be used for any number of people, though the calculations become difficult to compute due to the large numbers involved. For four randomly selected people, the probability that at least two of them share a birthday is

$$p = 1 - \frac{365 \times 364 \times 363 \times 362}{365 \times 365 \times 365 \times 365} = \frac{290,299,465}{17,748,900,625} = \frac{795,341}{48,627,125} \approx 0.016356,$$

which corresponds to about a 1.6% chance. For five randomly selected people, the probability that at least two of them share a birthday is

$$p = 1 - \frac{365 \times 364 \times 363 \times 362 \times 361}{365 \times 365 \times 365 \times 365 \times 365} = \frac{175,793,709,365}{6,478,348,728,125} = \frac{481,626,601}{17,748,900,625} \approx 0.027136,$$

which corresponds to about a 2.7% chance.

The probabilities are increasing, and this is reasonable because the more people observed, the more chances there are for birthdays to overlap. In fact, if there are 366 people in the group, there must be at least one pair who share a birthday as there are only 365 birthdays to choose from. So, somewhere between 5 people and 366 people, the probability increases considerably. What surprises most people is that it does not take too many people before the probability gets quite large. Table 7.5 shows these probabilities for between 10 and 70 people. For example, from the table we can observe that if there are as few as 23 people in a group, the chance that there are at least two people who share a birthday is above 50%. When you get 50 people in a room together, there is a 97% chance that at least two people in the room share a birthday. That means if you are sitting in a class with at least 50 other students, it is almost certain that there are at least two people in the room who share a birthday!

Table 7.5 The probability that at least two people in a group share a birthday assuming each person is equally likely to have one of

365 birthdays in the year excluding leap days.

Number of People	Probability at Least One Pair Shares a Birthday
10	0.1169
11	0.1141
12	0.1670
13	0.1944
14	0.2231
15	0.2529
16	0.2836
17	0.3150
18	0.3469
19	0.3791
20	0.4114
21	0.4437
22	0.4757
23	0.5073
24	0.5383
25	0.5687
26	0.5982
27	0.6269
28	0.6545
29	0.6810
30	0.7063
31	0.7305
32	0.7355
33	0.7750
34	0.7953
35	0.8144
36	0.8322
37	0.8487
38	0.8641
39	0.8782
40	0.8912
50	0.9704

60	0.9941
70	0.9992

The rapid increase of this probability is sometimes called a paradox because it does not seem possible that the probability would be so large for so few people. The main thing to remember is that this is not the probability that someone has the same birthdate as yours; this is the probability that at least two people in the room have a matching birthday.

What about the assumptions that we made at the beginning of this section? As we stated earlier, the distribution of birthdays is not uniform throughout the year, and hence there are some months of the year where birthdays are more common than others. In fact, recent birthdate data (1994–2014) compiled by the U.S. Centers for Disease Control and Prevention's National Center for Health Statistics (NCHS) and the Social Security Administration show that September 9 is the most common birth date in the United States. September is a very common month for birthdays as nine of the top ten birth dates occur in September, though the most common month for a birthday is August (Cahn 2024). How does the nonuniformity of birth dates throughout the year affect the calculations?

It turns out that the probability of at least one match increases when the distribution of birth dates is not uniform (Bloom 1973; Munford 1977; Rust 1976). This makes sense because if there are some birthdays that are more common than others, we would expect that there would be a better chance of at least two individuals having those birthdays. Some researchers have used the actual distribution of birthdays in the United States to determine how much these probabilities change. What they found is that the probabilities are relatively insensitive to the moderate variation in birthrates observed in the United States (Nunnikhoven 1992). For example, when there are 10 people in the group, the probability of at least one match is 0.116948 when birthdays are assumed to be uniformly distributed throughout the year, whereas this probability is 0.117100 when the actual distribution of birthdays is used.

This page titled 7.6.1: The Birthday Problem is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.6.2: Risk of Death from a Shark Attack

The Florida Museum devotes an entire section of their website (www.floridamuseum.ufl.edu) to the lifetime risk of fatality due to shark attack with comparisons to other types of accidental death. The purpose of the research is to emphasize the point that even though sharks have a deadly and dangerous reputation, the actual chance of becoming a victim of a shark fatality is rather low compared with many other activities that we may participate in regularly.

The website states that the odds of suffering a shark-related fatality within your lifetime is 1 in 4,332,817. This compares to 1 in 5 odds for dying from heart disease, 1 in 7 odds for dying from cancer, 1 in 84 odds for having a fatal car accident, and 1 in 79,746 for dying from lightning. The odds of a fatal shark stack are clearly much smaller than any of these other common causes of fatalities. We will compare these odds more directly by first converting them to probabilities and then by calculating the odds ratios of a shark fatality compared to these other types of fatalities.

We will first convert these odds to probabilities. The odds of dying from heart disease is 1 in 5 so that the probability of dying from heart disease during your lifetime is $1/(5+1) = 1/6 \approx 0.1667$. Similarly, the probabilities for dying from cancer is $1/8 = 0.1250$, having a fatal car accident is $1/85 \approx 0.0118$, and dying from a lightning strike is $1/79747 \approx 0.0000125$. Converting the odds of dying from a shark attack yields a probability of $1/4332818 \approx 0.000000231$. Clearly we can observe that the probability suffering a fatal shark attack is much less than each of these outcomes.

Now let us compute the odds ratio comparing a shark attack to a couple of the other outcomes. We first compare the odds of dying from heart disease against the odds of dying from a shark attack. The odds of dying from heart disease are reported as 1 in 5 so that $m = 1$ and $n - m = 5$. The odds of dying from a shark attack are 1 in 4,3323,817 so that $k = 1$ and $l - k = 4332817$. Therefore, the odds ratio is

$$\text{odds ratio} = \frac{m(l-k)}{k(n-m)} = \frac{1 \times 4332817}{1 \times 5} = \frac{4332817}{5} \approx 866563.4,$$

which shows of course that the odds of dying from heart disease are much greater than the odds of dying from a shark attack. The closest risk we can find in the data is when we compare the odds of dying from a lightning strike compared against the odds of dying from a shark attack. The odds of dying from a lightning strike are reported as 1 in 79746 so that $m = 1$ and $n - m = 79746$. Therefore, the odds ratio is

$$\text{odds ratio} = \frac{m(l-k)}{k(n-m)} = \frac{1 \times 4332817}{1 \times 79746} = \frac{4332817}{79746} \approx 54.33,$$

which shows that the odds of dying from a lightning strike are about 54 times the odds of dying from a shark attack.

The purpose of the information on the website is to alleviate fears about shark attacks on the Florida coast. As with any study of this type, one needs to interpret the odds and the probabilities in context. To compute the odds, the number of deaths of each type was compared to the population of the United States in 2021. The risk was averaged over about 70 years, the current life expectancy of the population. First and foremost, these odds then apply to individuals who live in the United States. Furthermore, this is a risk that aggregated over the entire population and may not apply singly to each individual person. For example, a person who practices a healthy lifestyle may have much lower odds of dying from a heart attack. Individuals who do not use tobacco products are likely to have much lower odds for cancer. The odds of dying in a fatal car accident may depend on where one lives and drives, and if they drive at all. Finally, one can lower the odds of being the victim of a fatal shark attack by simply not going in the ocean. Hence, while these odds make for interesting comparisons, it is much more accurate to simply report the number of fatalities of each type, noting that relatively few people die from shark attacks each year.

This page titled 7.6.2: Risk of Death from a Shark Attack is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.6.3: Bias in the Application of the Death Penalty

Capital punishment in the United States has long been criticized as historical data shows that it has been disproportionately applied to African American defendants. This led the Supreme Court to declare that capital punishment, as it was applied at the time, to be unconstitutional in 1972. After this ruling, many states changed their capital punishment statutes, and many of these statutes are still applied today. A short time after the state of Florida reenacted its program of capital punishment, a study of 600 homicide indictments in 20 counties in Florida from 1976 and 1977 were examined to determine if capital punishment was still being applied unevenly according to race (Radelet 1981).

Table 7.6 shows the number of individuals who were indicted of first-degree murder and sentenced to death according to their race and the race of the victim. For illustrative purposes, we have combined the data from primary homicides, which involve family and friends, and non-primary homicides, which typically involve strangers.

Table 7.6 Data showing the number of individuals indicted for first degree murder who were sentenced to capital punishment based on the race of the defendant and the victim. (Radelet 1981).

	Capital Punishment	
	Yes	No
White Victim		
African American Defendant	11	48
White Defendant	22	175
African American Victim		
African American Defendant	6	101
White Defendant	0	8

From the table we see that the odds ratio for an African American defendant being sentenced to capital punishment compared to a white defendant, when the victim is white, is

$$\text{odds ratio} = \frac{11 \times 175}{22 \times 48} = \frac{1925}{1056} \approx 1.82.$$

This indicates the odds that a black defendant is sentenced to capital punishment is 1.8 times the odds of a white defendant being sentenced to capital punishment. Similarly, the odds ratio for an African American defendant being sentenced to capital punishment compared to a white defendant, when the victim is African American, is

$$\text{odds ratio} = \frac{6 \times 8}{0 \times 101}. \tag{7.6.3.1}$$

In this case the odds ratio cannot be computed because the denominator of this fraction is 0. It is very difficult to draw any conclusions in this case. Looking closely at the data we see that a little less than 6% of the African American defendants were sentenced to capital punishment where there are 107 black defendants in all. Note that there are only 8 white defendants. If the sentencing rate were the same, then we would probably not observe any white defendants being sentenced to capital punishment because 6% of 8 is approximately 0.5. Therefore, even if the rates of capital punishment were the same, we might observe the data that is observed here. The relative lack of data means that we cannot make any conclusions in for the case of African American victims.

This page titled 7.6.3: Bias in the Application of the Death Penalty is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.6.4: The Vietnam War Draft Lottery

During the Vietnam war young men were subject to conscription by the United States military. To reduce the bias in the system of selection, which had been criticized for selecting the poor and undereducated, the government decided to use a random system of selection, which became known as the draft lottery (Erikson and Stoker 2011). Between December 1969 and February 1972, the United States Selective Service held four Vietnam draft lotteries. Each draft lottery randomly assigned men in the same birth cohorts order-of-induction numbers through a mechanical drawing of 366 birth dates. Following the lotteries, men in a particular birth cohort were called to report for possible induction into military service.

To select the birth dates at random, the 366 possible birth dates were printed on slips of paper, which were placed in plastic containers. The containers were opaque so that the dates could not be seen. The containers were shuffled and placed in a glass jar from which the containers were chosen one at a time and opened. The first date drawn was September 14, followed by April 24, December 30, February 14, October 18, and June 8. The first 195 birthdays drawn were later called to serve in the order they were drawn (Abney 2019).

The draft lottery has been heavily criticized as statisticians raised concerns that each of the dates did not have an equal likelihood of being selected. It appears that the mixing process used was not sufficient to ensure that the containers were equally likely to be drawn. An analysis of the drawing sequence shows, for example, that the first 4 months of the year appeared less frequently than the other 8 months in the first third of the dates drawn. Similarly, the last 4 months of the year appeared less frequently in the last third (Fienberg 1971). Additionally, the same research showed that if the containers were mixed so that each container was equally likely to be chosen, the probability of getting such a result is approximately 0.02. Hence this would be equivalent to rolling a 50-sided die and rolling a 1. Further, a more sophisticated statistical analysis appears to support the contention that the process used to select individuals for the draft was not fair in that some dates were more likely to be picked than others.

This page titled 7.6.4: The Vietnam War Draft Lottery is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.6.5: Ovid and Gambling

In this section we will consider a very old application of the idea of probability to the problem of rolling three fair, six-sided dice. Publius Ovidius Naso (43 BCE–18 AD), known in English as Ovid, was a Roman poet who lived during the reign of Augustus. Ovid was popular in Europe in the 12th and 13th centuries to the point that people composed works during that time and attributed them to Ovid (called pseudo-Ovidian works). *De Vetula* is a pseudo-Ovidian poem consisting of three books written in France in the mid-thirteenth century. The real author of the poem is unknown. It is written in the form of an autobiography of Ovid. The poem is a morality story in which the reader is led to transform his life from one centered on pleasure at the beginning to the one that is accepting the Christian faith by the end.

The first book describes Ovid's youth, including pursuits such as hunting and fishing. The next section deals with reasons to avoid dice playing, and this is where probability and chance are addressed. A section of the poem deals with the calculation of the chances of various throws of dice. The section of the poem begins by describing the outcomes of one die and three dice. Of particular interest is the case when three dice are thrown and the numbers represented on the upper faces of the dice are added together. The author notes that certain sums occur more often than others. Specifically, with three dice, there are sixteen possible sums that can be observed. For example, while the outcomes 3, 4, 17, and 18 occur less frequently as there is only one configuration that will yield these sums, the outcomes 9, 10, 11, and 12 occur more frequently as there are six configurations that yield these sums.

Figure **7.6.5.1** shows a page from an early printing of the book. The top row of the table shown in the photograph shows (6,6,6), that is, each die rolled is a 6, which is the only possible way of observing a sum of 18. The next row shows (6,6,5), that is, one die is a 5, and the other two dice are each 6, is the only possible way of observing a sum of 17. When we look at the third row of the table, we get our first clue as to why some sums occur more often than others. In the case of observing a sum of 16 there are two possible configurations of the dice that give this sum: (6,6,4) and (6,5,5). The reasoning is now that because a sum of 16 can occur with two possible configurations of the dice, and there is only one possible configuration that will produce each of the sums 17 and 18, the sum 16 will be observed more often than either of the sums 17 or 18. Working down the table we can observe that the sums 9, 10, 11, and 12 should be observed the most often in that there are six configurations that will produce each of these sums. For example, the six configurations that will produce a sum of 10 are (6,3,1), (6,2,2), (5,4,1), (5,3,2), (4,4,2), and (4,3,3). However, as we shall see, this does not tell us the complete story.

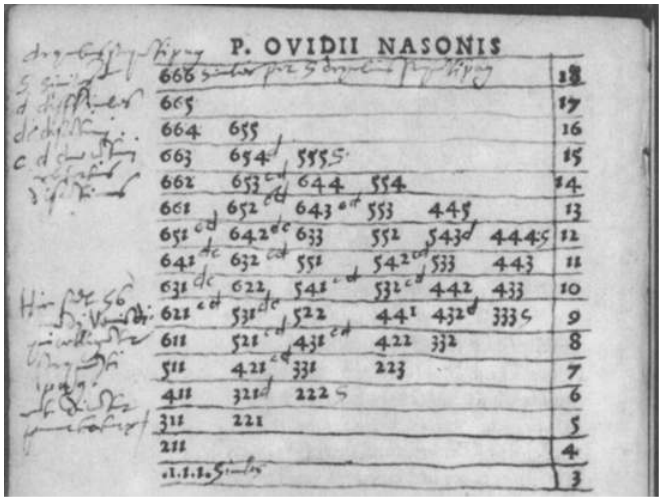


Figure **7.6.5.1**: A page from an early printing of *De Vetula* that shows why some sums of three dice occur more often than other (public domain image).

If you roll three dice many times and keep track of the sums, you will observe after a while that rolling a sum of 17 occurs more often than rolling a sum of 18, even though both sums only have one configuration that will produce the sum. The reason for this, as the author of *De Vetula* shows, is that for the sum of 18 all three dice must be equal to 6, while for the sum of 17 one of the dice is different than the other two. This makes a difference because we also need to consider which of the three dice is equal to 5. To see why this is true, suppose one of the dice is red, another is blue, and the third is green. Then there are three choices that will produce this configuration: the red die is 5 while the blue and green dice are 6, the blue die is 5 while the red and green dice are 6,

and the green die is 5 while the blue and red dice are 6. In fact, as this argument shows, we should observe a sum of 17 about three times as often as we observe a sum of 18.

Figure **7.6.5.2** presents a photograph of a table from an early printing of *De Vetula* showing these differences. The first group of configurations will only occur once because all the numbers are the same. The second group of numbers can each occur in three different ways because two numbers are the same and one is different. The third, fourth, and fifth group of numbers have each dice equal to a different number, and the numbers are in order. Each of these configurations can occur in 6 different ways. For example, the configuration (1,2,3) can occur as (1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), (3,2,1).

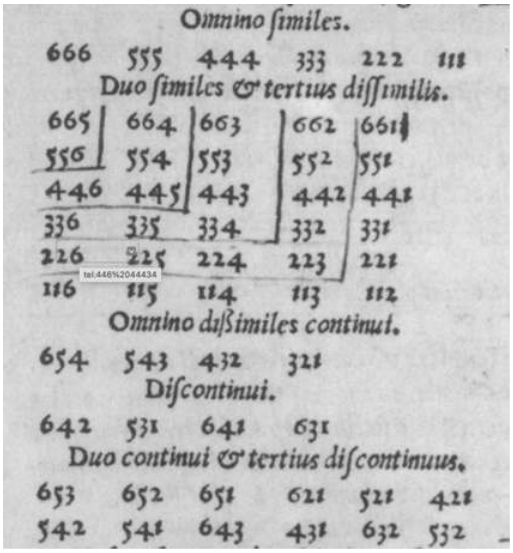


Figure **7.6.5.2**: A page from an early printing of *De Vetula* that shows the number of ways each configuration of three dice rolls can occur (public domain image).

In all there are $6 \times 6 \times 6 = 216$ possible ways that three dice can be rolled that accounts for all the configurations and orderings. To compute the probability of a particular sum we need to count how many configurations and orderings correspond to that sum and divide by 26. Starting with the sum of 18, we find an easy case because there is only one configuration and one ordering, which is (6,6,6). Therefore, it follows that the probability of observing a sum of 18 is $1/216$. The next largest possible sum is 17, which has one configuration but three possible orderings (6,6,5), (6,5,6), and (5,6,6). Therefore, it follows that the probability of observing a sum of 17 is $3/216$. This process of counting continues for the remainder of the possible sums from 16 down to 3. Near the middle of the range, the counting procedure can become rather complicated. When all the counting is done, we end up with a table such as the one shown in the photograph in Figure **7.6.5.3**, which is also from an early printing of *De Vetula*.

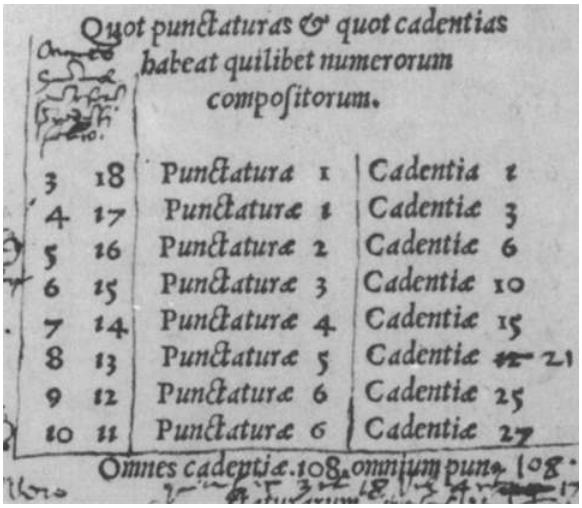


Figure **7.6.5.3**: The table of probabilities for the sums of three dice from an early printing of *De Vetula* (public domain image).

In Figure **7.6.5.3** we can focus on the last column of the pictured table where the number of configurations and orderings, taken together, is reported. From this table we can observe that the sums of 18 and 3 will both occur with a probability of $\mathbf{1/216}$. Similarly, the sums of 17 and 4 will both occur with a probability of $\mathbf{3/216}$ and the sums of 16 and 5 will both occur with a probability of $\mathbf{6/216}$. What one quickly observes from the table is that the probabilities of observing a sum near the middle of the range are much higher, for example a probability of $\mathbf{27/216}$ for getting a sum of 10 or 11, when compared to the sums near the end of the range. The table is quite helpful in that not only does it emphasize that some sums are more likely than others, but the calculations specifically enumerate how much more likely some sums are than others. For example, the chance of observing a sum of 5 is six times that of observing a sum of 3. While these calculations are quite routine using modern methods for counting configurations and orderings, it is notable that *De Vetula* is the first known exposition of this type of problem in probability.

This page titled 7.6.5: Ovid and Gambling is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.6.6: Sports Activities and Lower Back Pain in Adolescence

Lower back pain is a serious condition that can become a chronic lifetime issue. Research has shown that lower back pain is as prevalent in childhood and adolescence as it is in adulthood (Balagué et al. 1988; Balagué et al. 1995; Burton et al. 1996; Hestbaek et al. 2006; Jones et al. 2004; Kazemi 2025; Leboeuf-Yde and Kyvik 1998; Li et al. 2024; Salminen et al. 1992; Taimela et al. 1997). Throughout the years many potential sources of this problem have been identified, including lifestyle choices, playing video games, the weight of school bags, and physical health factors (Sato et al. 2011). One study considered survey data collected from 26,766 students in Niigata City, Japan, to determine whether participation in sports activities may increase the likelihood that lower back pain is developed in adolescents (Sato et al. 2011).

The study included all schoolchildren, including all elementary school students from the fourth to sixth grade and all junior high students from the first to third year, in Niigata City, located on the west coast of Japan. The children ranged in age from 9 to 15 years old. An anonymous questionnaire was distributed to each school. The data were collected from October to November in 2005. The survey collected basic demographic and health-related information along with any experience of lower back pain in the present and the past. Additionally, the details of sports activities other than school physical education classes, musical activities, and social activities were reported in the questionnaire. The amount of time spent participating in sports activities was also reported.

The responses were divided between those who did not participate in sports activities except for physical education in school and those who participated in sports activities. The prevalence of lower back pain was divided into three levels: level 1, no limitation in any activity; level 2, necessary to refrain from participating in sports and physical activities; and level 3, severe lower back pain, where a student needed to refrain from participating in sports and physical activities and was absent from school. The amount of time students spent participating in sports activities was divided into three time ranges: less than 6 hours per week, between 6 and 12 hours per week, and more than 12 hours per week.

The researchers in the study compared the odds of developing severe lower back pain for students who did not participate in sports against those who did and provided a detailed analysis for 21 different sporting activities. Statistical methods were used to remove the potential confounding effects of age, gender, and body mass index. When comparing all the students who participated in the sports against those who did not, the odds ratio for severe lower back pain was 1.57, indicating that the sports group had a slightly higher prevalence of severe lower back pain. The results varied by the type of sport that the student reported playing. For example, the odds ratio for students who played table tennis was 1.05, indicating that the odds of severe lower back pain were almost the same as students who did not participate in sports. On the other hand, the odds ratio for rugby was 2.58, indicating that the odds of having lower back pain for rugby players was more than twice that of students who did not participate in sports. One sport, sumo wrestling, had an odds ratio of 0.85, which indicates that students who participated in this sport had lower odds of severe lower back pain than students who did not participate in sports. Sumo wrestling was the only sport that had such a result.

This page titled 7.6.6: Sports Activities and Lower Back Pain in Adolescence is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.7: Summary

Randomness is an important aspect of everyday life that not only is part of the games we play but is a crucial part of the modern application of the scientific method. The likelihood of the occurrence of random events can be described using the mathematical measure called probability or, equivalently, chance. An intuitive understanding of probability is crucial to being able to properly interpret the statistical analyses that are commonly used in the application of empirical studies to the scientific method. In this chapter we introduced the concepts of probability and chance and discussed how they can be interpreted in terms of rolling a die with a specified number of sides. We also introduced the concept of odds, which is a measure of how often a random event occurs compared to how often it does not occur. In medical studies, the odds of two groups are often compared using an odds ratio.

This page titled 7.7: Summary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.8: About the Quotation

Sidney Poitier (1927–2022) was an actor, director, activist, and diplomat. He was the first black actor to win the Academy Award for Best Actor. He also won two Golden Globe Awards, a BAFTA Award, and a Grammy Award. The quote is from an interview with the *Washington Post* and appeared in a Reuters article published shortly after his death.

This page titled 7.8: About the Quotation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.9: Practice and Exploration

1. Suppose that six equally qualified employees are eligible for a promotion. Four of these employees are Asian American. Two employees are to be selected at random and neither are Asian American. Make a table like Table 7.1 that lists all the possible selections of two employees from the pool of six employees. Use the labels A_1, A_2, A_3, and A_4 to denote the Asian American employees. Use the labels N_1 and N_2 for the two employees who are not Asian American. You should get fifteen such choices. Remember that the order in which the employees is selected does not matter.
2. Determine how many of these possible selections correspond to choosing the two individuals who are not Asian American. Use this number to compute the probability of choosing the two individuals who are not Asian American.
3. In an employment discrimination case, a lawyer is presenting evidence that their client, in this case the company, has a racially unbiased method for evaluating and hiring employees. As part of this evidence the lawyer has had a statistician compute the probability that no Hispanic applicants would be hired from a specified applicant pool, under the assumption that there is no bias. The lawyer states that this probability is equal to 0.001, saying that this is not a rare occurrence. Using your intuitive ideas about probability, do you agree with the lawyer that this would not be a rare occurrence?
4. A ballot box contains one hundred ballots, seventy-five for Candidate *A* and twenty-five for Candidate *B*. Suppose that a ballot is selected at random in a way that each ballot of an equal chance of being selected. What is the probability that the selected ballot is for Candidate *A*? What are the odds that the selected ballot is for Candidate *A*? What is the odds ratio of selecting a ballot for Candidate *A* compared to selecting a Ballot for Candidate *B*?
5. The odds that a senior at a certain suburban high school will go to college is 2 to 1 while the odds that a senior at a certain urban high school will go to college is 1 to 3. Compute the probabilities corresponding to each of these odds. According to these odds, is it more likely that a high school senior will go to college from the urban or the suburban high school? What is the odds ratio comparing the suburban high school to the urban high school?
6. A professor at a large university is interested in whether students who use the free department tutoring center are less likely to fail their course. The professor includes an extra credit question on their final exam that asks whether the student included the free tutoring center as part of their usual study habits for the class. The professor then aggregated these responses with the final grades in the course to produce the following table:

Used Tutoring Center?	Passed Course	Failed Course
Yes	24	12
No	33	9

   Based on the observed data, compute the odds ratio for passing the course when using the tutoring center compared to not using the tutoring center. Interpret the result that you get. Are you surprised by the result? What confounding factor might explain why you got this result?

7. Suppose that a standard deck of cards is shuffled and the top card is selected. Assuming that each of the cards is equally likely to be on the top of the deck, what is the probability of drawing an ace? Convert this probability to chance and to approximate odds. Next compute the probability of drawing a diamond. Convert this probability to chance and to approximate odds. Compute and interpret the odds ratio for drawing an ace compared to drawing a diamond.
8. The *complement rule* is a result from probability that states that the probability that an outcome does not occur is one minus the probability that the outcome does occur. This result was used several times in this chapter without naming the result that we were using. The simplest case of applying the complement rule is for a coin flip. In terms of the experiment of flipping a coin, explain in your own words why the probability of flipping heads is equal to one minus the probability of not flipping heads, that is, flipping tails. How would you translate the complement rule from probability to chance?
9. You and a friend are preparing to play a role-playing game. When you unpack your dice, you notice that neither of you brought an eight-sided die. Your friend states that this is no problem because both of you brought a four-sided die, and if you roll both four-sided dice at one time and add the result that this will be equivalent to rolling a single eight-sided die. Do you think that this is true?

10. In the study of sports activities and lower back pain in adolescents, it was reported that the odds ratio for developing lower back pain for students who participated in soccer compared to the no sports group was 1.77, while the odds ratio for developing lower back pain for students who participated in archery compared to the no sports group was 1.10. Based on these two odds ratios, can we conclude that students who participate in soccer are more likely to develop lower back pain than students who participate in archery? Why or why not?
11. Earlier we reported that the odds of suffering a shark-related fatality within your lifetime is 1 in 4,332,817. The same report from the Florida Museum states that the odds of dying in a bicycle accident during your lifetime is 1 in 4,919. Based on these odds, compute the probability of dying in a bicycle accident during your lifetime. What is the odds ratio of dying in a bicycle accident compared to a fatal shark attack?
12. The epic poem *De Vetula* considered computing the probabilities associated with each of the possible sums observed when rolling three six-sided dice. Suppose that instead of a die, we have a coin where one side is labeled 0 and the other side is labeled 1. Doing the same analysis used in *De Vetula*, find the probabilities associated with each possible sum of three flips of this coin.

This page titled 7.9: Practice and Exploration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.10: Collaborative Activity

For this collaborative activity you will need a standard deck of cards, a ten-sided die, a twenty-sided die, and someone in your group who is good at shuffling cards. We will simulate an experiment where a group of patients with a minor medical condition are recruited for a medical study. Each patient will either receive a standard treatment or a new treatment where which treatment each individual receives will be randomly determined by drawing cards from the shuffled deck of cards. According to the treatment determined by the card draw, a dice roll will then determine whether the treatment was successful. The data gathered from the experiment will be used to calculate probabilities, odds, and odds ratios.

To begin, designate an individual in your group as the person who will shuffle the deck of cards and will successively draw cards from the top of the deck. Another individual will oversee rolling the dice, and a third individual will write down the simulated data on the provided data sheet shown in Table 7.7.

Table 7.7 Data collection table for the collaborative activity.

Replication	Card	Dice Roll	Treatment	Outcome
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				

The activity begins by shuffling the deck of cards. Shuffle them at least five times so that the order of the cards is mixed up well. Drawing from the deck will then be approximately equivalent to random draws from the deck of cards.

Next, repeat the following process for twenty-five replications:

1. Selected a card from the top of the deck and write the in the *Card* column of the data table on the line corresponding to the current replication. You can us a two-character abbreviation to write down the results using 2–10, J, Q, K, and A to represent the value of the card and C, D, H, and S to represent the card suit. For example, 2H will represent the two of hearts.
2. If the card that has been selected is a heart or a diamond, roll the twenty-sided die, and record the result in the *Dice Roll* column of the data table on the line corresponding to the current replication.

3. If the card that has been selected is a club or a spade, then roll the ten-sided die, and write the result in the *Dice Roll* column of the data table on the line corresponding to the current replication.

Table 7.8 gives an example of what the first two columns of the data table for ten hypothetical replications of this process. From the table we can observe that the first card drawn from the top of the deck was the five of spades, which has been represented by 5S on the table in the *Card* column for the first replication. Because this card is from the spade suit, the ten-sided die was rolled and the resulting roll was 3, which has been recorded in the *Dice Roll* column for the first replication. The second card drawn from the top of the deck was the two of hearts, which has been represented by 2H on the table in the *Card* column for the second replication. Because this card is from the heart suit, the twenty-sided die was rolled and the resulting roll was 5, which has been recorded in the *Dice Roll* column for the second replication. The remainder of the table was filled in for the remaining eight replications of this process.

Table 7.8 An example of ten replications of the collaborative activity. For the *Card* column the card value (A,2–10,J,Q,K) is given first and the suite second (C = club, D = diamond, H = heart, S = spade). The first card, marked by 5S represents the five of spades.

Replication	Card	Dice Roll	Treatment	Outcome
1	5S	3	N	C
2	2H	9	O	N
3	KC	8	N	N
4	QD	3	O	C
5	KD	16	O	N
6	AC	4	N	C
7	8D	20	O	N
8	10H	18	O	N
9	4D	19	O	N
10	9C	2	N	C

The next step is to translate the card draws and dice rolls into what is simulated in our medical study. Each replication represents an individual in the study. The suit of the card draw represents whether they are given the standard treatment or the new treatment for their medical condition. Each replication where a drawn card is either a club or spade corresponds to individuals who received the new treatment. Therefore, each time either a club or a spade is drawn, write N in the *Treatment* column of the data table in the line corresponding to that replication. The remaining replications, where either a diamond or a heart is drawn, corresponds to individuals who received the old treatment. Therefore, each time either a diamond or a heart is drawn, write O in the *Treatment* column of the data table in the line corresponding to that replication. In Table 7.8 we can observe that the first card was a spade so that N has been written in the *Treatment* column for the first replication. Similarly, the second card was a heart so that O has been written in the *Treatment* column for the second replication. The remaining replications in the table are treated in a similar manner.

Next, we need to determine whether the patient is cured or not. This will be determined by the dice roll. Regardless of whether the ten or the twenty-sided die is used we will conclude that the individual is cured of the medical condition if the dice roll is 7 or lower. For each replication where the dice roll is less than or equal to 7, write C in the *Outcome* column of the data table to symbolize that the individual is cured. If the die roll is 8 or larger, write N in the *Outcome* column of the data table to symbolize that the individual is not cured. In Table 7.8 we can observe that the first die roll was 3 so that C has been written in the *Outcome* column for the first replication. Similarly, the second die roll was 5 so that C has been written in the *Outcome* column for the second replication. For the third replication the die roll was 8 so that an N has been written in the *Outcome* column for the third replication. The remaining replications in the table are treated in a similar manner.

Calculations:

1. After the data table has been completed, construct a table of the same form as Table 7.3. Use the old treatment group as the control group and the new treatment group as the treatment group. Hence, using the notation in Table 7.3, the value of a, will be equal to the observed number of times that a person with the new treatment was not cured, the value of b will be equal to the number of times that a person with the new treatment was cured, the value of c will be equal to the number of times that a person with the old treatment was not cured, and the value of d will be equal to the number of times that a person with the old treatment was cured. Table 7.9 shows this table for the example data given in Table 7.8.

Table 7.9 Data for comparing the cure of the condition for the new treatment and the old treatment groups for the data given in Table 7.8.

	Condition Cured?	
	No	Yes
New Treatment	1	3
Old Treatment	5	1

2. Next, we will compute the probabilities and the odds that the old and new treatments cured the medical condition based on what was observed in the experiment. To compute the observed probability that the new treatment cured the medical condition, take the number of individuals cured by the new treatment and divide it by the number of people who received the new treatment. A similar calculation is used to compute the observed probability that the old treatment cured the medical condition. For the example data given in Tables 7.8 and 7.9, the observed probability that the new treatment cured the medical condition is $3/4$, and for the old treatment the observed probability is $1/6$. Using the methods outlined in this chapter, the odds that the new treatment cured the medical condition are 3 to 1, and the odds that the old treatment cured the medical condition are 1 to 5.
3. Finally, we will compute the observed odds ratio for not being cured with the new treatment compared to the old treatment using the formula given in this chapter. For the example data given in Tables 7.8 and 7.9, the observed odds ratio is

$$\text{odds ratio} = \frac{a \times d}{b \times c} = \frac{1 \times 1}{3 \times 5} = \frac{1}{15}$$

Questions:

1. Because we know how the random mechanisms work that created the data, we know the values of all the probabilities associated with the different outcomes in the data. Recall that everyone received the new treatment if the corresponding card drawn from the deck was a club or a spade, and they received the old treatment otherwise. Assuming that the cards are drawn at random and that each card has an equal chance of being drawn for everyone, what are the probabilities associated with being assigned the new and the old treatments? Convert these probabilities into chances and odds.
2. Suppose that an individual has been assigned to the new treatment. Based on how the data was generated, what is the probability that they will not be cured? What is the probability that they will be cured? Convert these probabilities into chances and odds.
3. Suppose that an individual has been assigned to the old treatment. Based on how the data was generated, what is the probability that they will not be cured? What is the probability that they will be cured? Convert these probabilities into chances and odds.
4. Using the probabilities and odds given above, compute the odds ratio for not being cured with the new treatment compared to the old treatment using the formula given in this chapter.
5. Now compare the probabilities, odds, and odds ratio computed from the observed data with what was computed based on the random model in the problems above. Comment on the differences and explain why the values computed from the observed data differ from what was computed using the true random model for the data.
6. What do the odds ratios indicate about how well the new treatment does when compared to the old treatment? Does this conclusion make sense based on how the date were created?

This page titled 7.10: Collaborative Activity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

7.11: References

Abney, W. 2019. *Random Destiny: How the Vietnam War Draft Lottery Shaped a Generation*. Vernon Press. https://vernonpress.com/book/225?srs...LvVk5FeIooL-j0

Anderson, T. W., D. B. Reid, and G. H. Beaton. 1972. "Vitamin C and the Common Cold: A Double-Blind Trial." *Canadian Medical Association Journal* 107, no. 6, 503–508. https://www.cmaj.ca/content/107/6/503

Balagué, F., G. Dutoit, and M. Waldbuger. 1988. "Low Back Pain in Schoolchildren. An Epidemiological Study." *Scandinavian Journal of Rehabilitation Medicine* 20, 175–179. https://pubmed.ncbi.nlm.nih.gov/2976526/

Balagué F., M. L. Skovron, M. Nordin, G. Dutoit, L. R. Pol, and M. Waldburger. 1995. "Low Back Pain in Schoolchildren. A Study of Familial and Psychological Factors." *Spine* 20, no. 11, 1265–1270. https://doi.org/10.1097/00007632-199506000-00012

Beltrami, Edward. 2020. *What is Random? Chance and Order in Mathematics and Life*. Springer. https://link.springer.com/book/10.10...799-2#overview

Bittoni, M. A., D. P. Carbone, and R. E. Harris. 2024. "Vaping, Smoking and Lung Cancer Risk." *Journal of Oncology Research and Therapy* 9: 10229. https://doi.org/10.29011/2574-710X.10229

Bloom, D. M. 1973. "Birthday Problem." *American Mathematical Monthly* 80, no. 10, 1141–1142. https://doi.org/10.2307/2318556

Burton, A. K., R. D. Clarke, T. D. McClune, and K. M. Tillotson. 1996. "The Natural History of Low Back Pain in Adolescents." *Spine* 21, no. 20, 2323-2328. https://doi.org/10.1097/00007632-199610150-00004

Cahn, Lauren. 2024. "These Are the Most Common Birthdays in the U.S." *The Readers Digest*. https://www.rd.com/article/september...r-birth-month/

Clark, Matthew P. A. and Brian D. Westerberg. 2009. "How Random is the Toss of a Coin?" *CMAJ* 181, no. 12, E306–E308. https://doi.org/10.1503/cmaj.091733

Collins L., A. M. Glasser, H. Abudayyeh, J. L. Pearson, and A. C. Villanti. 2019. "E-Cigarette Marketing and Communication: How E-Cigarette Companies Market E-Cigarettes and the Public Engages with E-Cigarette Information." *Nicotine & Tobacco Research* 21, no. 1, 14–24. https://doi.org/10.1093/ntr/ntx284

Erikson, Robert S. and Laura Stoker. 2011. "Caught in the Draft: The Effect of Vietnam Draft Lottery Status on Political Attitudes." *American Political Science Review* 102, no. 2, 221–237. https://doi.org/10.1093/poq/nfp024

Fienberg, Stephen E. 1971. "Randomization and Social Affairs: The 1970 Draft Lottery." *Science* 171, no. 3968. https://doi.org/10.1126/science.171.3968.255

Hestbaek L., C. Leboeuf-Yde, K. O. Kyvik, and C. M. Manniche. 2006. "The Course of Low Back Pain from Adolescence to Adulthood. Eight-Year Follow-Up of 9600 Twins." *Spine* 31, no. 4, 468–472. https://doi.org/10.1097/01.brs.0000199958.04073.d9

Jones, M. A., G. Stratton, T. Reilly, and V. B. Unnithan. 2004. "A School-Based Survey of Recurrent Non-Specific Low-Back Pain Prevalence and Consequences in Children." *Health Education Research* 19, 284–289. https://doi.org/10.1093/her/cyg025

Kazemi, S., 2025. "The Prevalence and Risk Factors of Low Back Pain in Children and Adolescents in Iran: A Systematic Review." *International Journal of Musculoskeletal Pain Prevention* 5, no. 1, 301–309. https://doi.org/10.52547/ijmpp.5.1.301

Kerr, Benjamin, Margaret A. Riley, Marcus W. Feldman, and Brendan J. M. Bohannan. 2002. "Local Dispersal Promotes Biodiversity in a Real-Life Game of Rock–Paper–Scissors." *Nature* 418, no. 6894, 171–174. https://doi.org/10.1038/nature00823

Kirkup, Benjamin C. and Margaret A Riley. 2004. "Antibiotic-Mediated Antagonism Leads to a Bacterial Game of Rock–Paper–Scissors in Vivo." *Nature* 428, no. 6981, 412–414. https://doi.org/10.1098/rspb.2019.2754

Leboeuf-Yde C. and K. O. Kyvik. 1998. "At What Age Does Low Back Pain Become a Common Problem? A Study of 29,424 Individuals Aged 12–41 Years." *Spine* 23, no. 2, 228–234. https://doi.org/10.1097/00007632-199801150-00015

Lewin-Epstein, Ohad and Lilach Hadany. 2020. "Host–Microbiome Coevolution Can Promote Cooperation in a Rock–Paper–Scissors Dynamics." *Proceedings of the Royal Society B*, 287, no. 1920, 20192754. https://doi.org/10.1098/rspb.2019.2754

Li, Q., L. Peng, Y. Wang, Y. Yang, and Z. Wang. 2024. “Risk Factors for Low Back Pain in the Chinese Population: A Systematic Review and Meta-Analysis.” *BMC Public Health* 24, no 1, 1181. https://doi.org/10.1186/s12889-024-18510-0

Menezes, J. 2021. “Antipredator Behavior in the Rock-Paper-Scissors Model.” *Physical Review E* 103, no. 5, 052216. https://doi.org/10.1103/PhysRevE.103.052216

Munford, A. G. 1977. “A Note on the Uniformity Assumption in the Birthday Problem.” *The American Statistician* 31, no. 3, 119. https://doi.org/10.1080/00031305.1977.10479214

Nunnikhoven, Thomas S. 1992. “A Birthday Problem Solution for Nonuniform Birth Frequencies.” *The American Statistician* 46, no. 4, 270–274. https://doi.org/10.1080/00031305.1992.10475901

Ometto, D. L. A. 2016. *Freedom and Self-Knowledge*. PhD thesis, University Utrecht. https://dspace.library.uu.nl/bitstre...393/Ometto.pdf

Peterson, I. 2001. “Lava Lamp Randomness.” *Science News*. May 2, 2001. https://www.sciencenews.org/article/...amp-randomness

Radelet, M. L. 1981. “Racial Characteristics and the Imposition of the Death Penalty.” *American Sociological Review* 46, no. 6, 918–927. https://doi.org/10.2307/2095088

Reichenbach, Tobias, Mauro Mobilia, and Erwin Frey. 2007. “Mobility Promotes and Jeopardizes Biodiversity in Rock–Paper–Scissors Games.” *Nature* 448, no. 7157, 1046–1049. https://doi.org/10.1038/nature06095

Rust, P. F. 1976. “The Effect of Leap Years and Seasonal Trends on the Birthday Problem.” *The American Statistician* 30, no. 4, 197–198. https://doi.org/10.1080/00031305.1976.10479179

Salminen J. J., J. Pentti, and P. Terho. 1992. “Low Back Pain and Disability in 14-Year-Old Schoolchildren.” *Acta Pædiatrica* 81, no. 12, 1035–1039. https://doi.org/10.1111/j.1651-2227.1992.tb12170.x

Sapru S., M. Vardhan, Q. Li, Y. Guo, X. Li, and D. Saxena. 2020. “E-Cigarettes Use in the United States: Reasons for Use, Perceptions, and Effects on Health.” *BMC Public Health* 20, 1518. https://doi.org/10.1186/s12889-020-09572-x

Sato T., T. Ito, T. Hirano, O. Morita, R. Kikuchi, N. Endo, and N. Tanabe. 2011. “Low Back Pain in Childhood and Adolescence: Assessment of Sports Activities.” *European Spine Journal* 20, 94–99. https://doi.org/10.1007/s00586-010-1485-8

Schreiber, Sebastian J. and Timothy P. Killingback. 2013. “Spatial Heterogeneity Promotes Coexistence of Rock–Paper–Scissors Metacommunities.” *Theoretical Population Biology* 86, 1–11. https://doi.org/10.1016/j.tpb.2013.02.004

Sinervo, Barry and Curt M. Lively. 1996. “The Rock–Paper–Scissors Game and the Evolution of Alternative Male Strategies.” *Nature* 380, no. 6571, 240–243. https://doi.org/10.1038/380240a0

Tashian, Carl. 2017. “A Brief History of Random Numbers.” https://tashian.com/articles/a-brief...andom-numbers/

Taimela S., U. M. Kujala, J. J. Salminen, and T. Viljanen. 1997. “The Prevalence of Low Back Pain Among Children and Adolescents. A Nationwide, Cohort-Based Questionnaire Survey in Finland.” *Spine* 22, no. 10, 1132–1136. https://doi.org/10.1097/00007632-199705150-00013

Vulovic, Vladimir Z. and Richard E. Prange. 1986. “Randomness of a True Coin Toss.” *Physical Review A* 33, no. 1, 576. https://doi.org/10.1103/PhysRevA.33.576

This page titled 7.11: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

CHAPTER OVERVIEW

8: Observing Data from Populations

“By a small sample we may judge of the whole piece.”

Miguel de Cervantes

8.1: What to Expect
8.2: Motivation
8.3: The Concept of Sampling
8.4: Randomness and Observing Data
8.5: More Complicated Sampling Methods
8.6: Nonresponse
8.7: Critical Analysis
8.7.1: The Achievement Gap
8.7.2: Tobacco Retailer Density
8.7.3: Gender and Race-Related Stress
8.7.4: Sampling Animals in Populations
8.7.5: Sports Activities and Lower Back Pain in Adolescence
8.7.6: Sleep and College Students
8.7.7: Visual Exercise Environments and Treadmill Running
8.8: Summary
8.9: About the Quotation
8.10: Practice and Exploration
8.11: Collaborative Activity
8.12: References

This page titled 8: Observing Data from Populations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.1: What to Expect

This chapter considers how researchers obtain the data that they use in their studies. Of essential importance is the fact that in most cases it is impractical to observe all the individuals or items in a population. Researchers then use a sample, a small part of the population, for their studies. It is from this sample that researchers use statistical methods to make conclusions about the entire population. We will demonstrate that it is crucial to take this sample from the population according to certain principles, or the results of the corresponding study will not be scientifically valid. The easiest and most effective methods for getting good samples are usually based on algorithms that choose individuals or items for the sample based on a random mechanism.

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

1. Understand the potential hazards of sampling from a population.
2. Define and understand the implications of sampling bias.
3. Understand the concept of random sampling.
4. Define and understand the implications of response bias.
5. Understand more advanced sampling techniques such as cluster sampling and stratified sampling.

This page titled 8.1: What to Expect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.2: Motivation

An empirical study of a problem begins with identifying a hypothesis. Once this has been done, the population corresponding to the hypothesis needs to be carefully identified, and the characteristics of the individuals that should be observed is determined. At this point, the researcher is ready to observe data. The question of how this is done can be the most difficult part of any study. Earlier we considered the nature of data, with an emphasis on what can be observed and how measurements can be taken from an individual in a population. The question considered in this chapter is how researchers decide who is observed in the population.

To serve as a hypothetical example, consider the problem of attempting to determine if there is a difference in perceived barriers to voting between the racial groups in a small city. As a working hypothesis, the researcher may feel that populations of color may perceive larger barriers to voting than the white population in the city. With this working hypothesis in mind, the researcher must now carefully define the population of interest while at the same time refining the working hypothesis of the study.

It is immediately apparent that there is a dynamic property to the hypothesis. Are the perceived barriers based on previous experiences, that is, participation in previous elections, or are they based on what is perceived for an upcoming election? This potential difference in viewpoint may be important to the interpretability of the conclusions of the study, and may have a fundamental effect on how the population is defined. For the sake of this example, suppose that the researcher has decided to focus on the most recent previous election. The researcher must also decide whether to focus on people who voted, attempted to vote, are registered voters, or voters who are eligible to vote whether they are registered or not. This is an important question because a person who did not attempt to vote may have done so because they perceived significant barriers to voting. Further, a person night not even register to vote if they perceive significant barriers to voting. In this case the researcher may decide to consider all voters who were registered to vote at the time of the previous election, while acknowledging that people who did not register to vote will be excluded. The hypothesis is now refined to consider the question of whether there are differences in the perceived barriers to voting for those who were registered to vote in the previous election between a specified set of racial groups in a small city. This process may continue through several more iterations until a very specific population has been identified.

Now that this population has been identified for our example, how does the researcher observe data in the population? There are several important issues with observing data from a population that must be considered by researchers. How do researchers know who is in the population? How do they contact and communicate with the individuals in the population? What if the population is too large for them to handle? What do they do about people whom they cannot contact or who do not want to participate in the study?

While several of these questions are considered in this chapter, the major question we will tackle is: What if the population is too large to contact everyone? Considering the example as a backdrop, suppose that the small city has around 15,000 registered voters. If the researcher could contact each of these people, and if they all decided to participate in the study, then the researcher could get a complete view of how the registered voters in the city perceived voting barriers for the election of interest.

A few calculations will show that this task is daunting even if everyone is accessible and cooperates. If the researcher has thirty days to complete the data collection part of the study, then they would need to talk to $15,000 \div 30 = 500$ registered voters per day to complete the study. Suppose that if the survey takes fifteen minutes to complete, that corresponds to 125 hours of interviews per day.

If the interviewers work in eight-hour shifts, the researcher would need to hire about sixteen interviewers to work every day for 30 days, assuming no down time between interviews. If an interviewer makes $15 per hour, the researcher would need to spend $56,250 just to pay the interviewers. This cost does not include materials, computer software and hardware, analysts, and the many other associated costs with conducting a survey.

This small example demonstrates only one of the problems with attempting to observe everyone in a population. There is a myriad of other potential problems. People will have moved, they may not be at home when you try to contact them, and many will not want to bother with taking a survey. In the absence of observing an entire population, we are necessarily obligated to observe only some of the population and hope that what we observe is representative of what the entire population looks like. This concept is called *sampling*.

This page titled 8.2: Motivation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.3: The Concept of Sampling

Suppose that we have identified a population for a research study. As discussed in the previous section, if we could contact everyone in the population, and if we could observe all our measurements of interest on each individual or item in the population, then we would definitively know the characteristic of interest, or parameter, for the population. The process of observing an entire population is called a census.

Definition: Census

Observing every individual or item in a population is called a *census*.

In the United States a type of census is conducted every ten years. Article I, Section 2, Clause 3 of the United States Constitution states:

> *The actual enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.*

The need for the census every ten years in the United States is for the express purpose of apportioning representation in the House of Representatives, though the demographic data subsequently gathered from the census is used for policymaking and social research. The key words in this clause with respect to the census are *actual enumeration*, which means that everyone should be counted in the census. This is a monumental task. According to the U.S. Census Bureau, in 2010 there were 635,000 individuals employed to obtain the census data at a cost of about thirteen billion dollars (see census.gov for more facts of the U.S. Census).

On a much smaller scale, a census is conducted anytime a course instructor reports the average grade on a class exam. The population consists of all the students in the course, where the measurement of interest for each student is the grade on the exam. The parameter is the average of all the scores on the exam. An example of this type of hypothetical census is discussed earlier where the population parameters for two small classes are known because all the student grades have been observed.

As an example, consider a major concern in education. Many studies have demonstrated that African American students have lower grade point averages than their white counterparts after the first year of college (Mannan et al. 1986; Spenner et al. 2004; Jay and D'Augelli 1991; Paige and Witty 2010; Condron et al. 2013; Lorah and Ndum 2013; Miksic 2014). This is known as the *achievement gap*. This problem is important as success during the first year of college is associated with measures of long-term success such as graduation rates (Gershenfeld et al. 2016; Oguntunde et al. 2018; Olneck 2005).

This type of study is quite frequently achieved using a full census of the populations involved. Many universities and colleges provide real-time assessment of grade point averages so that the achievement gap between different racial and ethnic groups can be monitored. These studies are used to shape policy, assess programs, and design intervention methods to aid in student achievement. A census is practical in these cases because the university has easy electronic access to all the records of the students, and a statistician or data scientist can produce the relevant average of the grade point averages based on the complete census of all the students. However, the results of any study would only apply to the students at that college or university.

Now consider a researcher attempting to do the same type of study on a much wider population. For example, suppose that we are interested in comparing the achievement gap between community colleges and universities at the end of the first year. The very thought of attempting to contact everyone, and moreover, obtaining sensitive information like grade point averages may seem overwhelming, and indeed it is. Not only would the process be time consuming and expensive, but there would also be serious legal issues associated with attempting to get information on grade and race, unless the data were coded in some way. So how do researchers perform studies like this in a way that the results provide more universal conclusions?

Let us examine a study of overall college performance, which includes observations on student demographics, pre-college academic and racial experiences, and in-college attitudinal and behavioral characteristics (Nettles et al. 1986). The conclusions of the study are based on a survey of 4,094 students and 706 faculty from 30 colleges and universities in the southern and eastern regions of the United States. The first thing we notice is that the researchers did not observe the characteristics of all the students

and faculty at the colleges and universities. In fact, they only observed the characteristics of a relatively small proportion of these students and faculty. This is called a sample from the population.

Definition: Sample

A *sample* is any observed part of a population.

Using a representative sample is how the researchers in this study were able to study these effects without getting data from every student from 30 colleges and universities in the southern and eastern regions of the United States. The researchers only observed a small part of the population.

The concept of sampling is quite old (Stephan 1948). One can imagine that any chef, upon preparing a meal for guests, will try some of what was prepared to ensure quality. Many avid readers will scan a few pages of a book to see if they think they will enjoy reading it. These are both examples of sampling. While a census is mandated by the constitution in the Unites States, sampling was often used to estimate populations in other countries. The first census in Great Britain occurred 1801, and prior to that time the population was estimated using sampling techniques. Sampling continued to be used throughout the 19th century for agricultural crop and livestock estimates, economic statistics, social surveys, and health surveys. The formal mathematical development of statistical inference in the early part of the 20th century also formalized methods for sampling. A few of these methods will be highlighted in this chapter.

Given that not all the individuals in the samples were observed, how can the researchers conclude that the trends observed in the data can be generalized to the entire population? The short answer is that they can never be completely sure the data they observe in a sample from a population reflects what is going on in the entire population. How well the conclusions of the study accurately exemplify the population depends on how well the sample reflects the overall characteristics of the population. If the characteristics of a sample are a good reflection of the characteristics of a population, then such a sample is said to be representative.

Definition: Representative Sample

A sample from a population is *representative* if the characteristics of the sample are like those of the population.

Suppose that a friend brings you a puzzle composed of 50 pieces with a blank box. That is, you have no idea what picture the puzzle would make if you put all the pieces together. Your friend will let you look at ten pieces, from which you will try to describe the picture that the puzzle represents. You friend hands you the ten pieces given in Figure **8.3.1**. What do you guess? You can only use the information from the pieces you observe. They are all shades of blue or white, and a couple of the pieces show what appears to be clouds. At this point your best guess may be, since there is no other information available, that the picture is of a blue sky with clouds.

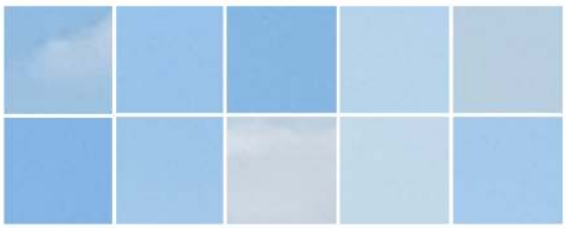


Figure **8.3.1**: Ten pieces from the unknown puzzle picture (Public domain image created by Alan M. Polansky).

Now suppose your friend had given you the ten pieces shown in Figure 8.2 instead. Now there is some variety in the pieces that you observe. It may be difficult to see, but two pieces show bare tree branches with a sky background, one piece shows trees with dense leaves at a distance, and two other show water ripples. Now we could guess that the picture is of a lake or pond with dense trees in the background and a bare tree closer in the foreground. Since the trees in the background have lush dark green leaves, we could guess that the tree in the foreground is, perhaps, dead.



Figure **8.3.2**: Ten other pieces from the unknown puzzle picture (Image created by the author).

The full picture of the puzzle is given in Figure 8.3. This is a dead tree in the middle of Shabbona Lake in northern Illinois, approximately seventy miles west of Chicago, an artificial lake created in 1975 by damming a tributary of the Fox River. Its name derives from the Potawatomi leader Shabbona. The tree was on the land when the land was flooded by the dam, and the preserving nature of the lake water has kept the tree standing for over 45 years. At this point you might rightly say that the pieces selected in Figure 8.1 were not selected fairly, as they were all taken from an area of the picture that with just sky, no trees or water. This is what happens when a researcher chooses a non-representative sample from a population. You get a distorted picture of what the population looks like, and the conclusions that you make can be way off. Hence the puzzle pieces selected in Figure 8.1 are not a representative sample of the corresponding population.



Figure **8.3.3**: The actual picture that the puzzle pieces were taken from (photograph by author).

The pieces selected in Figure 8.2 give us much more accurate information about what the picture looks like. Note that we cannot reconstruct the entire picture from those pieces. For example, none of the pieces we observe show the reflection of the dead tree in the water, so we would not be able to guess that the dead tree is close enough to the lake to have its reflection in the water. Hence, we can still be fooled, but nonetheless we were able to obtain quite a bit of information. One can even notice one piece with the leafy trees in the distance, reflective of the fact that trees takes up a small portion of the entire picture, whereas most of the observed pieces have either water or sky, which takes up most of the picture. This is an example of a sample that could be representative.

Now let us consider a slightly more practical example of sampling. Suppose that a professor is giving a final exam in a large lecture class of 250 students. As the students turn in the exam, the papers are stacked in the order in which they are turned in. That is, the students who turned the exam in first are at the bottom of the stack and the students who turned the exam in last are at the top of the stack. The department chair catches the professor in their office right after the exam and is anxious to know how the students did on the exam. The professor tells the chair that there are 250 exams, and it will take some time to grade them, but that you can take a sample of ten exams, grade them in an hour, and report to the chair how those ten students did on the exam. Now the professor has an important choice to make. Let us suppose that they want to give a clear and fair indication of how the class did to the chair. Otherwise, reporting too good or too bad results will give a biased version of what the chair will find out later when all the exams have been graded. How should the professor choose the exams?

One obvious choice is to take ten exams from the top of the stack and grade them. But is this a good idea? Specifically, will this method of choosing the exams provide a fair representation of how the class did as a whole? That is, would this sample be representative? The main question here is whether the order in which the exams were turned in is associated with the students' performance on the exam. It may be argued that students who are less familiar with the material may take longer to complete the exam, while it could also be the case that some students who perform better may be more careful, may have more to write, and will take longer to complete the exam. The key here is that the professor really doesn't know if these two things are related, and

therefore they need to be careful. If there is a relationship, any choice made using turn-in order may give a biased idea as to how the class did.

To see how this could happen, consider the exam scores given in Table 8.1. This is a contrived example of a class of twenty-five students who took an exam with exam scores given in the order that they were turned in. The scores were simulated on a computer in such a way that the exam scores tend to be higher for exams that are turned in later. Suppose that the professor graded the first five exams that were turned in. The average score on the first five exams is equal to

$$\frac{1}{5} \times (65 + 76 + 74 + 58 + 68) = 68.2, \tag{8.3.1}$$

which indicates and average score in the D range. Now suppose that we graded the last five exams that were turned in. The average score on the last five exams is equal to

$$\frac{1}{5} \times (95 + 81 + 97 + 80 + 97) = 90.2, \tag{8.3.2}$$

which indicates and average score in the A- range. But neither of samples provides a good indication of how the class did as a group, as neither sample is representative. If the average is computed on all the grades, then we get 78.0.

Table 8.1. Twenty-five exam scores simulated in a way that the exam scores tend to increase with the time it took for the student to complete the exam.

Order	Score
1	65
2	76
3	74
4	58
5	68
6	59
7	74
8	71
9	67
10	74
11	89
12	74
13	75
14	75
15	74
16	79
17	88
18	87
19	82
20	91
21	96
22	81
23	97
24	80
25	97

Of course, the professor could attempt to construct a representative sample in a very contrived way, so that the average comes out near the true average. But this would require the professor to know all the exam scores in advance, but such an approach is impractical. In essence, the professor needs a way to pick the exams to grade using a method that is not related in any way to the

trend in the exam scores. The answer that statisticians and data scientists have used for the last century is to use a random method of picking items for a sample. If the random mechanism is chosen carefully, such a method can be an effective and relatively simple method for choosing items to observe in a sample.

This page titled 8.3: The Concept of Sampling is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.4: Randomness and Observing Data

The concept of using a random mechanism to aid in obtaining a representative sample dates to the early twentieth century (Stephan 1948). Early examples of using random sampling methods include the work of William Sealy Gosset (Figure **8.4.1**). Gosset worked for the Guinness Brewery in Dublin, Ireland, and was a pioneer in using statistical methods to solve manufacturing problems. In his research, the concept of sampling items at random verified important theoretical calculations that became what is known as Student's t test (Student 1908). The technique that Gosset used was based on drawing cards from a large bin. This research was published in 1908. Another early example of the use of a random mechanism in sampling was in research published by the English statistician and economist Sir Arthur Lyon Bowley (Figure **8.4.2**) in 1906. He studied bond rates by taking a sample where the random mechanism was based on the final digits in one of the tables in the *Nautical Almanac* (Bowley 1906).



Figure **8.4.1**: William Sealy Gosset (1876–1937) was an English statistician, chemist and brewer who for the Guinness Brewery. He and was a pioneer in the development of modern statistical methods (Public domain image).



Figure **8.4.2**: Sir Arthur Lyon Bowley (1869–1957) was an English statistician and economist who pioneered the use of sampling techniques in social surveys (Public domain image).

How is it possible that choosing a sample randomly will help researchers obtain a sample that is representative of the population? In the example of choosing which exams should be graded, we pointed out that what was really required was a method of choosing the exams to grade in a way that is not related to any possible trend in the exam scores. How can the professor guard against any type of trend when there is no way to know what the trends may be? This is where the idea of randomness is helpful.

Suppose the professor could randomly choose five exams in such a way that each exam has an equal chance of being selected. If that could be accomplished, then the method of selection would be completely unrelated to the order in which the exams were turned in. In fact, this method of selection is unrelated to any possible trend that could exist in the exam scores.

The professor could implement taking such a sample by taking twenty-five index cards and numbering them from one to twenty-five. They could then shuffle the deck many times to ensure that the ordering of the cards was random and that each ordering was equally likely. The first five cards could then be selected and the exams corresponding to the number on the cards would be selected for the sample. This type of sample is known as a simple random sample.

Definition: Simple Random Sample

A *simple random sample* is any random method of choosing a sample from a population in such a way such that every individual or item in the population has an equal chance of being selected.

The method of shuffling a deck of playing cards in card games is used to ensure that every possible selection of cards is equally likely. When you pick a card from the top of a well shuffled deck, each card has an equal chance of being selected. In the same way, shuffling the index cards with the exam numbers on them ensures that each exam has an equal chance of being selected for the sample.

In the past, statisticians have used physical objects such as decks of cards to implement random sampling in practical situations. The use of large tables of random sequences was also used before the advent of digital computers. Practical applications for modern applications depend on computer-based algorithms.

The advantage of simple random sampling is that the method guarantees that the corresponding sample will be very likely to be a representative sample. It may happen every time, but the probability of this occurring is very small. In many cases, using a simple random sample is the standard method for taking a sample from a population.

As an example, consider the data given in Table 8.2, which corresponds to a small population of residents in a senior living complex. The data in the table gives the unit number, gender identity, and the average response time in minutes for non-emergency calls to the staff over the past month. The female and non-binary residents contend that the average response time is larger for residents who do not identify as male. Rather than observing the entire population, the management decides to take a simple random sample of five units to study the situation. Since there are twenty units, the manager uses a 20-sided die to choose the units for the sample. The manager rolls the die five times, getting 18, 4, 17, 5, and 4. Noticing that the roll 4 was repeated, the die is rolled once more and 1 is observed. Hence the units 1, 4, 5, 17, and 18 are include in the random sample. If another number had been repeated, the die would have been rolled again until five distinct numbers had been observed. It has been proven mathematically that if the die is fair, this method will result in a simple random sample.

Table 8.2 The unit number, gender identity, and the average response time in minutes for non-emergency calls to the staff over the past month at a senior living complex. The gender identities are coded as F (female), M (male), and N (nonbinary).

Unit	Gender	Response
1	M	11.03
2	N	12.15
3	F	11.21
4	F	11.45
5	F	11.61
6	M	10.83
7	M	10.65
8	N	12.62
9	M	10.53
10	F	11.88
11	F	12.00
12	M	10.34
13	M	10.37
14	N	13.26
15	F	12.01
16	M	10.16
17	M	10.65
18	F	10.90
19	F	10.83
20	F	10.49

For these die rolls, the corresponding simple random sample of data is presented in Table 8.3. How representative is this sample? Let us consider the percentage of each gender identity in the population and in the sample. The population contains eight males, or $8 \div 20 \times 100$ identify as male. The simple random sample contains two males, or $2 \div 5 \times 100$ identify as male. From this viewpoint, the sample seems to be representative. On the other hand, there are no nonbinary people included in the sample.

Table 8.3 The unit number, gender identity, and the average response time in minutes for non-emergency calls to the staff over the past month at a senior living complex included in the simple random sample. The gender identities are coded as F (female), M (male), and N (nonbinary).

Unit	Gender	Response
1	M	11.03
4	F	11.45
5	F	11.61
17	M	10.65
18	F	10.90

Few samples will perfectly represent the population from which they came. The hope when taking a sample is that it represents the characteristics of the population well enough for researchers to answer the questions that they have about the population. Without knowing the entire population, it is very difficult to determine if a sample represents a population. As we shall see later, statistical methodology can help determine the size of the researcher's potential error, and this information determines whether the sample provides sufficient evidence to prove a research hypothesis. In cases where a simple random sample cannot be taken, other sampling methods have been developed.

This page titled 8.4: Randomness and Observing Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.5: More Complicated Sampling Methods

While simple random sampling is both powerful and useful in practice, it is not always practical to implement. The problem is that in some cases the researcher is not exactly sure who the members of the population are, and in these cases, it can be difficult to ensure that all the members have an equal chance of being included in the sample. For example, suppose that a researcher is interested in taking a simple random sample of residents in a small town. A list of residents in the town is not available. But they may have a map that details the location of each residence in the town. In theory, they could take a simple random sample of these homes and talk to the residents. But how do they decide who to speak with in each home? If they randomly choose a resident in each home, then the sample is no longer a simple random sample because some homes have fewer residents than others. What about apartments that are not included on the map? What about people who are homeless, or those who live in nonstandard housing? Without a list of residents in the town, along with their locations, implementing a simple random sample can be very difficult.

To make sampling easier, researchers can divide up a population into groups called clusters. The researcher will then take a simple random sample from these clusters. Once the clusters have been sampled, the researchers can then either take a census within the cluster if it is small enough, or they can take a simple random sample within the cluster.

Definition: Cluster Sample

A sample from a population is a *cluster sample* if the population is first divided into groups called *clusters*. A simple random sample of these clusters is then taken, and individuals within each cluster are observed either through a census or through an additional simple random sample.

The advantage of using a cluster sample is that the population can be divided into smaller and easier-to-manage groups. There are two disadvantages to using a cluster sample. The first is that the potential for statistical errors can be larger when using cluster sample. The second is that the statistical techniques used by researchers must be modified to account for the fact that this type of sampling method was used.

Cluster samples work best when each of the clusters themselves are small representatives of the population. In this way it does not matter so much if not all the clusters are observed. If the clusters are divided in such a way that the clusters represent important differences in the population, then not observing some of these clusters will likely result in a nonrepresentative sample. So, for example, dividing a city by neighborhoods may inadvertently create clusters based on race, which could result in sample that is not representative of the population.

As another example, suppose researchers are planning a cluster sample of a city that is based on census blocks. For simplicity, suppose that these census blocks are known to be approximately representative of the entire population of the city and are appropriate for using as clusters. There are twenty of these blocks numbered 1–20, as indicated in Figure **8.5.1**. A simple random sample of the clusters chooses the blocks numbered 2, 8, 10, 11, and 17. Each block contains roughly 100 residents. Because these blocks are relatively small, the researchers can study the blocks closely and determine who the residents are in each of the blocks, obtaining enough information to take a simple random sample of five residents in each block. This creates a total sample of twenty-five residents from the city based on a cluster sample.

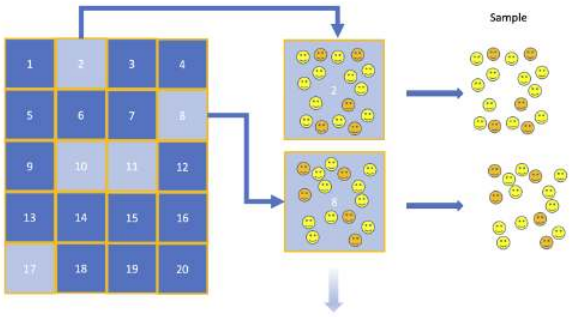


Figure **8.5.1**: A graphical depiction of cluster sampling. The population has been divided into twenty clusters, and a simple random sample of clusters is taken. Only the samples from the first two clusters are depicted. (Public domain image created by Alan M. Polansky)

An alternative to cluster sampling and simple random sampling is stratified sampling. Stratified sampling may look a lot like cluster sampling at first, but there are two major differences, and the purpose of stratified sampling is somewhat different. Like cluster sampling, stratified sampling also divides the population into groups, which in this case are called strata. In contrast to cluster sampling, in which the clusters should be representative of the population, the strata should be as homogeneous as possible. Hence, when using stratified sampling it is advantageous to divide the population into groups containing individuals who are as alike as possible. The other main difference with cluster sampling comes in the second phase of the sampling process. In stratified sampling, a simple random sample is taken within each of the strata, whereas in cluster sampling, only randomly sampled clusters are used.

Definition: Stratified Sample

A sample from a population is a *stratified sample* if the population is first divided into groups called *strata*. Individuals within each *stratum* are observed through a simple random sample.

As with cluster sampling, the corresponding statistical methods need to be modified to take advantage of this methodology, but the resulting conclusions can often be more reliable than even with simple random sampling. The other advantage of stratified sampling is that with cluster sampling, it is often easier to take a simple random sample on smaller groups than it is for the entire group.

For example, suppose researchers are planning a stratified sample of a city. Suppose that there are five residential neighborhoods in the city that are highly segregated so that those living in each neighborhood tend be of the same race or ethnicity and economic class. These blocks are numbered 1–5, as indicated in Figure **8.5.2**. Within these neighborhoods the researchers can determine who the residents are, obtaining enough information to take a simple random sample of five residents in each neighborhood. This creates a total sample of twenty-five residents from the city based on a stratified sample.

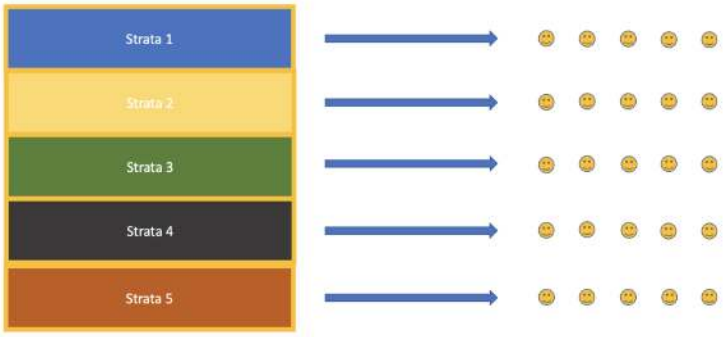


Figure **8.5.2**: A graphical depiction of stratified sampling. The population has been divided into five strata. From each strata a simple random sample of individuals is taken. (Public domain image created by Alan M. Polansky)

There are many other methods used for sampling, including combinations of cluster sampling, stratified sampling, and simple random sampling. These methods can be very effective if they are implemented with care and the corresponding statistical methods are adjusted to reflect how the data were observed. Many surveys are implemented without any specific plan on how the sample will be taken and how the sample may be related to the corresponding population. Lack of care with the sampling phase of the study can yield results that are very difficult to interpret. Furthermore, applying statistical methods to data that have not been observed according to a well understood, random sampling method can be very dangerous. As we shall see in later chapters, statistical methods allow one to assess the risk that one takes from observing a sample instead of taking a census of a population. These risk assessments will not be valid if the correct sampling method has not been used.

This page titled 8.5: More Complicated Sampling Methods is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.6: Nonresponse

Imagine that some friends who are all investment bankers are sitting at a trendy restaurant, hanging out, when one of them gets a phone call. They look at the number, they don't recognize it, but they go ahead and answer. The voice that comes through and asks politely if they would be interested in taking a survey about the economic conditions in their city. They decline and go back to talking with their friends. They never even give it another thought.

What happens for the researcher when this happens is more complicated. The researcher had carefully designed a study along with a plan for sampling people at random from the population. Certainly, the researcher could not have forced anyone to take the survey, and so there is nothing they can do. But now that someone has not responded, the sampling plan has been violated. The researcher has other responses, so does it make a difference?

The answer to that question depends on if the reason a participant does not respond to a survey is related to the issues being studied by the survey. The investment banker is a busy young person with many things going on. The survey they decided not to participate in was about economic conditions in the city. What type of person might take part in such a survey?

Investment bankers may be busy, but other people might not be. For example, it is well known that retired people often have more free time at their disposal and are more likely to take surveys. Do you think that these individuals may respond differently to the survey than someone who is too busy to take the survey? Remember, this survey is about economic conditions, and many retired persons are living on a fixed income. They may have very different views of economic conditions than a young investment banker. This creates a problem in that the survey may end up giving a skewed view of economic conditions if some people are overrepresented in the sample.

Earlier we discussed the presidential election of 1936 where we considered the methods used by the *Literary Digest* magazine. We looked at the political poll that predicted Alf Landon would beat Franklin Roosevelt. The poll was in error partially because the population from which the *Literary Digest* observed their data was not the same as the population they wanted to study. However, this poll also suffered from non-response. The *Literary Digest* sent out 10 million surveys, but only 2.4 million individuals participated in the survey.

Non-response is a common problem in surveys, and special studies have historically shown that the middle class tends to be overrepresented in typical surveys. Other research has shown that individuals who have strong feelings about a particular survey subject are more likely to participate. Hence, surveys on controversial issues are likely to have high response rate from individuals who have extreme views on the topic. Personal interviews have historically had better response rates, but non-response can still be a problem. Many modern surveys are implemented over social networks or other internet-based resources, and researchers are studying the demographic profiles of those who respond and do not respond to these surveys.

If non-response is a problem in surveys, what do researchers do to get reasonable results from their surveys? Statistical researchers have developed sophisticated methods to account for this bias. For example, if it is known that certain segments of the population tend not to respond to surveys, then respondents from those segments may be weighted more heavily in the results of the survey. The researchers need to be careful to figure out how much additional weight to give these responses. Similarly, even when an individual elects not to participate in a survey, some information may be known about them, such as demographic information. Researchers can use this information along with the remaining information about those who responded to the survey to attempt to guess how those individuals would have filled out the survey. These methods are known as imputation methods. Once again, these methods must be used very carefully.

Many surveys do not try to correct for non-response at all, and the results of these studies must be interpreted very cautiously. Even when the response rate is relatively high, non-response bias can have a large effect on the results when those who do not respond differ greatly from those who do respond.

This page titled 8.6: Nonresponse is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.7: Critical Analysis

Many research articles are based on some type of sampling scheme. When assessing the sampling method used in a research article, you should look for methods such as simple random sampling, cluster sampling, stratified sampling, or some combination of these methods. If the research is based on an existing database, you may need to do some additional reading to understand how the data was originally obtained. In any case, if a sampling method that is based on statistical theory is used, some credibility can be afforded to the research study.

One should proceed with caution for any study that does not use careful random sampling techniques. This is particularly true when statistical methods are used to elicit conclusions from the data. Statistical methods rely on careful random sampling techniques to be valid, and a lack of such sampling techniques can render the interpretation of the results of the statistical techniques null.

In summary, the key questions in critically analyzing sampling in a study are:

1. Are individuals observed in the population based on a census or a sample?
2. If individuals were observed based on a census, does it seem reasonable that the authors would have access to all the individuals in the population?
3. If individuals were observed based on a sample, was a random sampling technique such as simple random sampling, cluster sampling, stratified sampling, or some combination thereof used?
4. If a random sampling technique was not used, what method was used? Does it seem likely that the method would provide a representative sample?
5. If a random sampling technique was not used, were the conclusions of the study based on statistical techniques?
6. Overall, does it seem like there may be any bias in the way that the sample was taken?

This page titled 8.7: Critical Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.7.1: The Achievement Gap

The socioeconomic status of parents has been shown to be associated with the school achievement of their children (Abrams and Kong 2012). One study considered the role of race, school location, and parental education on the black–white achievement gap for high school students in the United States (Henry et al. 2020). The particular emphasis of this study was to test the hypothesis that the effect of parental education attainment diminishes in urban and suburban high schools in the United States. The context of this question is the idea that parental education attainment provides further home-based resources for educational opportunities. However, the study shows that this effect is diminished for black students in urban schools.

The data for this study was obtained from the Education Longitudinal Study funded by the United States Department of Education. The Educational Longitudinal Study is one of the main studies of youth education in the United States. The term *longitudinal* means that some students were followed by the study throughout their educational career, creating a storyline of the student for the entire time they are in the public schools. The Educational Longitudinal Study is a nationwide study based on a cluster sample (Ingels et al. 2004). Schools were used as clusters. Of all the schools in the United States, 1,221 were selected from the population of approximately 27,000 schools containing 10th grade students. Because schools can vary in size, the schools were not selected based on a simple random sample. Instead, schools were selected randomly such that larger schools had a greater chance of being selected and smaller schools had a lower chance of being selected. Hence, students from smaller schools would not be overrepresented in the sample. Of the selected schools, 752 participated in the study. These schools provided 10th grade enrollment lists. Next, approximately 26 students per school were selected from these lists based on a simple random sample.

The sampling method for this study is designed well and, implemented by United States Department of Education, should be trustworthy. The response rate among schools may be of some concern. The authors of the study report that extensive data were available for non-responding schools, and these data were used to help reduce potential non-response bias.

This page titled 8.7.1: The Achievement Gap is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.7.2: Tobacco Retailer Density

Earlier we considered a research study that examined the association between disparities in the density of tobacco retailers and the disparities in the prevalence of tobacco use among youth and adults (Glasser et al. 2022). The study focused on the state of Ohio, and to obtain information about cigarette smoking among adults there, the study used data from the 2017 Ohio Behavioral Risk Factor Surveillance System, an annual random-digit dial telephone survey. Random-digit dialing is a methodology of selecting individuals for a survey that essentially calls random valid phone numbers, and can access anyone with a phone number, whether it is publicly available or not. This type of random digit dialing methodology essentially takes a simple random sample of individuals who have phone numbers in the area. While the response rate is not addressed in the research, it appears that careful weighting of the responses was used to reduce the effect of non-response bias. Specifically, the authors state that the data are “a state-representative sample of adults” (Glasser et al. 2022).

This page titled 8.7.2: Tobacco Retailer Density is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.7.3: Gender and Race-Related Stress

Some research has considered the possibility that the impact of race-related stress may depend on gender (Greer et al. 2009). The research is based on a survey of 183 individuals who were self-identified African American students at a mid-sized university. The participants for this study were from introductory undergraduate psychology courses, African American studies courses, and university events in which the audiences were primarily African American students. The participants from the courses received course credit. Those who were recruited at campus events did not receive credit as part of a course but received monetary compensation instead.

The researchers hypothesized that because African American women are exposed to stress associated with both race and gender that African American women would have higher levels of race-related stress and more severe mental health outcomes compared to African American men. However, statistical analyses of the observed data revealed that African American men had higher stress levels for institutional racism than women. Additionally, no gender differences were observed for cultural and individual racism.

The results of this study should be interpreted with caution. No sampling mechanism was indicated in the research paper, so we can surmise that all eligible students in the classes participated on a voluntary basis in exchange for course credit. Similarly, it appears that anyone willing to participate at the targeted campus events was used in the study. As these samples are not random, the applicability of the statistical analysis that follows would need to be interpreted carefully and with caution.

It should be noted that the paper also points out several limitations for the study related to the choice of sampling and population. The sample is made up of African American college students, and the article points out that other adults may have different life stressors. For example, older adults may have to deal with parenting, divorce, and significant financial struggles that may render them more vulnerable to the negative mental health effects of racism. Secondly, the sample for this study was comprised of individuals from the southeastern region of the United States and that future research studies should compare gender differences in race-related stress among African Americans from other geographic locations (Greer et al. 2009).

This page titled 8.7.3: Gender and Race-Related Stress is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.7.4: Sampling Animals in Populations

Earlier we learned about a study that considered how well a sampling method used in wildlife studies performed when applied to studying fish species in their natural habitat. Recall that in this study four lakes with no fish in them were stocked with rainbow trout that had three different growth rates (slow, intermediate, and fast). At the end of the season, fish were sampled from the lakes using experimental nets designed to eliminate bias due to the size of the fish. The researcher hypothesized that the fast-growing fish, independent of their size, would be more likely to be captured in nets. This hypothesis was verified by the results of the experiment.

This study shows that even when one is very careful about sampling from a population, selection bias can still be a problem. The culprit is usually a confounding variable that is unknown to the researcher, or a variable that the researcher cannot control. Researchers must always be aware that some confounding variable might affect what individuals or items will end up in their sample. It may be impossible to correct the sampling bias, but knowing what these confounding variables are may be important to interpreting the results of the study. In the example of sampling trout, the researchers using the capture technique then would need to be aware that any conclusions drawn from the data may be affected by the speed at which the fish grow.

This page titled 8.7.4: Sampling Animals in Populations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.7.5: Sports Activities and Lower Back Pain in Adolescence

In the last chapter we considered a study that investigated the role that sports participation may have in the development of lower back pain in adolescents (Sato et al. 2011). The study considered survey data collected from 26,766 students in Niigata City, Japan. The study included all elementary school students from the fourth to sixth grade and all junior high students from the first to third year, which corresponded to children ranging from 9 to 15 years of age.

This study is notable in that it contacted all the students in the selected grade levels, which is a significant scope for the study. However, the authors of the study point out that non-response is potentially a problem. Even though the survey was sent to every student, those who filled out the survey were limited, and therefore the percentage of valid responses was only 61.3%. In this study, evaluations were performed only for the students who provided valid responses, and therefore, the conclusions regarding the students who gave invalid responses are not available. The results of this study therefore apply to students in the school district who completed the survey. The researchers did not report any additional effort to find out any further information about those who did not complete the survey and therefore it is unknown whether the results for those students would be similar to those who completed the form. Without further information it is not known if completing the survey is related to the prevalence of lower back pain, or participation in sports.

This page titled 8.7.5: Sports Activities and Lower Back Pain in Adolescence is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.7.6: Sleep and College Students

In section 3.7.6 we considered a study that investigated whether an economic incentive could be used to promote healthy sleep habits. The researchers indicated that they recruited a total of 508 participants at the University of Oxford (163 subjects) and the University of Pittsburgh (345 subjects), though there is no specific information as to how these students were recruited. It can be inferred that these students were not part of a random sample, and hence the students probably responded to a recruitment advertisement for the study. The study does involve a small economic incentive that probably helped with recruitment.

This results from this type of study are not based on a random sample from the population of students, so one must be careful in attempting to generalize the findings to the population of all college students. First and foremost, the students in this study were recruited from two specific universities. Second, this was a self-selected sample. That is, the students were recruited and elected to participate in the study on their own. Therefore, if one would like to generalize the results of the study, one would need to argue that the students from these universities who elected to participate in the study are representative of all college students.

The researchers also indicated that not all of the individuals provided usable data for the study, usually due to the monitoring device not functioning properly. This could indicate serious problems with the interpretability of the results if the reason that the data was unusable is related to the sleep habits of the participants in the study. To address this issue, the researchers compared the demographics of the participants who provided usable data against those who did not, and they did not find any difference. This provides evidence that the reason for having unusable data was not related to demographics factors that could be related to sleep habits.

This page titled 8.7.6: Sleep and College Students is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.7.7: Visual Exercise Environments and Treadmill Running

In section 3.7.2 we learned about a study that investigated the effect of visual stimulation on the performance of individuals exercising on a treadmill. This was a research study based on recruitment of participants who, in this case, were individuals who exercised at least 150 minutes per week. At the most, the results of this study are applicable to individuals who exercise regularly and were willing to take part in the study. Unfortunately, few details are given on how the individuals were recruited and what specific population they were recruited from. This raises serious concerns about the general applicability of the study to the larger population.

This page titled 8.7.7: Visual Exercise Environments and Treadmill Running is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.8: Summary

Except in specific cases, observing every member of a population can be a time-consuming and resource-intensive process, though such methods are used in some applications. Sampling part of a population provides a practical alternative if the method of sampling is done correctly. The simplest sampling method from a conceptual standpoint is the simple random sample, which guarantees that each member of the population has an equal chance of being selected for the sample. If the simple random sample is done correctly, the probability that the sample is representative of the population is quite high, and statistical methods can be used to determine how large any errors might be from taking the sample.

If a simple random sample cannot be used, then there are alternative methods for sampling that can make sampling easier and more efficient. Both methods studied in this chapter are based on dividing a population into smaller, easier-to-handle groups. In cluster sampling, the groups should each be representative of the population, and a simple random sample of these groups is used in the study. In stratified sampling, the groups should be homogeneous within the groups, and simple random samples are taken from each group. Both methods allow the researcher to break the population into groups for which a simple random sample may be easier to implement. Stratified sampling has the added benefit that the resulting statistical procedures will provide better results, assuming the methods are applied correctly.

Non-response is a critical problem that can reduce the universality of analysis results even with a sound sampling procedure. The crucial aspect of non-response is whether the reason for individuals electing not to participate in the survey relates to the issues covered in the survey. If non-respondents differ greatly from those who respond, the results of the survey will be biased and the corresponding conclusions from the research study may contain significant errors. When non-response is a problem, researchers often employ techniques to find out as much as they can about the those who elect not to participate.

This page titled 8.8: Summary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.9: About the Quotation

Miguel de Cervantes Saavedra (Figure **8.9.1**) was a Spanish writer and is regarded as the greatest writer in the Spanish language. He is best known for his novel *Don Quixote*, which is often considered to be the first novel. The quote at the beginning of the chapter is taken from this book.



Figure **8.9.1**: Miguel de Cervantes Saavedra (1547–1616) was a Spanish writer best known for his novel *Don Quixote*. He is depicted here at the Battle of Lepanto in a painting by Augusto Ferrer-Dalmau Nieto (Public domain image).

This page titled 8.9: About the Quotation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.10: Practice and Exploration

1. The data in the Table 8.4 represent the entire population of workers at a small industrial plant. The information included in the table corresponds to the employee identification number, age, gender identity, race or ethnicity, and hourly wage of those employed at the company. Write the employee identification number of each employee on separate index cards. Shuffle the cards a few times and then pick the first five cards from the top. The employees corresponding to these identification numbers are an approximate simple random sample of employees from the company. Compute the percentage of each gender identity in the population and in the sample. Are the population and sample percentages approximately the same? Compute the percentage of each race or ethnicity in the population and in the sample. Are the population and sample percentages approximately the same? What do these results say about the representativeness of the sample that you took?

Table 8.4 The population data for the employees at a small industrial plant. Gender identities are coded as F (female), M (male), and N (nonbinary). Race and ethnicities are coded as AA (African American), AS (Asian), HI (Hispanic), NA (Native

American), and WH (White). Salary refers to the hourly salary of the employee.

Number	Age	Gender	Race	Salary
085	55	M	WH	26.52
359	21	N	HI	16.50
246	57	F	HI	21.95
589	25	F	WH	21.00
168	54	M	AS	25.75
204	33	F	HI	20.80
242	52	M	WH	26.45
908	48	M	WH	26.20
033	43	M	WH	26.65
591	31	F	WH	20.45
259	34	F	NA	15.90
498	22	F	WH	19.20
068	48	M	HI	21.65
729	38	F	AA	14.95
433	23	F	WH	26.85
544	31	F	AA	18.20
197	50	M	WH	28.00
414	54	M	WH	26.70
570	57	M	WH	28.80
004	61	F	AA	23.10
014	65	M	WH	30.55
814	65	F	WH	20.80
225	41	M	WH	21.00
423	25	M	WH	26.50
139	64	F	AA	23.60
632	41	M	HI	23.45
071	35	F	HI	23.20
502	51	F	HI	25.55
602	61	M	WH	27.60
464	25	F	AA	15.55
335	34	M	WH	27.05
724	25	F	AA	22.40

628	30	F	AA	15.90
597	34	N	HI	19.90
155	26	F	WH	25.20

2. A student at a small college wants to take a simple random sample of the students at the college. They decide that they will go to the common lunchroom at noon and wait until it is very crowded. They will then randomly pick students based on the available seats in the lunchroom using a simple random sample. Do you think that the resulting sample will be a simple random sample of students at the college? Why or why not?
3. A researcher is interested in establishing a political profile of residents in a large city. To gather data, they decide to stand in front of an independent bookstore with the hope of interviewing residents who are planning to visit the bookstore. Do you think the resulting sample will be representative of the population of residents in the city? Why or why not?
4. A researcher wishes to sample people for a survey as they enter a small political rally. As there is only one entrance, they seat themselves by the door and as each person enters, they flip a coin. If the coin flip shows heads, the person is directed to fill out a survey before entering. Otherwise, the person is not selected, and they can enter the rally without filling out the survey. Will this result in a simple random sample of those attending the rally? Why or why not?
5. A researcher intends to study the number of campus services that students at a large university use with the hopes of testing hypotheses about possible differences due to gender, race, and ethnicity, and sexual orientation. Using the semester schedule, the researcher decides to take a cluster sample with the classes taken to be the clusters. The researcher will then obtain the course roster for a random sample of classes. Within each of these classes, they will then take a random sample of students within each class. Do you think that using the classes as clusters is reasonable for this study? Why or why not?
6. The manager of a community center would like to survey its patrons about what additional needs the center could fulfill. The center offers lunch to those who live in the downtown area at noon, and so they go at lunchtime and randomly select twenty individuals using a simple random sample from those at lunch to take a survey. Do you think that this sample will be representative of the population? Why or why not?
7. The superintendent of a local school district is worried that non-white students in the school district are not being educated to the same level as the white students, creating an achievement gap that the school district must address. To study this issue the superintendent takes a simple random sample of ten students from each school in the district to be included in a study of educational achievement. What type of sampling method is the superintendent using?
8. A professor at a large private university is interested in studying gender pay gaps at their university. Because salary data is not available to them, they decide to randomly sample ten departments on campus and take a simple random sample of five faculty within each department. They will then email these faculty and ask them their salary and other relevant demographic information as part of a survey. What type of sampling method is this? Do you think there will be a problem with non-response? Could the faculty who decide not to take part in the survey be different from those who do in terms of salary?
9. The human resources department of a large company is interested in assessing whether discrimination against members of the LGBTQIA+ community is common at the company. To study this issue the director of the department proposes setting up an office where members of this community can come speak with counselors who will gather their information and assess the discrimination level for that individual. These interviews are to take place during normal working hours, so individuals who wish to participate will need to notify their supervisor that they will be taking part in the study. Do you think this study will have a problem with non-response? Do you think individuals who do not respond to the survey will be different from those who respond? In what ways? Can you propose a better way of getting people to participate in the study?
10. The manager of a survey organization has been put in charge of a survey of potential voters in the mayoral election in a large city. The current mayor has been critical of some of the practices of the police force regarding interactions with people of color. The survey has been designed to measure how this may change support for the mayor as a candidate in an upcoming reelection campaign. The survey will be implemented online, and the manager is aware of the need for a good response rate. The online interface used the title “Opinion Poll on the Policies of the Mayor.” The manager argues that this title is too boring and that

people will not wish to participate in the survey as a result. The manager suggests the alternate title “Do You Care that the Mayor Keeps Criticizing the Police?” Do you think that this new title will increase the response rate? How might this new title affect the participation rate of those who support the mayor? How about those who do not support the mayor? Do you think the title change is a good idea? Why or why not?

11. Think again about the example of the puzzle described in earlier in this chapter. You are shown a fully assembled 1,000-piece puzzle, and you are asked to take a sample of 100 pieces in such a way that another person would have the best chance possible of guessing what the puzzle picture is without seeing the completed puzzle. How would you pick the 100 pieces? Would you use a simple random sample? Would you purposely choose pieces that best represent the picture, or give clues as to what the puzzle depicts? How does this situation differ from choosing individuals in a population for a survey?
12. In Problem 11 you are asked to take a sample of 100 puzzle pieces in such a way that another person would have the best chance possible of guessing what the picture is without seeing the completed puzzle. Now suppose the puzzle was lying flat on its face so that you could not see the picture, only the back of the puzzle. Now how would you pick the 100 pieces? Would you use a simple random sample? How does this situation differ from the one described in Problem 11? Is this situation closer to that of choosing individuals in a population for a survey? How does it differ?
13. Using an online news service, find a news article about an issue that is important to you personally that is based on an empirical analysis. Identify, as clearly as you can, the population that the authors of the news article have used to observe their data and gather information and how that information is gathered. Was a random sampling procedure used? If so, what type of random sampling technique was used? If a random sampling procedure was not used, then how was the data gathered? Do you think the sample is representative? Do you think there may be a problem with bias from non-response?
14. Using a library database, find an academic research article about an issue that is important to you personally that is based on an empirical analysis. Identify, as clearly as you can, the population that the authors of the research article have used to observe their data and gather information and how that information is gathered. Was a random sampling procedure used? If so, what type of random sampling technique was used? If a random sampling procedure was not used, then how was the data gathered? Do you think the sample is representative? Do you think there may be a problem with bias from non-response?

This page titled 8.10: Practice and Exploration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.11: Collaborative Activity

This collaborative activity will explore sampling from a small population. For this activity you will be supplied with a ten-sided die and a twenty-sided die. The population that you will be sampling from is represented by the entries in Table 8.5. There are 200 cells in this table representing 200 individuals that work at a small manufacturing company. Within each cell is a salary that has been coded in thousands of dollars and a designator as to the type of job that the individual performs. The job designators are *M* for management, *E* for engineering, *A* for assembly line worker, *D* for data processor, and *H* for human resources. For example, the entry A42 is used to indicate an assembly line worker whose salary is $42,000 per year, rounded to the nearest thousand dollars. The data have been assembled in the table so that the first four rows are all management employees, the next first four rows are all engineers, the next six rows are assembly line workers, the next four rows are data processors and analysists, and the final two rows are human resources workers.

Table 8.5 The hypothetical population to be used for the collaborative activity.

M102	M98	M99	M98	M114	M99	M98	M116	M105	M111
M106	M127	M100	M111	M117	M111	M117	M107	M130	M119
M109	M97	M119	M114	M123	M110	M119	M100	M117	M105
M213	M125	M98	M99	M105	M111	M99	M97	M111	M107
E106	E91	E94	E110	E106	E97	E105	E112	E117	E104
E107	E109	E94	E98	E99	E91	E98	E98	E97	E118
E96	E100	E98	E99	E111	E100	E112	E122	E104	E96
E91	E104	E91	E104	E107	E111	E112	E110	E109	E110
A45	A48	A45	A47	A47	A48	A44	A46	A47	A47
A45	A47	A42	A44	A47	A46	A51	A54	A48	A48
A49	A44	A46	A50	A52	A43	A46	A49	A50	A49
A48	A44	A47	A49	A46	A46	A50	A51	A46	A42
A45	A55	A46	A46	A50	A49	A49	A48	A45	A45
A46	A43	A47	A48	A48	A45	A51	A50	A48	A58
D85	D89	D72	D87	D78	D73	D88	D96	D80	D89
D91	D101	D73	D72	D90	D76	D71	D87	D80	D79
D72	D84	D93	D76	D95	D89	D88	D78	D79	D95
D75	D101	D88	D75	D75	D75	D72	D85	D73	D76
H64	H61	H70	H77	H77	H78	H64	H72	H72	H79
H80	H101	H71	H67	H72	H77	H78	H70	H83	H68

Your group will first take a simple random sample of 20 of the 200 individuals at the company. To do this you will need one person to roll the dice and a second to write down the data. To take the simple random sample, use the following process:

1. Roll both dice.
2. Locate the individual in the column of Table 8.5 indicated by the ten-sided die and the row that is indicated by the twenty-sided die. Include this individual in the sample.

3. Check this selection against the individuals already in the sample. If this individual has already been included, re-roll the dice and choose again. If this individual has not been selected yet, proceed to the next step.
4. Write down the data for the selected individual on a table like the one shown in Table 8.6.
5. Repeat this process until you have completed your sample of 20 individuals.

Table 8.6 Data Collection Sheet for the Collaborative Activity

Sample Type	Cluster or Strata	Sample
Simple Random Sample		
Cluster Sample		
Stratified Sample	M	
	E	
	A	
	D	
	H	

As an example, suppose that we begin the process by rolling an 8 on the twenty-sided die and a 9 on the ten-sided die. This corresponds to the E109 entry. This data would then be added to the data sheet, as show in Table 8.7. Next, suppose that we roll a 2 on the twenty-sided die and a 7 on the ten-sided die. This corresponds to the M117 entry. This data would then be added to the data sheet, as show in Table 8.7. Assume the next two rolls are (19,1) and (3,9). Note that we get the same data for the rolls (2,7) and (3,9), but we include both data points. This is because two individuals have the same salary in the same category. Because this corresponds to two different individuals from the population, we include both data points. Suppose the next five rolls are (12,9), (5,3), (18,2), (5,9), and (7,10), with the corresponding data shown in Table 8.7. At this point we roll (5,3). In this case, not only is the data the same but the roll that selected the data is the same as we previously rolled. Because this individual has already been selected for the sample, we skip it and roll again, getting (1,3), which corresponds to the data M99. The process continues in a similar manner yielding the example data shown in Table 8.7.

Table 8.7 Data Collection Sheet for the Collaborative Activity

Sample Type	Cluster or Strata	Sample
Simple Random Sample		E109, M117, H64, M117, A46, E94, D101, E117, E96, M99, M106, A45, A47, M102, M114, M213, M109, D87, M125, E111
Cluster Sample	10	M105, D79, H79, E96
	2	M125, A47, D101, D101
	5	E107, H77, D75, M123
	9	M130, A45, D80, M105
	3	M98, M100, D93, D72
Stratified Sample	M	M98, M100, M119, M109
	E	E100, E110, E112, E122
	A	A46, A46, A49, A42, A47, A58
	D	D78, D73, D76, D88
	H	H72, H70

Your group will next take a cluster sample of 20 of the 200 individuals at the company. To do this, you will need one person to roll the dice and a second to write down the data. The clusters will be taken to be the columns of Table 8.5. The process detailed below will take a simple random sample of five clusters, and then take a simple random sample of four individuals from each cluster. To take a cluster sample, use the following process:

1. Roll the ten-sided die.
2. Locate the column corresponding to the outcome of the roll of the ten-sided die. This is the sampled cluster. If this cluster has been previously sampled, return to Step 1. Write the column number in the *Cluster* column of Table 8.6.
3. Roll the twenty-sided die and locate the roll corresponding to this outcome in the column corresponding to the sampled cluster. This is the individual to be included in the sample.
4. Check this selection against the individuals that are already in the sample. If this individual has already been included in the sample, re-roll the dice and choose again. If this individual has not been selected yet, proceed to the next step.
5. Write down the data for the selected individual on a table like the one shown in Table 8.6 in the row corresponding to the current cluster.
6. If you have selected fewer than four individuals from this cluster, proceed to Step 3. If you have selected four individuals from this cluster, and fewer than five clusters have been sampled, proceed to Step 1. If you have selected four individuals from this cluster and five clusters have been sampled, the process is complete.

For example, suppose that we roll the ten-sided die and the result is 10. This indicates that the tenth column is the first cluster used in our cluster sample. This is indicated on the example data sheet in Table 8.7. We now roll the twenty-sided die four times. Suppose that these rolls are 3, 3, 16, and 19, which correspond to the data points M105, D79, H79, and E96, as indicated in Table 8.5. Notice that the roll equal to 3 is repeated, so the second one is skipped and we need another roll. Suppose that this roll is equal to 7. We now select the data values corresponding to these rows in the tenth column of the data table. These values are recorded in the example data sheet in Table 8.7. Now that we have five observations for that cluster, we need to sample another cluster. Suppose that we roll the ten-sided die and observe a 2. This indicates that the second column is the second cluster used in our cluster sample. This is indicated on the example data sheet in Table 8.7. We now roll the twenty-sided die four times. Suppose that these rolls are 4, 10, 15, and 17, which correspond to the data points M125, A47, D101, and D101, as indicated in Table 8. Note that even though the data D101 is repeated, the data come from two different individuals.

Finally, your group will take a stratified sample. The strata will correspond to the job types. Stratified sampling takes some additional work as the sizes of the strata are different. For example, there are 40 management employees while there are only 20 human resource employees. This means that we need to sample twice as many management employees as human resource employees. Because we need a sample of size 20, the sample sizes for the strata will correspond to the number of rows in the table for each job type. Therefore, your group should sample 4 management employees, 4 engineers, 6 assembly line workers, 4 data analysts, and 2 human resource employees. To take the sample of management employees, use the following process:

1. Roll the ten-sided die twice. The first roll will select the row that the individual will be sampled from, and the second roll will select the column. If the first roll is greater than four, continue rolling until you get a number between 1 and 4 because there are only four rows of management employees.
2. Once the row and column have been selected, make sure that the individual has not already been selected for the sample. If they have, return to Step 1. If they have not, write the data for that individual in the *M* row of the *Stratified Sample* part of the data table.
3. If you have selected four individuals, you have completed the sample for the management employees and you will next sample from the engineering employees. If you have not yet selected four individuals, return to Step 1.

This process is then repeated for each of the remaining strata, yielding a stratified sample of size 20. Remember that the sample sizes differ depending on which strata you are sampling from: 4 management employees, 4 engineers, 6 assembly line workers, 4 data analysts, and 2 human resource employees. Once the process is complete, the data table assembled by your group should appear similar to the one shown in Table 8.7.

Take the management stratum, for example. Suppose that we roll a ten-sided die and observe a 1. This indicates that we will sample an employee from the first row of management employees. The ten-sided die is rolled again and this time we observe 7. The we pick the management employee from the first row and the seventh column which is the data value M98, as indicated in Table 8.7. Next, we roll 9, but because there are only four rows of management employees, we roll again and observe a 3. Therefore, we will sample an employee from the third row of management employees. Rolling the die again, suppose we observe 8, then we pick the data from the third row of management employees in the 8th column, which is the data point M100, as indicated in Table 8.7. We continue this process until we have sampled four individuals from the management employees, as indicated by the example data in Table 8.7.

Calculations

1. We will first calculate the percentage of individuals in each sample from each job type. Take the number of individuals for a job type and then divide by the number of individuals in the sample, and then multiply by 100%. For example, using the data in Table 8.7, nine managers were selected for the simple random sample. Because the total number of individuals in the simple random sample is 20, the total percentage of individuals in the sample that are managers is $9 \div 20 \times 100$. These percentages can be reported on a table like that of Table 8.8. The percentages for the example data can be found in Table 8.9.

Table 8.8 Percentage of each job type in the observed samples .

Sample Type	Job Type	Percentage	Average
Simple Random	M		
Sample	E		
	A		
	D		
	H		
Cluster Sample	M		
	E		
	A		
	D		
	H		
Stratified	M		
Sample	E		
	A		
	D		
	H		

Table 8.9 Percentage of each job type in the observed samples for the example data.

Sample Type	Job Type	Percentage	Average
Simple Random	M	45	122.44
Sample	E	25	105.40
	A	15	46.00
	D	10	94.00
	H	5	64.00
Cluster Sample	M	35	112.29
	E	10	101.50
	A	10	46.00
	D	35	85.86
	H	10	78.00
Stratified	M	20	106.50
Sample	E	20	111.00
	A	30	48.00
	D	20	78.75
	H	10	71.00

2. We will next calculate the sample average salary for each job type and for each sample type. Beginning with the management employees in the simple random sample, first identify the management employees in the simple random sample and consider the salaries for those employees. For the example data, these salaries are: 117, 117, 99, 106, 102, 114, 213, 109, 125. To find the average of these values, add them up and divide by the number of data values:

$$(117 + 117 + 99 + 106 + 102 + 114 + 213 + 109 + 125) \div 9 = 1102 \div 9 = 122.44.$$

Next, move to the engineering employees with the same process. Identify all the engineers in the sample and consider their salaries. Add these together and divide by the number of data values. For the example data, these values are 109, 94, 117, 96, and 111. To find the average of these values, add them up and divide by the number of data values:

$$(109 + 94 + 117 + 96 + 111) \div 5 = 527 \div 5 = 105.4.$$

This process will then continue for the remaining job types in the simple random sample. Then apply the same process for the cluster sample and the stratified sample. The remaining averages for the example data are given on Table 8.9.

Questions

1. Now that you have taken each type of sample, think about comparing them. Suppose that someone asked you why a cluster sample is not the same as a simple random sample. How would you explain the idea that they are different?
2. Suppose that someone asked you why a stratified sample is not the same as a simple random sample. How would you explain the idea that they are different?
3. Now consider the percentage of each job type represented in your three samples, and compare them to the percentage of each job type in the population. Which of your three samples seemed to do the best job of representing the population in terms of these percentages?

4. Finally, consider the average salary for each job type represented in your three samples, and compare them to the average salary for each job type in the population. For the population, the average salary is 111.575 for managers, 103.450 for engineers, 47.416 for assembly line workers, 82.525 for data analysts, and 74.050 for human resource employees. Which of your three samples seemed to do the best job of representing the population in terms of these averages?

This page titled 8.11: Collaborative Activity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

8.12: References

Abrams, Karen, and Fanying Kong. 2012. “The Variables Most Closely Associated with Academic Achievement: A Review of the Research Literature.” Progressive Research Institute of Nebraska. https://static1.squarespace.com/stat...+Sept+2012.pdf

Bowley, Arthur Lyon. 1906. “Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science, York, 1906.” *Journal of the Royal Statistical Society* 69, no. 3, 540–558. https://doi.org/10.1111/j.2397-2335.1906.tb02334.x

Condron, Dennis J., Daniel Tope, Christina R. Steidl, and Kendralin J. Freeman. 2013. “Racial Segregation and the Black/White Achievement Gap, 1992 to 2009.” *The Sociological Quarterly* 54, no. 1, 130–157. https://doi.org/10.1111/tsq.12010

Gershenfeld, Susan, Denice Ward Hood, and Min Zhan. 2016. “The Role of First-Semester GPA in Predicting Graduation Rates of Underrepresented Students.” *Journal of College Student Retention: Research, Theory & Practice* 17, no. 4, 469–488. https://doi.org/10.1177/1521025115579251

Glasser, Allison M., Nathaniel Onnen, Peter F. Craigmile, Elli Schwartz, and Megan E. Roberts. 2022. “Associations Between Disparities in Tobacco Retailer Density and Disparities in Tobacco Use.” *Preventive Medicine* 154, 106910. https://doi.org/10.1016/j.ypmed.2021.106910

Greer, Tawanda M., Adrian Laseter, and David Asiamah. 2009. “Gender as a Moderator of the Relation Between Race-Related Stress and Mental Health Symptoms for African Americans.” *Psychology of Women Quarterly* 33, no. 3, 295–307. https://doi.org/10.1111/j.1471-6402.2009.01502.x

Henry, Daphne A., Laura Betancur Cortés, and Elizabeth Votruba-Drzal. 2020. “Black–White Achievement Gaps Differ by Family Socioeconomic Status from Early Childhood Through Early Adolescence.” *Journal of Educational Psychology* 112, no. 8, 1471–1489. https://doi.org/10.1037/edu0000439

Ingels, S. J., D. J. Pratt, J. E. Rogers, P. H. Siegel, and E. S. Stutts. 2004. *Education Longitudinal Study of 2002: Base Year Data File User's Manual*. NCES 2004-405. National Center for Education Statistics. https://files.eric.ed.gov/fulltext/ED484410.pdf

Jay, Gina M., and Anthony R. D’Augelli. 1991. “Social Support and Adjustment to University Life: A Comparison of African-American and White Freshmen.” *Journal of Community Psychology* 19, no. 2, 95–108. https://doi.org/10.1002/1520-6629(199104)19:2<95::AID-JCOP2290190202>3.0.CO;2-2

Lorah, Julie, and Edwin Ndum. 2013. “Trends in Achievement Gaps in First-Year College Courses for Racial/Ethnic, Income, and Gender Subgroups: A 12-Year Study.” *ACT Research Report Series 8*. ACT Inc. https://www.act.org/content/dam/act/...T_RR2013-8.pdf

Mannan, Golam, Lillian Charleston, and Behrooz Saghafi. 1986. “A Comparison of the Academic Performance of Black and White Freshman Students on an Urban Commuter Campus.” *The Journal of Negro Education* 55, no. 2, 155–161. https://doi.org/10.2307/2294878

Miksic, Mai. 2014. *The Persistent Achievement Gaps in American Education*. CUNY Institute for Education Policy. https://www.researchgate.net/profile...-Education.pdf

Nettles, Michael T., A. Robert Thoeny, and Erica J. Gosman. 1986. “Comparative and Predictive Analyses of Black and White Students’ College Achievement and Experiences.” *The Journal of Higher Education* 57, no. 3, 289–318. https://doi.org/10.2307/1981554

Oguntunde, Pelumi E., Hilary I. Okagbue, Omoleye A. Oguntunde, Abiodun A. Opanuga, and Sola J. Oluwatunde. 2018. “Analysis of the Inter-Relationship Between Students’ First Year Results and Their Final Graduating Grades.” *International Journal of Advanced and Applied Sciences* 5, no. 10, 1–6. https://doi.org/10.21833/ijaas.2018.10.001

Olneck, Michael. 2005. “Economic Consequences of the Academic Achievement Gap for African Americans.” *Marquette Law Review* 89, 95. https://scholarship.law.marquette.ed.../vol89/iss1/8/

Paige, Rod, and Elaine Witty. 2010. *The Black-White Achievement Gap*. Amacom. https://www.google.com/books/edition...MsR3VtMC?hl=en

Sato T., T. Ito, T. Hirano, O. Morita, R. Kikuchi, N. Endo, and N. Tanabe. 2011. “Low Back Pain in Childhood and Adolescence: Assessment of Sports Activities.” *European Spine Journal* 20, 94–99. https://doi.org/10.1007/s00586-010-1485-8

Spenner, Kenneth I., Claudia Buchmann, and Lawrence R. Landerman. 2004. “The Black–White Achievement Gap in the First College Year: Evidence from a New Longitudinal Case Study.” *Research in Social Stratification and Mobility* 22, 187–216. https://doi.org/10.1016/S0276-5624(04)22007-8

Stephan, F. F. 1948. “History of the Uses of Modern Sampling Procedures.” *Journal of the American Statistical Association* 43, no. 241, 12–39. https://doi.org/10.1080/01621459.1948.10483247

Student. 1908. “The Probable Error of a Mean.” *Biometrika* 6, no. 1, 1–26. https://doi.org/10.2307/2331554

This page titled 8.12: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

CHAPTER OVERVIEW

9: Numerical Methods for Summarizing Data

“While the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to.”

Sir Arthur Conan Doyle from *The Sign of the Four*

9.1: What to Expect
9.2: Summarizing Data
9.3: Location
9.4: Variation
9.5: Percentiles
9.6: Critical Analysis
9.6.1: Race and COVID-19 Mortality
9.6.2: Effects of a New Special Formula Fertilizer on Potato Growth
9.6.3: Visual Exercise Environments and Treadmill Running
9.6.4: Satisfaction with Sport Participation
9.7: Summary
9.8: About the Quotation
9.9: Practice and Exploration
9.10: Collaborative Activity
9.11: References

This page titled 9: Numerical Methods for Summarizing Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.1: What to Expect

This chapter marks the beginning of considering what types of calculations researchers use to extract information from data after it has been observed. Summarizing the key characteristics of observed data is an important first step in any formal data analysis, and these types of summaries are very common in published studies. In this chapter, we will consider data that has been observed on an interval or ratio measurement scale. Summaries of data can be either numerical or visual. The focus of this chapter is on the numerical summaries for data and the conceptual interpretation of these summaries. There are some numerical calculations involved in this chapter, though most of the calculations are straightforward. By the end of this chapter, you will be able to:

1. Describe the conceptual ideas behind the location, variation, and percentiles.
2. Compute and interpret an average, the median, and the range.
3. Discuss the characteristics of data that the standard deviation measures.

This page titled 9.1: What to Expect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.2: Summarizing Data

When you get home in the evening and your partner, spouse, roommate, or friend asks you about your day, you usually provide a summary of what happened during the day. This summary gives an impression of the entire day without getting into a lot of the details, unless they are called for. Working with data is very similar. The data observed from a study is usually a large mass of numbers: several variables, hundreds or thousands of observations, all listed out or stored in computer memory somewhere. How can we look at this data in a way that will helps us find the information about research problems that interest us?

To better visualize the issue we face, let us consider a hypothetical study of first-year salaries for those graduating with a bachelor's degree at a particular university. The main issue being considered is whether students of color are thought to have lower first-year salaries after graduation than white students. To study this potential effect, 100 students of color and 100 white students are randomly sampled from the previous graduating class at the university. Each of the sampled students are contacted and the university can observe their first year starting salary. For the sake of simplicity, assume that all the contacted students were employed and reported their salary. The observed salaries, reported in thousands of dollars, are given in Tables 9.1 and 9.2.

Table 9.1 Observed first-year salaries for graduates of color from the university for the hypothetical study discussed in the text (in thousands of dollars).

45	52	50	42	57	49	57	61	52	55
59	53	51	53	58	33	59	64	46	59
53	55	60	59	62	48	48	49	37	49
58	43	53	45	47	42	54	45	38	46
52	53	53	38	53	66	48	37	53	40
47	45	47	54	50	58	41	43	51	59
48	42	43	44	42	48	39	52	45	49
55	61	43	40	45	52	33	45	61	46
37	54	57	57	52	53	43	38	59	53
42	51	57	44	45	45	56	45	47	46

Table 9.2 Observed first-year salaries for white graduates from the university for the hypothetical study discussed in the text (in thousands of dollars).

57	65	56	55	65	64	62	58	68	51
57	57	63	64	46	59	62	45	64	58
62	58	56	49	42	68	59	45	73	54
52	64	67	45	51	72	52	60	62	56
70	49	53	51	59	76	60	63	50	52
59	64	68	64	63	67	69	61	47	57
67	75	64	78	51	58	62	59	55	68
82	62	58	58	49	64	64	60	69	56
56	66	61	45	66	69	56	58	58	65
51	69	55	64	59	63	71	63	64	61

Looking at the tables, what can we learn about the first-year salaries? Briefly, we can observe that in Table 9.1 there seem to be a lot of salaries between $30,000 and $60,000, while in Table 9.2 there seem to be a lot of salaries between $50,000 and $70,000. These observations certainly give one the feeling that in some way the salaries reported in Table 9.1 are generally lower than the salaries reported in Table 9.2, but conceptually the conclusion is somewhat abstract. What would be helpful is a way to summarize the salaries in these tables to make such a comparison easier.

A summary, such as those given at the end of each of the chapters, provides a brief review of the essential points in the information presented. A summary of a table of data will briefly review the essential characteristics of the data. One such summary you are probably familiar with is the computation of an average grade. For Table 9.1 the average income is 49.5, while for Table 9.2 the average income is 60.0. Most people then are very comfortable with the interpreting the result on some intuitive level. The average income of the graduates of color is 49.5 while the average income for white graduates is 60.0, so that the graduates of color, on average, make a little more than $10,000 less than the white graduates in the first year after graduation.

But what have we really done here? What have we really compared? What does an average tell us—what does it summarize about the data in each table? These are important questions. In each case we have taken 100 observed data values and summarized them by a single number. Obviously, much of the detail contained in the data is lost during this calculation and so it is important for us to know what characteristic of the data is being summarized, and what details are left out.

This page titled 9.2: Summarizing Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.3: Location

While there are many characteristics of data that can be considered, most studies concentrate on a few key characteristics that are easy for both experts and novices to interpret. One of the primary concerns with summarizing a set of data is to get some idea as to how big a typical value in the data may be. This is particularly relevant when comparing two or more sets of data. Earlier we considered two lists of salaries and wished to determine if the individuals in one of the groups tended to earn more than the individuals in the other group. Of course, some data points in a set of data may overlap with the other. That is, individuals in one group may make more than the other group and vice-versa, but we are interested in an overall summary of the size of the data in each table. Statisticians and data scientists answer these types of questions by summarizing the location or central tendency of the data.

The idea that a set of data has a location means that there is some mathematical structure to the data. At the very least we should be able to determine if one datum is greater than, equal to, or less than another datum. In this section we will insist on a little bit more structure as it is most common to measure the location of data that has been measured on an interval or ratio measurement scale. While there are some measures of location for ordinal data, they are much less common in practice.

Definition: Location

A measure of *location* of a set of data is a measure that summarizes the set of data with a single value that represents the middle or center of the data.

This is an abstract concept, and there is not one unique method for finding the location of a set of data. Rather, there are several methods that attempt to measure this concept, all in slightly different ways. We will consider the two most popular methods for measuring location that are used in most simple statistical analyses.

The first method for summarizing the location of a set of data is one that most individuals are familiar with: an average of the numbers. In the field of statistics and data science, the average is usually called the mean.

Definition: Mean

To compute the *mean* of a set of data, add all the values in the data set and divide by the number of values in the data set.

Example 9.3.1

Suppose that we have five incomes (in thousands of dollars): 23, 54, 44, 77, and 26. Find the mean.

Solution

To find the mean income we first add all the data values to get

$$23 + 54 + 44 + 77 + 36 = 234. \tag{9.3.1}$$

Once the values have been added, we divide by the number of observations. In this case there are five, so the mean is equal to

$$234 \div 5 = 46.8. \tag{9.3.2}$$

The relevant question is how does the mean measure the location of a set of data? The idea behind the mean is based on the concept of center of mass used in physics. For the data given in the example, visualize a balance beam that has been marked like a ruler at every whole number. For this example, we will only include the part of this ruler needed to show the data. Now think about placing a small weight on the ruler at each point where there is a data point in the data set. For this to work, all the weights would need to have the same mass. Now the question is, if the ruler and the weights were placed on a point, where could we locate the point so that the ruler and weights would be perfectly balanced? As demonstrated in Figure 9.3.1, the answer to this turns out to be the mean!

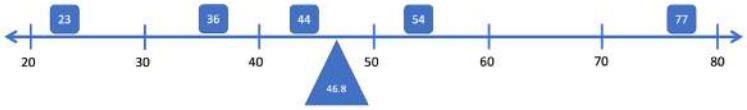


Figure 9.3.1: An interpretation of the mean as the center of mass for a set of data. A weight is placed at each data point on a hypothetical number line. The mean is where the line would balance (Public domain image created by Alan M. Polansky).

In Figure 9.3.1 we can visually observe how this interpretation works. Note that the balance point is slightly to the right of the third largest value in the data set to make up for the fact that the largest datum is farther out to the right than the bulk of the data points. This interpretation of the mean can tell us a lot about its properties and can be useful when considering whether the mean is a good measure of location for a specific set of data when evaluating a research study. For example, it should be clear that if we take the largest data point in the previous example and add to it, essentially moving it farther right on the number line in Figure 9.3.1, then the mean will also have to move to the right for the line to keep balance.

Example 9.3.2

Suppose that we have five incomes (in thousands of dollars): 23, 54, 44, 97, and 26. This is the same data as was used in the previous example with 77 replaced by 97. Find the mean.

Solution

To find the mean income, we first add all the data values to get

$$23 + 54 + 44 + 97 + 36 = 254. \tag{9.3.3}$$

Once the values have been added, we divide by the number of observations. There are five observations, so the mean is equal to

$$224 \div 5 = 50.8. \tag{9.3.4}$$

We can observe that the mean increased from 46.8 to 50.8 to keep balance when the datum 77 was moved to the right, now 97 (Figure 9.3.2). We can now imagine moving the largest data point even farther to the right, for example to 135.

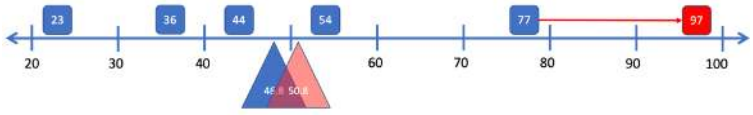


Figure 9.3.2: This is the same data as in Figure 9.1 with 77 being replaced by 97. The corresponding effect on the value of the mean is shown by the shift in the red balance point. (Public domain image created by Alan M. Polansky)

Example 9.3.3

Suppose that we have five incomes (in thousands of dollars): 23, 54, 44, 150, and 26. This is the same set of data used in the previous examples with 77 replaced by 135. Find the mean.

Solution

To find the mean income we first add all the data values to get

$$23 + 54 + 44 + 135 + 36 = 292. \tag{9.3.5}$$

Once the values have been added, we divide by the number of observations. In this case, there are five observations, so the mean is equal to

$$292 \div 5 = 58.4. \tag{9.3.6}$$

From the previous example and Figure 9.3.3 we can observe that when the largest value is 135 the mean is greater than four of the five data points, or 80% of the data. In this case, many statisticians and data scientists question whether the mean is really providing a good measure of the location of the data. Indeed, the problem gets worse the farther the rightmost point is shifted. For

example, if 77 is replaced by 1,000, the mean would be equal to 231.4, which is more than six times the second largest value in the set of data. Statisticians say that the mean is not robust to outliers because of this problem.

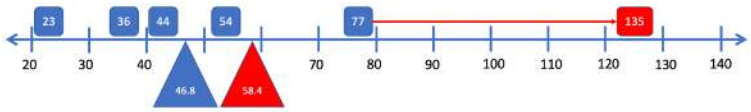


Figure **9.3.3**: This is the same data as in the previous example with 77 being replaced by 135. The corresponding effect on the value of the mean is shown by the shift in the red balance point. (Public domain image created by Alan M. Polansky)

Is the behavior of the mean a problem in practical applications? Consider the data given in Table 9.3, which corresponds to 100 observations of salaries from a large employer that have been simulated for the purpose of the example. One can observe from the table that the salaries have quite a range, starting at around $20,000 and going all the way up to $919,000. One can also observe, and we will see this better in subsequent chapters, that many salaries are relatively low, around $20,000 to $40,000, and only a few that are higher. This type of behavior is quite typical of salary data at any company. There are usually many relatively low salaries with a few very large salaries.

Table 9.3 Simulated salaries of 100 employees in thousands of dollars.

20	44	23	23	27	30	24	110	28	141
51	23	24	31	29	26	21	325	97	42
21	33	24	19	21	27	31	24	58	21
20	36	55	34	30	27	58	47	24	39
28	42	162	29	46	21	99	89	44	40
32	41	69	29	25	31	43	23	37	44
27	34	40	24	32	29	19	919	41	44
22	25	35	27	26	29	34	27	35	27
20	78	26	36	22	24	34	46	20	70
21	26	24	29	31	34	114	39	26	29

If the mean is computed on the salary data, we get that the mean salary for the company is $50,070. That is, the president of the company can boast that the average worker at their company earns about $50,000. While there are certainly employees at the company who make that salary, and more, a close look at Table 9.3 reveals that 84% of the employees attending the company make less than that salary. From this viewpoint the mean seems to imply that the salaries at the company are much better than they seem to most of the employees.

The mean is generally not considered a good measure of the location of data when there are values that are far away from the bulk of the data. These data values, called outliers, can either be very large (to the right of most of the data) or can be very small (to the left of most of the data), and the mean will be pulled in the direction of the outlying value. In these cases, a measure of location that divides the data in two halves, an upper half and a lower half, is often considered a better measure of location. This measure is called a median.

Definition: Median

A *median* is any point which divides the lower half from the upper half of the data.

Before proceeding to some examples there are some things that should be pointed out about a median. If there are an odd number of data values, then the median will be one of the data points. If there are an even number of data values, then the median is any value between the two halves of the data. Traditionally, for the case of an even number of data points, the average of the two data

points nearest the middle of the data is the median. One should be careful in that not all researchers or statistical computing software will use this convention, but we will use it in this book.

Example 9.3.4

Suppose that we have five incomes (in thousands of dollars): 23, 54, 44, 77, and 26. Find the median.

Solution

To find the median income we first need to sort the data values from smallest to largest: 23, 26, 44, 54, and 77. Because there is an odd number of data points, the median will be the data point that divides the upper half of the data from the lower half. This point is 44 since there are exactly two data points less than 44 and exactly two data points above 44.

Example 9.3.5

Suppose that we have six incomes (in thousands of dollars): 23, 54, 44, 77, 34, and 26. Find the median.

Solution

To find the median income we first need to sort the data values from smallest to largest: 23, 26, 34, 44, 54, and 76. Because there are an even number of data points, the median will be any point that divides the upper half of the data from the lower half, that is, any point greater than 34 but less than 44. Following the usual convention, the two values closest to the middle are averaged together so that the median is

$$(34 + 44) \div 2 = 39. \tag{9.3.7}$$

In the previous examples the interpretation of the median is very simple: the median is the point that divides the upper half of the data from the lower half. That is, half of the salaries are above the median and half are below.

Note that the median does not have the same problem that is shared by the mean when considering outliers in the data. In the previous two examples the largest value in the data set 23, 54, 44, 77, and 26 was first changed from 77 to 97, and then from 77 to 135. In each case the mean became larger to balance the increasing value of the largest data point. The median does not change when only the largest value is increased.

Example 9.3.6

Suppose that we have five incomes (in thousands of dollars): 23, 54, 44, 97, and 26. Find the median.

Solution

To find the median income we first need to sort the data values from smallest to largest: 23, 26, 44, 54, and 97. Because there is an odd number of data points, the median will be the data point that divides the upper half of the data from the lower half. This point is 44 since there are exactly two data points less than 44 and exactly two data points above 44. Note that the value of the median for this example matches the median for the same set of data where 77 was used in place of 97.

Because the median is more resistant to outliers, it is often preferred to the mean when the data may have very large, or very small, data values compared to the bulk of the data. Salaries and income data are two instances where the median is often preferred over the mean because of the potential for such outliers. For the simulated salary data given in Table 9.3, the median is $30,000. This means that half of the employees earn more than $30,000 and half of the employees earn less than $30,000. For this data, this value is a more realistic view of the typical salary at the company.

If the mean has problems with outliers, why not always use the median instead of the mean? Indeed, the mean is the most common measure of location, and the reason for this has to do with some theoretical properties of the mean that are beyond the scope of this book. What you need to know is that if a set of data is generally well behaved, the mean is a much better choice and has many theoretical properties that support its preference. In many of these cases the two measures will be close to one another. If there are outliers in the data, and under some additional conditions that will be discussed later, the median may be used in its place. You may find that some studies report both measures.

Are there other ways to measure the location of a set of data? Statisticians have developed many other methods for measuring the location of a set of data, and some of these have very impressive theoretical properties. The usual problem with these methods is that they can be complicated to compute and knowing when to use them is not always clear. For these reasons, the use of alternative measures is quite rare and familiarity with these methods is usually not required. Indeed, none of the research in any of the studies we use in this book rely on any measure of location other than the mean or median.

This page titled 9.3: Location is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.4: Variation

The second most important characteristic of a set of data is how far away from the location the data points typically are found. This characteristic is called variation. To demonstrate what differences in variation look like given different data sets, consider the data found in Tables 9.4 and 9.5. The data in these tables correspond to simulated data that represents the waiting times for emergency services at two hospitals. The data in Tables 9.4 and 9.5 correspond to the waiting times of 25 randomly sampled cases at a suburban and urban hospital, respectively. A comparison of the locations of the data in these tables show that the mean waiting time at the suburban hospital is about 100, which matches that of the urban hospital. Similarly, the median waiting time at the suburban hospital is about 101 and the median of the urban hospital is 103. Hence, both sets of data have very similar locations. However, a close look at the data listed in the two tables reveals some fundamental differences.

Table 9.4 Twenty-five simulated waiting times (in minutes) in the emergency room of a suburban hospital emergency room.

91	116	98	104	81
103	114	105	99	105
91	84	81	79	105
103	103	125	98	104
94	101	101	101	110

Table 9.5 Twenty-five simulated waiting times (in minutes) in the emergency room of an urban hospital emergency room.

134	122	125	114	102
39	72	97	110	103
119	109	74	124	124
109	86	68	102	75
104	93	126	86	80

In Table 9.4 the lowest data point is 79, while the highest data point is 125, and in Table 9.5 the lowest data point is 39, while the highest data point is 134. But it is not with just these extreme data points that we see a difference. If we look at the data in Table 9.5, there seems to be a wider “variety” in the data. The difference between the data in the two tables is that the data in Table 9.5 have more variation than the data in Table 9.4.

Definition: Variation

A measure of *variation* of a set of data is a measure that summarizes the set of data with a single value that represents typically how far data points can be found from the location of the data.

We will discuss two methods for measuring the variation in a set of data. The first one is simply computed by taking the largest datum in the dataset and subtracting from it the smallest datum. This measure is called the range.

Definition: Range

To compute the *range* of a set of data, subtract the smallest value in the set of data from the largest value in the set of data.

Example 9.4.1

In a previous example we were given five incomes (in thousands of dollars): 23, 54, 44, 77, and 26. Find the range.

Solution

To find the range of the incomes we first note that the largest value is 77 and the smallest value is 23. The range is then computed as $77 - 23 = 54$.

Example 9.4.2

In a follow-up example, the largest value is replaced by 97 so that the data are 23, 54, 44, 97, and 26. Find the range.

Solution

To find the range for this data we first note that the largest value is now 97, and the smallest value is still 23. The range is then computed as $97 - 23 = 74$.

Example 9.4.3

As another example, the largest value is replaced by 135 so that the data is 23, 54, 44, 135, and 26. Find the range.

Solution

To find the range for this data we first note the largest value is 135 and the smallest value in the data is still 23. The range is then computed as $135 - 23 = 112$.

In each of the calculations in the previous example the range increases as the largest value increases. This is not a surprise as the range is computed based on this value. Let us consider once again the hypothetical study of first-year salaries for those graduating with a bachelor's degree at a particular university with the data given in Tables 9.1 and 9.2. For the data in Table 9.1, the largest value is 66 while the smallest value is 33, and therefore the range is computed as $66 - 33 = 33$. For the data in Table 9.2, the largest value is 82 while the smallest value is 42, and the range is computed as $82 - 42 = 40$. Therefore we can conclude that there is slightly more variation in the data in Table 9.1 than in Table 9.2. For the simulated waiting room data in Table 9.5, the largest value is 134 while the smallest value is 39 so that the range is $134 - 39 = 95$, while for the data in Table 9.4, the largest value is 125 while the smallest value is 79 so that the range is $125 - 79 = 46$. The calculations verify the general trend observed earlier in that the waiting times for the urban hospital emergency room have much more variation than the emergency room waiting times for the suburban hospital emergency room.

In the previous example calculations, the range appears to do a decent job of telling us about variation in the data, but it is not used very often as a measure of variation in research studies. There are many reasons for this that have to do with statistical theory, but from an intuitive viewpoint it is easy to see that the range only uses information from the data corresponding to two data points: the maximum and minimum data points. In the definition of variation, we get a sense that variation is about how far all the values in the data are typically found from the location. We will consider in some detail a measure of variation that is used in most statistical applications: the standard deviation.

To develop the idea of the standard deviation, we will consider the simple set of data from a previous example consisting of the values 23, 54, 44, 77, and 26. Variation describes how far the data points are from a measure of location, and to develop the standard deviation we will use the mean as the measure of location. It is easy to compute how far away each data value is from the mean. These distances, as we will compute them here, are called the absolute deviations from the mean. Referring to Figure 9.4 will aid in this development. As before, the data are indicated on the number line using the boxes and the mean is indicated by the tip of the triangle below the number line. We begin with the first value, which is equal to 23. The difference between the mean, which is equal to 46.8, and this value will tell us how far the data is from the mean. In this case the value of the absolute deviation is $46.8 - 23.0 = 23.8$. This is indicated by the distance between the dashed green vertical line and the dash blue vertical line in Figure 9.4.1. The second value is equal to 36. The difference between the mean and this value is $46.8 - 36.0 = 10.8$. This is indicated by the distance between the dashed orange vertical line and the dash blue vertical line in Figure 9.4.1. Similarly, the absolute deviation for the next value is $46.8 - 44.0 = 2.8$, as indicated in grey in Figure 9.4.1.

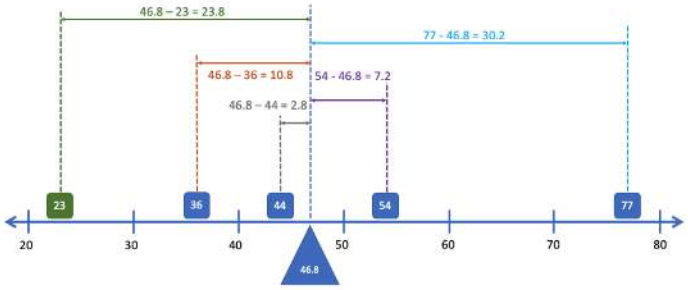


Figure 9.4.1: Computing the absolute deviations for the data from the example. The vertical dashed line indicates the location of the data as measured by the mean. The absolute deviations are indicated by the distances between each point and the mean (Public domain figure created by Alan M. Polansky).

The calculation of the absolute deviations changes slightly when we get to the next value, which is 54. Because distances are always greater than zero, to get the absolute deviation we will take the mean and subtract it from 54. That is, the absolute deviation is $54.0 - 46.8 = 7.2$, as indicated in purple in Figure 9.4.1. The last absolute deviation is $77.0 - 46.8 = 30.2$, indicated in light blue in Figure 9.4.1. The absolute deviations are listed in Table 9.6.

Table 9.6 The calculations used in computing the standard deviation of the of the data from the example. The data values are given in the first column, while for convenience the mean is listed in the second column. The squared absolute deviations are given in the fourth column.

Value	Mean	Absolute Deviation	Squared Absolute Deviation
23	46.8	23.8	566.44
36	46.8	10.8	116.64
44	46.8	2.8	7.84
54	46.8	7.2	51.84
77	46.8	30.2	912.04
Sum	1654.80		

Now that we have measured the distance between each point, what do we do with them? We would like to summarize the variation in the data with a single number, but at this point we have computed the distance between each data value and the mean, so we have as many of these absolute deviations as data values. One very tempting possibility is to compute the mean of the absolute deviations to get a summary of variation, called the mean absolute deviation. This measure seems to have a good intuitive foundation, but you will find that it is almost never used in practice. The reason for this is complicated, but it turns out that the mean absolute deviation has statistical properties that make it very inconvenient to work with (Casella and Berger 2024). So, what do we do? The standard deviation is based on computing something like the mean square distance between each value and the mean. For example, the first absolute deviation is equal to 23.8, and if we multiply that value by itself, that is we square it, we get $23.8 \times 23.8 = 23.82^2 = 566.44$.

Continuing, the next absolute deviation is equal to 10.8, and if we multiply that value by itself, that is we square it, we get $10.8 \times 10.8 = 10.82^2 = 116.64$. The remaining squared absolute deviations are computed in a similar manner and are listed in the third column of Table 9.6.

Now that we have computed all the squared absolute deviations we can add them together, indicated as the sum in Table 9.6 as 1654.80. We then take the sum of the squared absolute deviations and by divide by one less than the number of data points, to get $1654.80 \div 4 = 413.7$. Once again, it might be a little confusing as to why we would divide by one less than the number of data points. It has been shown mathematically that dividing by the sample size understates the amount of variation in the data by a very specific amount. It has also been shown mathematically that dividing by one less than the sample size provides a good measure of the variation in the data (Casella and Berger 2024). This value is called the variance of the data. However, we are not quite done as

there is still a problem with this measure in that it is in square units, instead of the same units that we started with. For example, for this data we are considering here the variance is measured in squared thousands of dollars, whereas the original data is in thousands of dollars. This means that we cannot compare the variation of the data to its mean value, which is measured in dollars. To take care of this, we take the square root of the variance to obtain $\sqrt{413.7} = 20.33$, to get the standard deviation.

Technically the standard deviation is the square root of the unbiased mean square deviation, but this complicated name does not help us interpret what is going on with the data too much. What is important is getting an approximate intuitive feeling for what the standard deviation is telling us about the data. The easiest interpretation is based on what is commonly known as the empirical rule or also as the three-sigma rule which states that for many sets of data, the standard deviation can be used to imply how much of the data is within a certain distance from the mean.

Theorem 9.4.1 [The Empirical Rule]

For most data:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- Approximately 99% of the data is within three standard deviations of the mean.

The first thing to note about the result is that the conclusions are approximate, and do not hold exactly. If the data has a strange structure, or if there aren't a lot of data points, then the percentage stated in the result will not be exact, and in some cases are not accurate at all. Regardless, the empirical rule provides a handy method for intuitively interpreting the standard deviation even if the result is not always exact.

We first need to understand what the result means starting with the first statement that approximately 68% of the data is within one standard deviation of the mean. Data points that are within one standard deviation of the mean include data points as small as the standard deviation subtracted from the mean, or

$$\text{mean} - \text{standard deviation}, \tag{9.4.1}$$

or as large as the standard deviation added to the mean, that is,

$$\text{mean} + \text{standard deviation}. \tag{9.4.2}$$

What this means is that the distance between any data point in this range and the mean is less than the standard deviation of the data. The empirical rule then states that about 68% of the data points will have a distance to the mean less than the standard deviation.

Example 9.4.4

The mean and the standard deviation of the data presented in Table 9.4 are 99.8 and 11.1, respectively. Use the Empirical Rule to find ranges that contain 68%, 95%, and 99% of the data.

Solution

In Table 9.7 we have listed each data value, along with the distance between the data value and the mean. Values that are within one standard deviation of the mean will have a value as small as

$$\text{mean} - \text{standard deviation} = 99.8 - 11.1 = 88.7, \tag{9.4.3}$$

and as large as

$$\text{mean} + \text{standard deviation} = 99.8 + 11.1 = 110.9. \tag{9.4.4}$$

Therefore, any datum with a value between 88.7 and 110.9 is within one standard deviation of the mean. For this data, the eighteen values 91, 98, 104, 103, 105, 99, 105, 91, 105, 103, 103, 98, 104, 94, 101, 101, 101, and 110 are within one standard deviation of the mean. This corresponds to about 72% of the data, a little more than is predicted by the empirical rule.

Table 9.7 The data from Table 9.4 along with the distance between each value and the mean (99.8), and whether the datum is within one, two, or three standard deviations (11.1) of the mean.

Datum	Distance	Standard Deviations
91	8.8	1
116	16.2	2
98	1.8	1
104	4.2	1
81	18.8	2
103	3.2	1
114	14.2	2
105	5.2	1
99	8.8	1
105	5.2	1
91	8.8	1
84	15.8	2
81	18.8	2
79	20.8	2
105	5.2	1
103	3.2	1
103	3.2	1
125	25.2	3
98	1.8	1
104	4.2	1
94	5.8	1
101	1.2	1
101	1.2	1
101	1.2	1
110	10.2	1

Now we can consider the meaning of the remaining results for the Empirical Rule. The second statement states that approximately 95% of the data is within two standard deviations of the mean. This means that the datum could be as small as two times the standard deviation subtracted from the mean, that is, the datum could be as small as

$$\text{mean} - 2 \times \text{standard deviation} = 99.8 - 2 \times 11.1 = 77.6, \tag{9.4.5}$$

or as large as two times the standard deviation added to the mean, that is, the datum could be as large as

$$\text{mean} + 2 \times \text{standard deviation} = 99.8 + 2 \times 11.1 = 122.0. \tag{9.4.6}$$

Therefore, any datum with a value between 77.6 and 122.0 is within two standard deviations of the mean. For this data there are 24 values within two standard deviations of the mean. This corresponds to about 96% of the data, only a little more than is predicted by the empirical rule.

Similarly, the third statement that approximately 99% of the data is within three standard deviations of the mean indicates that the datum could be as small as three times the standard deviation subtracted from the mean, that is, the datum could be as small as

$$\text{mean} - 3 \times \text{standard deviation} = 99.8 - 3 \times 11.1 = 66.5, \tag{9.4.7}$$

or as large as three times the standard deviation added to the mean, that is, the datum could be as large as

$$\text{mean} + 3 \times \text{standard deviation} = 99.8 + 3 \times 11.1 = 133.1. \tag{9.4.8}$$

Therefore, any datum with a value between 66.5 and 133.1 is within three standard deviations of the mean. All the data is within three standard deviations of the mean, again very close to what is predicted by the empirical rule.

What all this means is that when you observe the mean and the standard deviation for a set of data, you now have information about how much data is clumped within certain ranges of the mean. From intuition, this interpretation can be taken to be an informal rule to give you an idea about what can be expected in the data. For example, suppose that you have taken an exam and the instructor reports that the mean grade on the exam was 72 with a standard deviation equal to 5. You look at your score and find that you have scored a 97. This is a good score, but how does it compare with your classmates? From the empirical rule you know that roughly two thirds of the class (68%) scored within one standard deviation of the mean. That is, roughly two thirds of the class scored between $72.0 - 5.0 = 67.0$ and $72.0 + 5.0 = 77.0$. This tells you that you scored better than anyone in that range, which is two thirds of the class. That is pretty good! But the empirical rule tells you even more, that roughly 95% scored within two standard deviations of the mean. That is, roughly 95% of the class scored between $72.0 - 10.0 = 62.0$ and $72.0 + 10 = 82.0$. This tells you that you scored better than anyone in that range, which is roughly 95% of the class. That is very good! But the empirical rule is not done yet. It tells you even more in that roughly 99% scored within three standard deviations of the mean. That is, roughly 99% of the class scored between $72.0 - 15 = 57.0$ and $72.0 + 15.0 = 87.0$. This tells you that you scored better than anyone in that range, which is roughly 99% of the class. That is excellent!

We will conclude our discussion by assessing the location and variation of the data considered in this section. Recall that Tables 9.1 and 9.2 contain the first-year salaries for people of color and white graduates from a university, respectively. Table 9.8 shows the mean, median, range, and standard deviation for the data in these tables. From the information in the table, we can observe that the typical person of color will have a salary about 10,000 less than the white alumni. This is consistently indicated by both the mean and the median. Both measures of variation indicate that the white salary data have slightly more variation.

Table 9.8 Summary measures of location and variation for the first-year salaries of graduates from a university.

Alumni Group	Mean	Median	Range	Standard Deviation
People of Color	49.5	49.0	33.0	7.2
White	60.0	60.0	40.0	7.7

For the simulated salary data given in Table 9.3 we had previously computed the mean and the median to be 50.1 and 30.0 (in thousands of dollars), respectively. The range and standard deviation of the data are 900 and 95.7, respectively. For this data we can assess how unusual the top salaries for this company really are. According to the empirical rule, virtually all the data (99%) should be within three standard deviations of the mean. Hence, 99% of the data should be between $50.1 - 3 \times 95.7 = -237.0$ and $50.1 + 3 \times 95.7 = 337.2$.

We can already see a problem with the lower number as it is negative, and we know that salaries cannot be negative. So, we will cut this value off at zero, and we then conclude that all the salaries should essentially be between 0 and 337. We can see from Table 9.3 that there is one salaries above this value: 919. The salary of 919 is nearly three times the upper limit set by the empirical rule. This is an indication of just how usual this value is—it is well outside what would usually be expected. Such a conclusion may not

be too surprising as we had commented before that this very large value was dragging the mean toward it, and why we suggested using the median instead in this case.

Finally, let us consider the waiting room data contained in Tables 9.4 and 9.5. The summary location and variation measures are presented in Table 9.9. We have already seen that both the mean and medians indicate that the typical waiting for both data sets is around 100 minutes. The ranges and the standard deviations however indicate that there is considerably more variation in the waiting times for the urban hospital. So, for example, a waiting time of 125 minutes at the urban hospital might not be too unusual since that waiting time is only a little more than one standard deviation above the mean, whereas the same waiting time at the suburban hospital would be unusual as 125 is more than two standard deviations above the mean.

Table 9.9 Summary measures of location and variation the waiting times in the emergency rooms and a suburban and an urban hospital.

Setting	Mean	Median	Range	Standard Deviation
Suburban	99.9	101.0	46.0	11.1
Urban	99.9	103.0	95.0	22.9

This page titled 9.4: Variation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.5: Percentiles

If you have ever taken a standardized exam, you will be familiar with the concept of a percentile. Standardized exams are designed so that your grade is computed relative to a population of students who have taken the same exam. That is, your grade on a standardized exam reflects how well you did relative to the other people who took the exam. Hence, such an exam is not a reflection of your overall knowledge of a subject but is really a reflection of your knowledge of a subject relative to a specified population. Usually, these exams include two indicators of how well you did. The first is typically some type of raw score which can be difficult to interpret without specific knowledge about how the exam is graded and scored. The second indicator usually reports your percentile, which is what percentage of the population got a score equal to or below your own. Therefore, if you scored at the 80th percentile, then 80% of the population got your score or below on the exam. Knowing your percentile score lets you know how well you did relative to the population of other students who took the exam. For those who look at your scores, for example a prospective university, the percentile is interpreted as a measure of exceptionalism.

A percentile is a number that identifies the point at which a specified percentage of a set of data is less than or equal to that number.

Definition: Percentile

The pth *percentile* of a set of data is the smallest point for which at least p% of the data is less than or equal to that point.

The idea of a percentile is a generalization of the concept of the median. The median is a point that divides the data in half, so that 50% of the data should be less than or equal to the median, and hence the median is the 50th percentile of a set of data. It should be noted, however, that this method for computing the median is slightly different than what is outlined in above due to computing conventions, so we will make a distinction between the median and the 50th percentile.

Example 9.5.1

Consider the following ten observations from the simulated salary data: 20, 44, 21, 23, 27, 30, 24, 110, 28, and 141. Find the 50th and the 25th percentiles.

Solution

To make our calculations easier we will sort the data from smallest to largest to get: 20, 21, 23, 24, 27, 28, 30, 44, 110, and 141. Suppose we would like to compute the median—then we would use the fact that there are an even number of observations and compute the median as the mean of the two middle observations. That is, the median is $(27 + 28) \div 2 = 27.5$. The 50th percentile, as defined above, is the smallest value in the data for which at least 50% is less than or equal to the value. Note that the value 27 has exactly 50% of the data less than or equal to that value, and that any value less than 27 will have a smaller percentage less than or equal to that value. Hence, the 50th percentile is 27. Suppose now we wish to find the 25th percentile. We need to find the smallest value for which at least 25% of the data is less than or equal to that value. Note that the value 21 is too small since only 20% of the data is less than or equal to that point. This is true for any value that is greater than 21 but less than 23. For the value 23, 30% of the data is less than or equal to that value. This is the smallest value such that at least 25% of the data is less than or equal to that value, and hence 23 is the 25th percentile.

While percentiles are conceptually easy to define, the computation of percentiles for sets of data with few observations can be problematic in that the result is difficult to interpret in terms of the data. In the previous example we found the 25th percentile to be the value 24. But this value also corresponds to the 27th percentile, and the 30th percentile. The fact that these percentiles are all equal can cause confusion and there are many alternative suggestions on how percentiles can be computed, though there is no absolute consensus on how the 25th percentile would be computed in this case. One of the most common suggestions is to use linear interpolation. For example, it seems like the 25th percentile should be halfway between 21 and 23, which would be 22. The good news is that when there are many data points, these problems quickly go away, and in practice you will only need to be able to interpret what the percentile means in terms of the data and not compute it on your own. Generally, technical problems with how a percentile is computed will rarely influence the overall interpretation of the results of a research study.

This page titled 9.5: Percentiles is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.6: Critical Analysis

Many research studies will provide measures of location and variability as well as percentiles to summarize certain aspects of the data that they have observed. Measures of location can be interpreted as representing the typical or center of the data, and measures of variation can be interpreted as the typical distance that data points in the data set will vary from the location. The relationship between the location and the variation is interpreted in terms of the empirical rule. Percentiles provide information about the rank of individual points, regarding the percentage of data below a specified point, within the set of data.

In summary, the key questions in critically analyzing measures of location, variation, and percentiles in a study are:

1. What do the measures of location indicate about the typical sizes of the values in the set of data?
2. What do the measures of variation indicate about the how far data values in the set of data may be from the measure of location?
3. What does the empirical rule indicate about how much data is concentrated within ranges around the mean?
4. What do percentiles indicate about how small or large data values are compared to the rest of the data values?

This page titled 9.6: Critical Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.6.1: Race and COVID-19 Mortality

The COVID-19 pandemic in the United States has had a disproportionate effect on African Americans with three times the rate of infection and six times the rate of death in communities where white residents comprise the majority of the population (Golestaneh et al. 2020). Indeed, while African Americans constitute 18% of the population, they represent 33% of the COVID hospitalizations (Garg et al. 2020). The mortality rates are also disproportionate. For example, in New York City, the African American mortality rate is more than twice that of the white population when these rates have been adjusted for variables like age (Golestaneh et al. 2020).

One of the most important questions when considering the differing rates of infection and mortality is to assess what the contributing factors are that cause these differences. This question is crucial because these contributing factors can tell researchers what type of interventions are required to help alleviate the problem. To that end, individuals who received outpatient care at least once at the Montefiore Health System between January 1, 2018, and January 1, 2020, were studied in terms of race, COVID-19 infection, and mortality. Other data such as demographics and the presence of preexisting conditions were also observed. In all, data from a total 505,992 individuals were observed in the study (Golestaneh et al. 2020).

We will focus on a few of the reported numerical characteristics of the data. The original research paper provides information about patients seen at least once in the ambulatory centers during a certain period, for example, the mean ages of the patients seen based on race. For white patients, the mean age is 53.2; for African American patients, the mean age is 40.6; for Hispanic patients, the mean age is 37.6; and for patients who do not fall into those categories the mean age is 34.5 years old. The standard deviation for the ages is around 24 years for all patients.

One can note the rather large standard deviation in comparison with the means for these results. If we typically think of ages of humans having a range between 0 and 100, a standard deviation is quite large, indicating that we could expect that there is a representation of ages throughout the entire range in the set of data. We can also observe that white patients tend to be older than the African American and Hispanic patients, and the individuals who do not fall into these categories.

This page titled 9.6.1: Race and COVID-19 Mortality is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.6.2: Effects of a New Special Formula Fertilizer on Potato Growth

Previously we considered a study that explored the potential of three fertilizers that were designed to enhance the yield and quality of potato crops in a region of China (Xu et al. 2025). These fertilizers were compared to using no fertilizer and to using a conventional fertilizer on the crops using many measures of plant growth and yield. Some of these measures are technical in nature, so we will consider only a few selected measures. All the measures are based on computing the mean of the measures over the observed plants grown using each of the treatments.

For most of the measures the means are reported through graphs in the original paper. For our purposes here we will report numerical averages that have been approximated from the graphs. Table 9.10 shows these values for the measures of plant growth and yield. The first measure we consider is the mean plant height at the final growing stage of the crop, called the starch accumulation stage. From Table 9.10 we can observe that the mean plant height for plants with no fertilizer was 55.5cm while plants treated with the common fertilizer had a mean height of 63.3cm. Therefore, the use of the common fertilizer appears to increase the mean plant height in this stage. The three new fertilizers produced mean plant heights of 68.9cm, 72.5cm, and 71.3cm, respectively. All of these plants are larger than what was observed for the common fertilizer, but there is not much difference between the three mean values. From these results it appears that any of the specially formulated fertilizers are effective in increasing the mean plant height over using the common fertilizer or no fertilizer.

Table 9.10 Mean values for the measures of growth and yield of potato plants using different fertilizers (Xu et al. 2025). *None* indicates that no fertilizer was used. *Common* indicates that a common fertilizer was used. *New 1*, *New 2*, and *New 3* indicate the three new formulations of fertilizer.

Measurement	None	Common	New 1	New 2	New 3
Plant Height (cm)	55.5	63.3	68.9	71.3	72.5
Stem Diameter (mm)	15.8	18.2	19.0	20.7	18.5
Dry Weight (g)	33.4	35.4	38.9	43.6	41.1
Yield (t per ha)	26.8	44.0	46.8	59.4	55.4

The second measure we consider is the mean stem diameter at the final growing stage, which, from Table 9.10, we see for plants with no fertilizer was 15.8mm, while plants treated with the common fertilizer was 18.2mm. Therefore, the use of the common fertilizer appears to increase the mean stem diameter. The three new fertilizers produced mean stem diameters of 19.0mm, 20.7mm, and 18.5mm, respectively. Once again, all of these plants are larger than what was observed for the common fertilizer, but there is not much difference between the three mean values. From these results it appears that any of the specially formulated fertilizers are effective in increasing the mean stem diameter over using the common fertilizer or no fertilizer, though it is notable that the second fertilizer once again produces the largest mean value.

The third measure is the dry weight of the plant. The weight measure corresponds to the part of the plant that is above ground. From Table 9.10 we can observe that the mean dry weight for plants with no fertilizer was 33.4g while plants treated with the common fertilizer had a mean dry weight of 35.4g. Therefore, the use of the common fertilizer appears to increase the mean dry weight, but not to a great extent. The three new fertilizers produced dry weights of 38.9g, 43.6g, and 41.1g, respectively. Once again, all of these are larger than what was observed for the common fertilizer, but there is not much difference between these three mean values. From these results it appears that any of the specially formulated fertilizers are effective in increasing the mean dry weight over using the common fertilizer or no fertilizer, though it is notable that the second fertilizer once again produces the largest mean value, as with the two previous measures.

The last measure we will consider here is the yield measured in tons of potatoes per hectare. Plants treated with no fertilizer had a yield of 26.8 tons per hectare while the plants treated with the common fertilizer had a yield of 44.0 tons per hectare, clearly indicating that the common fertilizer increases yield considerably over not using any fertilizer. The three new fertilizers had yields

of 46.8 tons per hectare, 59.4 tons per hectare, and 55.4 tons per hectare, respectively. Once again, the new fertilizers have improved yields over the common fertilizer, and the second new fertilizer shows the most improvement.

The second specialized fertilizer appears to be the best, as it outperforms all the other options for all the measures. As we shall see later in the book, one must be careful when making this type of conclusion, as many of the means are rather close. One key concept we considered earlier is the fact that the means computed here are estimating some conceptual idea about the average performance of each of the fertilizers. These estimates will have errors associated with them, and those errors must be accounted for when comparing means.

This page titled 9.6.2: Effects of a New Special Formula Fertilizer on Potato Growth is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.6.3: Visual Exercise Environments and Treadmill Running

Earlier we considered a study that assessed the effect that different visual and auditory environments have on athletic performance (Yeh et al. 2017). In particular, the study considered measures of athletic performance for individuals running on a treadmill in different visual environments. The visual environments included a video that represents a first-person perspective of moving through a park and a static image from the same park. These environments were compared to an environment selected by the participants, including visual and auditory entertainment. Each participant used the treadmill for a twenty-minute session under each of these environments. The speed of the treadmill was controlled by the participant in each case.

The two primary measurements taken on each participant were the distance ran and the mean heart rate during the run. The mean heart rate was computed for everyone by measuring the heart rate in one-minute intervals during the run. The mean value of these measurements was then computed. The mean distances and heart rates for each of the environments is reported in Table 9.11.

Table 9.11 Mean values for the measures of athletic performance and emotional state for the study of visual exercise environments and treadmill running (Yeh et al. 2017).

		Dynamic		Static		Self-Selected	
Measure	Sample Size	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Distance (m)	30	2891.6	631.4	2767.2	6626	3066.8	688.5
Heart Rate (bpm)	24	141	18	138	21	147	188
Pre Anxiety	30	0.48	0.68	0.52	0.74	0.40	0.52
Post Anxiety	30	0.17	0.26	0.22	0.34	0.26	0.31
Pre Happiness	30	1.81	0.86	1.78	0.88	1.38	0.85
Post Happiness	30	2.10	0.84	2.19	0.86	2.04	0.88

The mean distances for the dynamic, static, and self-selected are 2891.6m, 2767.2m, and 3066.8, respectively. Therefore, we can observe that the self-selected environment was associated with the farthest distance, followed by the dynamic environment, and then the static environment. The researchers also report the standard deviations for these measurements as indicated in Table 9.11. These standard deviations have a range between approximately 630m and 690m. Recalling the empirical rule, we can conclude that almost all the observed data is within three standard deviations of the means, which corresponds to approximately 1900m to 2100m. For example, the mean distance for the dynamic environment is 2891.6m and the standard deviation is 631.4m. This means that almost all the observed distances will be between

$$2891.6m - 3 \times 631.4m = 2891.6m - 1894.2m = 997.4m,$$

and

$$2891.6m + 3 \times \times 631.4m = 2891.6m + 1894.2m = 4785.8m.$$

Similarly, for the static environment the mean distance is 2767.2m and the standard deviation is 662.6m. This means that almost all the observed distances will be between

$$2767.2m - 3 \times 662.6m = 2767.2m - 1987.8m = 779.4m,$$

and

$$2767.2m + 3 \times 662.6m = 2767.2m + 1987.8m = 4755.0m.$$

Finally, for the self-selected environment the mean distance is 3066.8m and the standard deviation is 688.5m. This means that almost all the observed distances will be between

$$3066.8m + 3 \times 688.5m = 2767.2m - 2065.5m = 1001.3m,$$

and

$$3066.8m + 3 \times 688.5m = 2767.2m + 2065.5m = 5132.3m.$$

The researchers in the study additionally gathered data on the emotional state of the participants using a survey that was administered before and after each exercise session. We will consider two of these measures here. The first is a measure of anxiety, where a smaller number indicates less anxiety. The mean pre-run anxiety level for the dynamic environment was 0.48 compared to a mean of 0.17 for the post-run anxiety level. Similar results are observed for the remaining two environments. The mean pre-run anxiety level for the static environment was 0.52 compared to a mean of 0.22 for the post-run anxiety level and for the self-selected environment the mean pre-run anxiety level was 0.40 compared to a post-run level of 0.26. Therefore, the exercise session reduced anxiety to some extent using each of the environments. The anxiety levels between environments are very close and therefore we shall not try to interpret them.

The second is a measure of happiness, where a larger number indicates more happiness. The mean pre-run happiness level for the dynamic environment was 1.81 compared to a mean of 2.10 for the post-run happiness level. Similar results are observed for the remaining two environments. The mean pre-run happiness level for the static environment was 0.86 compared to a mean of 0.84 for the post-run happiness level. For the self-selected environment, the mean pre-run happiness level was 1.38 compared to a post-run level of 2.04. Therefore, the exercise session increased happiness for each of the environments. It is interesting to note that the happiness level for the static environment is much lower than for the two other environments, and that the exercise session does not appear to increase happiness to the same extent as it does for the other two environments. It is also interesting to note that the happiness levels for the dynamic and the self-selected environments are nearly the same for both the pre-run and post-run measurement times.

This page titled 9.6.3: Visual Exercise Environments and Treadmill Running is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.6.4: Satisfaction with Sport Participation

Earlier we considered a study that investigated how success, personality, emotional competence, and sport type (team or individual) are associated with sports satisfaction (Lepir and Lakić 2025). For the study athletes were surveyed through various organizations. The survey was designed to measure various personality traits as well as satisfaction with sport participation. The researchers report summary statistics for several of the measurements. See Table 9.12.

Table 9.12. Summary statistics for several of the measures used in the study of sports satisfaction (Lepir and Lakić 2025).

Measure	Mean	Median	Standard Deviation
Athletic Satisfaction	0.00	0.18	1.00
Extraversion	3.78	3.88	0.62
Agreeableness	3.65	3.67	0.54
Conscientiousness	3.82	3.89	0.59
Neuroticism	2.73	2.75	0.59
Openness	3.40	3.44	0.63

The first measure that we will consider is athletic satisfaction. We can observe from Table 9.12 that the mean level of satisfaction is 0 and the standard deviation is 1.00. It might seem a little strange that the data have this exact type of structure, but there is a very good reason for this. This measure has been specifically designed so that the mean is 0 and the standard deviation is 1, through a process called standardization. This makes individual satisfaction measurements very easy to interpret. For example, if an individual has a satisfaction measurement equal to 2.3, then we know that their level of satisfaction is above the mean satisfaction level because it is positive. The empirical rule will give us even more relevant information. We know from the empirical rule that approximately 95% of the data is within two standard deviations of the mean. Because the standard deviation is 1.0 in this case, the measurement of 2.3 is more than two standard deviations above the mean. This means that this rate of satisfaction is unusually high as it is beyond the range around the mean that contains 95% of the data. We can also gain more information by noting that the median is 0.18. This means that 50% of the data is above 0.18 while 50% of the data is below 0.18. This indicates that more than half of the respondents had a positive satisfaction level.

The next measurement is that of extraversion, which refers to whether an individual is focused externally, such as on social interaction rather than being more focused internally on their own thoughts and feelings. The measure is somewhat difficult to interpret so we will simply consider the summaries of the characteristics of the data. The mean value of extraversion is 3.78 and the median is 3.88. Once again, we can observe that the median is larger than the mean, which indicates that more than half of the values are larger than the mean. The standard deviation is 0.62. This indicates that almost all the measures of extraversion will be between $3.78 - 3 \times 0.62 = 3.78 - 1.86 = 1.92$ and $3.78 + 3 \times 0.62 = 3.78 + 1.86 = 5.64$. The summaries for the remaining measures can be interpreted in a similar way.

This page titled 9.6.4: Satisfaction with Sport Participation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.7: Summary

Numerical summaries of data are useful for providing information about the characteristics of data. This chapter considered two main characteristics: location and variation. The location of the data represents the center or typical value of the data and can be measured in several different ways. We discussed the mean (also known colloquially as the average), which is a measure of the center of mass of the data and the median—any point that divides the upper half of the data from the lower half. We also discussed variation, which is a measure of how far data can be found from the mean. Two measures, the range and the standard deviation, were discussed as measures of variation. Relating these two characteristics, we discussed how location and variation can be used together to approximately determine how much data lies within three specified distances from the mean. Finally, we considered percentiles, which are numbers that are above a specified percentage of the values in a set of data.

This page titled 9.7: Summary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.8: About the Quotation

Sir Arthur Conan Doyle (1859–1930) was a British writer and physician who is perhaps most famous for creating the character Sherlock Holmes in 1887. See Figure **9.8.1**. In the stories Holmes was known for the use of logical deduction to the point that it almost seemed magical.



Figure **9.8.1**: Sir Arthur Conan Dolye (1859–1930), the British writer who created the character of Sherlock Holmes (Public domain image).

This page titled 9.8: About the Quotation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.9: Practice and Exploration

1. Compute the mean, median, range, and standard deviation of the following five data values: 51, 79, 2, 80, and 38. To compute the standard deviation it is helpful to use a table like the one that is shown in Table 9.6.
2. Compute the mean, median, range, and standard deviation of the following six data values: 39, 28, 95, 42, 86, and 45. To compute the standard deviation it is helpful to use a table like the one that is shown in Table 9.6.
3. Compute the mean and the median of the following five data values: 51, 79, 2, 80, and 380. Explain why the mean and median are quite different from another.
4. Compute the mean and median of the following five data values: 1, 179, 120, 807, and 380. Explain why the mean and median are quite different from another.
5. Suppose that two exams from two different classes both have means equal to 81. The standard deviation was 15 for the first exam and 7 for the second exam. If a student must score at least 60 to pass the exam, which class would have a larger percentage of students who passed the class? Justify your answer.
6. Suppose a group of people has the mean age of 25 and the median age of 18 years old. Knowing that the oldest person in the room is 95, which measure of location will provide a better summary of the typical age in the group, the mean or the median?
7. Suppose that you have taken an exam in a class. The instructor has reported that the mean grade on the exam is 88 with a standard deviation equal to 7. You scored 90 on the exam. Using the empirical rule, decide if your score is unusually high. Explain the logic you used to come to your conclusion.
8. While having dinner with some friends, someone boasts that their first-year salary after graduating was $150,000. You know that for the alumni from their department the average starting salary is $67,000 with a standard deviation of $8,000. Given this information, how likely is it that this person is being truthful about their starting salary?
9. You and a friend are both taking the same history class at different times. Suppose in your class, the average is 85 with a standard deviation of 5 on the first exam. For your friend's class, the average is 75 with a standard deviation of 7. You scored a 90 and your friend scored an 89. Using the empirical rule, who did better? Explain.
10. If you flip a fair coin 100 times, the mean number of times you expect to observe a heads is 50 with a standard deviation of 5. Suppose your friend says they flipped a coin 100 times and got a heads 57 times. Using the empirical rule, state how unusual you think it would be to get heads this many times in 100 flips of a fair coin.
11. Jessica and Jonathan are planning to go to the same university in the fall. Jessica takes the ACT in January and Jonathan takes it in March. For the math portion of the exam, Jessica scores at the 92nd percentile, and Jonathan scores at the 94th percentile. Who did better on the exam in comparison to the other students who took the same exam?
12. A family in Illinois gave birth to twins, a boy and a girl. When they are 13 years old, their daughter is in the 82nd percentile for height and their son is in the 73rd. Who is taller for their gender?
13. Suppose that there are 150 people in a room attending a party and they have written their ages on slips of paper that they have given to the host. The host then computes that their own age is at the 77th percentile for the ages in the group. Approximately how many people at the party are younger than them?
14. The temperatures over the past year for a small town have been observed, and it is reported that the 25th percentile of the temperatures is 37.5 degrees Fahrenheit and the 75th percentile is 79.2 degrees Fahrenheit. What percentage of the temperatures are between these two numbers? If 365 temperatures have been observed, approximately how many of those will between these two temperatures?

This page titled 9.9: Practice and Exploration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.10: Collaborative Activity

This collaborative activity will consider computing the numerical summary measures studied in this chapter on some data that will be simulated by rolling some dice. You will be provided with a six-sided, a ten-sided, and a twenty-sided die. You will need to designate one person in your group to oversee rolling the dice, while another person will collect the data. This activity will allow you to investigate how these measures work for three different types of known models for generating data which correspond to the three dice.

For the first part of the activity, you will use the six-sided die. The process for obtaining the data is quite simple. The six-sided die will be rolled ten times, and the outcomes will be listed on the data sheet provided in Table 9.13. This process will then be used for the ten-sided die and the twenty sided-die.

Table 9.13. The data table for the collaborative activity.

Dice	1	2	3	4	5	6	7	8	9	10
Six-Sided										
Ten-Sided										
Twenty-Sided										

For example, suppose that when the six-sided die is rolled, you observe the rolls 6, 4, 1, 3, 1, 4, 4, 3, 6, and 5. The first roll of 6 is entered into the table in the row corresponding to the six-sided die under the column for the first roll. See Table 9.14. The remaining rolls are then listed under the appropriate columns in this roll as shown in the table. Next, suppose that when the ten-sided die is rolled you observe the rolls 7, 10, 8, 8, 2, 5, 1, 6, 5, and 7. As with the six-sided die, these rolls are listed under the appropriate column and row. Finally, the process is repeated for the twenty-sided die. The example rolls are shown on the table.

Table 9.14. Example data table for the collaborative activity.

Dice	1	2	3	4	5	6	7	8	9	10
Six-Sided	6	4	1	3	1	4	4	3	6	5
Ten-Sided	7	10	8	8	2	5	1	6	5	7
Twenty-Sided	7	2	18	13	12	6	18	5	15	17

Once the rolls have been completed, the group can work on calculating the mean, median, range, and standard deviation of the observed data. Begin by computing the mean of the rolls for each of the dice. This is done as demonstrated in this chapter: Add up all the observations and then divide by the number of observations. For the example data shown in Table 9.14, the mean roll for the six-sided dies is computed as

$$(6+4+1+3+1+4+4+3+6+5) \div 10 = 37 \div 10 = 3.7. \tag{9.10.1}$$

Similarly, for the ten-sided die the mean roll is

$$(7+10+8+8+2+5+1+6+5+7) \div 10 = 59 \div 10 = 5.9. \tag{9.10.2}$$

Finally, for the twenty-sided die the mean roll is

$$(7+2+18+13+12+6+18+5+15+17) \div 10 = 113 \div 10 = 11.3. \tag{9.10.3}$$

The means can be reported on a table like the one shown in Table 9.15.

Table 9.15. The data table for the mean, median, range, and standard deviation for the collaborative activity.

Dice	Mean	Median	Range	Standard Deviation
Six-Sided				
Ten-Sided				
Twenty-Sided				

To compute the median roll for each dice, you will first need to order the values from smallest to largest. Once they have been ordered, you will need to find a value that separates the smallest half of the data from the largest half of the data. Because there are ten values, the median will be somewhere between the fifth largest roll and the sixth largest roll. Following the suggestion from the chapter, we will average these two values together.

For the example data, the rolls for the six-sided die sorted from smallest to largest are 1, 1, 3, 3, 4, 4, 4, 5, 6, and 6. The fifth largest value is 4 and the sixth largest value is also 4, so when we take the mean of the two values we get $(4 + 4) \div 2 = 4$. Therefore, the median roll for the six-sided die is 4. Continuing with the rolls for the ten-sided die, the sorted values are 1, 2, 5, 5, 6, 7, 7, 8, 8, and 10. The fifth largest value is 6 and the sixth largest value is also 7, so when we take the mean of the two values we get $(6 + 7) div 2 = 6.5$. Therefore, the median roll for the six-sided die is 6.5. Finally, for the rolls for the twenty-sided die, the sorted values are 2, 5, 6, 7, 12, 13, 15, 17, 18, and 18. The fifth largest value is 12 and the sixth largest value is 13, so when we take the mean of the two values we get $(12 + 13) \div 2 = 12.5$. Therefore, the median roll for the six-sided die is 12.5.

To compute the range, we will use the ordered data values that we used when computing the median. The range is computed by subtracting the smallest value from the largest value. For the example data, the rolls for the six-sided die sorted from smallest to largest are 1, 1, 3, 3, 4, 4, 4, 5, 6, and 6. The largest value is 6 and the smallest value is 1; therefore the range is $6 - 1 = 5$. Continuing with the rolls for the ten-sided die, the sorted values are 1, 2, 5, 5, 6, 7, 7, 8, 8, and 10. The largest value is 10 and the smallest value is 1, so the range is $10 - 1 = 9$. Finally, for the rolls for the twenty-sided die, the sorted values are 2, 5, 6, 7, 12, 13, 15, 17, 18, and 18. The largest value is 18 and the smallest value is 2. The range is $18 - 2 = 16$.

The final calculations required for this activity are to compute the standard deviation of the observed rolls for each die. Because the standard deviation is somewhat more complicated to compute than the other measures, it is helpful to use a table like the one shown in Table 9.6. We will consider computing the standard deviation for the example data for the six-sided die. The corresponding table is shown in Table 9.16. From this table we can observe that the variance of the example data for the six-sided die is 3.12. To compute the standard deviation, we take the square root of this value to get $\sqrt{3.12} \approx 1.77$. The corresponding tables for the ten-sided and twenty-sided die are shown in Tables 9.17 and 9.18. From these tables we can observe that the variance for the example data for the ten-sided die is 7.65, which corresponds to a standard deviation of $\sqrt{7.65} \approx 2.76$, and the variance for the example data for the twenty-sided die is 34.67, which corresponds to a standard deviation of $\sqrt{34.67} \approx 5.89$.

The values of the summary statistics for the example data are shown in Table 9.19.

Table 9.16 The calculations used in computing the standard deviation of the of the data for the example data from the collaborative activity for the six-sided die. The data values are given in the first column, while for convenience the mean is listed in the second column. The squared absolute deviations are given in the fourth column.

Value	Mean	Absolute Deviation	Squared Absolute Deviation
6	3.7	2.3	5.29
4	3.7	0.3	0.09
1	3.7	2.7	7.29
3	3.7	0.7	0.49
1	3.7	2.7	7.29
4	3.7	0.3	0.09
4	3.7	0.3	0.09
3	3.7	0.7	0.49
6	3.7	2.3	5.29
5	3.7	1.3	1.69
Sum	28.1		
Variance	3.12		

Table 9.18 The calculations used in computing the standard deviation of the of the data for the example data from the collaborative activity for the ten-sided die. The data values are given in the first column, while for convenience the mean is listed in the second column. The squared absolute deviations are given in the fourth column.

Value	Mean	Absolute Deviation	Squared Absolute Deviation
7	5.9	1.1	1.21
10	5.9	4.1	16.81
8	5.9	2.1	4.41
8	5.9	2.1	4.41
2	5.9	3.9	15.21
5	5.9	0.9	0.81
1	5.9	4.9	24.01
6	5.9	0.1	0.01
5	5.9	0.9	0.81
7	5.9	1.1	1.21
Sum	68.9		
Variance	7.65		

Table 9.19 The calculations used in computing the standard deviation of the of the data for the example data from the collaborative activity for the twenty-sided die. The data values are given in the first column, while for convenience the mean is listed in the second column. The squared absolute deviations are given in the fourth column.

Value	Mean	Absolute Deviation	Squared Absolute Deviation
7	11.3	4.3	18.49
2	11.3	9.3	86.49
18	11.3	6.7	44.89
13	11.3	1.7	2.89
12	11.3	0.7	0.49
6	11.3	5.3	28.09
18	11.3	6.7	44.89
5	11.3	6.3	39.69
15	11.3	3.7	13.69
17	11.3	5.7	32.49
Sum	312.10		
Variance	34.68		

Table 9.19. Example data table for the mean, median, range, and standard deviation for the collaborative activity.

Dice	Mean	Median	Range	Standard Deviation
Six-Sided	3.7	4.0	5	1.76
Ten-Sided	5.9	6.5	9	2.76
Twenty-Sided	11.3	12.5	16	5.89

Questions

1. Compare the means of the ten rolls of the three different dice. Given what you know about the dice, namely the number of sides of each and the fact that each face has an equal probability, do the relative values of the means provide good information about the relative values of a typical roll from each die? If you rolled each die a very large number of times, what value do you think the mean roll for each die would be equal to?
2. Compare the medians of the ten rolls of the three different dice. Given what you know about the dice, namely the number of sides of each and the fact that each face has an equal probability, do the relative values of the medians provide good information about the relative values of a typical roll from each die? How do the medians compare to the means? If you rolled each die a very large number of times, what value do you think the mean roll for each die would be equal to?
3. Compare the ranges of the ten rolls of the three different dice. Given what you know about the dice, namely the number of sides of each and the fact that each face has an equal probability, do the relative values of the ranges provide good information about the relative variability of the rolls from each die? If you rolled each die a very large number of times, what value do you think the range of the rolls for each die would be equal to?
4. Compare the standard deviations of the ten rolls of the three different dice. Given what you know about the dice, namely the number of sides of each and the fact that each face has an equal probability, do the relative values of the ranges provide good information about the relative variability of the rolls from each die?

This page titled 9.10: Collaborative Activity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

9.11: References

Casella, G. and R. L. Berger. 2024. *Statistical Inference*. Chapmand and Hall.

Garg, Shikha, Lindsay Kim, Michael Whitaker, Alissa O'Halloran, Charisse Cummings, Rachel Holstein,Mila Prill, Shua J. Chai, Pam D. Kirley, Nisha B. Alden, et al. 2020. "Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019—COVID-Net, 14 States, March 1–30, 2020." *Morbidity and Mortality Weekly Report* 69, no. 15, 458.

Golestaneh, Ladan, Joel Neugarten, Molly Fisher, Henny H. Billett, Morayma Reyes Gil, Tanya Johns, Milagros Yunes, Michele H. Mokrzycki, Maria Coco, Keith C. Norris, et al. 2020. "The Association of Race and COVID-19 mortality." *EClinicalMedicine* 25, 100455.

Lepir, D., and S. Lakić. 2025. "Correlates of Sport Satisfaction: The Role of Success Level, Personality Traits, and Emotional Competence in Team and Individual Sports." *Psihologija* 3 (online issue). https://doi.org/10.2298/PSI240723003L

Xu, F., A. Meng, Y. Liu, J. Li, and N. Wu. 2025. "Effects of New Special Formula Fertilizer on Potato Growth, Yield, and Fertilizer Utilization Efficiency." *Plants* 14, 627. https://doi.org/10.3390/plants140406

Yeh, H.-P., J. A. Stone, S. M. Churchill, E. Bymer, and K. Davids. 2017. "Physical and Emotional Benefits of Different Exercise Environments Designed for Treadmill Running." *Environmental Research and Public Health* 14, no. 7, 752. https://doi.org/10.3390/ijerph14070752

This page titled 9.11: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

CHAPTER OVERVIEW

10: Frequency Distributions

“All interesting issues in natural history are questions of relative frequency, not single examples. Everything happens once amidst the richness of nature. But when an unanticipated phenomenon occurs again and again—finally turning into an expectation—then theories are overturned.”

–Stephen Jay Gould

10.1: What to Expect
10.2: Qualitative Data
10.3: Quantitative Data
10.4: Cross Classified Data
10.5: Simpson's Paradox
10.6: Critical Analysis
10.6.1: African American Student Debt
10.6.2: Race and School Discipline
10.7: Summary
10.8: About the Quotation
10.9: Practice and Exploration
10.10: Collaborative Activity
10.11: References

This page titled 10: Frequency Distributions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.1: What to Expect

The previous chapter demonstrated how data can be summarized with respect to several key characteristics, namely location, variation, and as percentiles. For qualitative data these types of summaries are often not informative, as the mathematical operations of summing or ranking may not be valid for the data. In other cases, we may have quantitative data for which we would like some additional detail, and these summaries are not enough. A frequency distribution can often be helpful in these cases. By the end of this chapter, you will be able to:

1. Construct and interpret frequency tables for qualitative data.
2. Construct and interpret frequency distributions for qualitative data.
3. Construct and interpret frequency distributions for cross-classified data.
4. Understand and interpret the importance of Simpson's paradox.

This page titled 10.1: What to Expect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.2: Qualitative Data

Suppose that we have observed data from a large survey-based study of alumni from the past decade of a large university. Among those, the variables observed in the study are the debt at graduation and a self-reported racial and ethnicity classification with the categories African American, Asian, Hispanic, and white. Sampling one hundred of these observations at random, we obtain the data given in Table 10.1. These data are hypothetical but were simulated to have characteristics like summary data given in (Espinosa et al. 2019).

Table 10.1 Data from a hypothetical alumni study. *Debt* indicates the debt load (in thousands of dollars) of the student at graduation

with an undergraduate degree. *Race* is coded as AF (African American), AS (Asian), HI (Hispanic), and WH (White).

Observation	Race	Debt		Observation	Race	Debt		Observation	Race	Debt
1	WH	15		34	WH	26		67	HI	25
2	WH	13		35	WH	0		68	AF	54
3	HI	0		36	AF	44		69	WH	14
4	WH	48		37	AF	4		70	HI	22
5	HI	23		38	WH	35		71	WH	0
6	WH	36		39	HI	5		72	WH	27
7	WH	32		40	AF	57		73	AF	6
8	HI	34		41	WH	32		74	WH	22
9	WH	40		42	WH	5		75	AF	0
10	WH	23		43	WH	0		76	WH	34
11	WH	16		44	WH	24		77	WH	0
12	HI	24		45	WH	0		78	WH	48
13	HI	28		46	WH	0		79	WH	31
14	AF	36		47	WH	0		80	WH	0
15	WH	37		48	WH	0		81	WH	42
16	AF	0		49	WH	3		82	HI	0
17	WH	0		50	HI	13		83	HI	0
18	WH	0		51	WH	0		84	WH	0
19	AS	38		52	HI	18		85	WH	47
20	WH	0		53	WH	21		86	WH	0
21	HI	0		54	AS	0		87	AS	0
22	WH	0		55	HI	25		88	AF	14
23	WH	50		56	AS	0		89	WH	47
24	WH	24		57	HI	0		90	WH	21
25	WH	59		58	WH	57		91	WH	0
26	WH	38		59	WH	0		92	HI	12
27	HI	22		60	WH	23		93	WH	34
28	HI	43		61	WH	42		94	WH	23
29	WH	35		62	WH	7		95	WH	45
30	HI	4		63	WH	11		96	WH	36
31	WH	0		64	HI	22		97	WH	26

32	AS	0		65	AF	54		98	WH	15
33	AF	27		66	HI	0		99	WH	37
								100	WH	26

In looking at Table 10.2 we can easily summarize the debt data in many ways. The table shows two measures of location, the mean and the median, as well as a measure of variation, the standard deviation. Comparing the means and the medians, we can conclude that African American students have the largest amount of debt at graduation, followed by white, Hispanic, and then Asian American students. The measures of variation tend to follow the same pattern, though the measures are much closer. We will not attempt to explain why these differences occur; at this point we are only interested in summarizing the observed data.

Table 10.2 Summary Statistics of Debt by Race. *Debt* indicates the debt load (in thousands of dollars) of the student at graduation with an undergraduate degree. *Race* is coded as AF (African American), AS (Asian), HI (Hispanic), and WH (White).

Race	Mean	Median	Standard Deviation
AF	26.9	27.0	23.1
AS	7.6	0.0	17.0
HI	14.5	15.5	13.1
WH	21.9	23.0	18.8

The next question is how to summarize the data on race. As this data is qualitative and on a nominal measurement scale, no mathematical calculations can be performed. Hence, we cannot compute an average because we cannot add the data. We cannot compute a median because we cannot sort the data. The only mathematical comparison that can be performed is to determine whether two observed individuals have the same race, and hence the only summary that is valid for us to do is to compute how many of each race are represented in the data. When this information is presented for each of the categories in a tabular format, the resulting table is called a frequency distribution and the individual counts for each category are called the frequencies.

Definition: Frequency Distribution

The *frequency distribution* of a set of qualitative data is a table that contains the number of times each category of the data is observed. The individual counts for each category are called the *frequencies* of the categories.

The frequency distribution for the race data observed in Table 10.1 is given in Table 10.3. The first row of this table corresponds to the data for African American alumni. The frequency reported in the table indicates that of 100 observations, 11 correspond to African American alumni: 14, 16, 33, 36, 37, 40, 65, 68, 73, 75, and 88. Similarly, the frequency reported in the table indicates that there are 5 Asian American observations for those alumni: 19, 32, 54, 56, and 87. The reported frequencies of 22 for Hispanic and 62 for white alumni are interpreted in similar ways. Note that adding all these frequencies equals 100, which is the total number observations in the set of data.

Table 10.3 Frequencies, relative frequencies, proportions, and percentages for each of the categories of the race data contained in Table 10.1.

Race	Frequency	Relative Frequency	Percent
AF	11	0.11	11%
AS	5	0.05	5%
HI	22	0.22	22%
WH	62	0.62	62%
Total	100	1.00	100%

A frequency distribution is useful in determining if there are categories for a variable that corresponds to a large part of the data, and in turn if there are categories that are quite rare. For some sets of data, a frequency distribution may indicate that all the categories may roughly occur an equal number of times. The frequency distribution for the race data observed in Table 10.3 indicates that most of the Alumni observed in the Alumni data were white, followed by Hispanic, African American, and Asian American. It is noteworthy in this case that Asian American Alumni were quite rare in this data. Whether this is reflective of a similar trend in all the alumni from the university or if it is an artifact of this set of observed data would have to be researched further. Note that if the frequency distribution shows that the data do not generally reflect the characteristics of the population of alumni, this may be an indication of a nonrepresentative sample.

Comparing frequencies directly may not be as convenient as looking at what proportion or percentage of the data is in each category. This can be particularly true when there are more than a few categories, and when there is a large range of variation of the frequencies. If a frequency is divided by the total number of observations in the set of data, the result is the relative frequency, which measures the proportion of the data is in the category. Each proportion can be converted to a percentage by multiplying it by 100%. A table of either relative frequencies or percentages is usually still called a frequency distribution.

Definition: Frequency Distribution

The *frequency distribution* of a set of qualitative data is a table that contains the number of times each category of the data is observed, the proportion of data that was observed in each category, or the percentage of data that was observed in each category. The individual counts for each category are called the *frequencies* of the categories. The *relative frequencies* are computed by dividing each frequency by the total number of observations in the set of data. The *percentages* are computed by multiplying the relative frequencies by 100%.

The relative frequencies and percentages for the race variable in the Alumni study are given in the third and fourth columns of Table 10.3. It should be noted that because there were exactly one hundred observations in these data, the frequencies, relative frequencies, and percentages look pretty much like the frequencies, and the conclusions for this table are basically the same as before.

As another example, consider a research study where twenty-five students who identified as female were randomly sampled from the graduating class at a small midwestern college. Of concern to the researcher is the level of gender discrimination the students experienced during their time at the college. To measure discrimination, the study used eight questions based on the perceptions regarding gender discrimination using a standard scale developed and validated by other researchers. Each indicator in the survey was scored using a five-point Likert scale to ask the respondents how strongly they agree or disagree with a statement. The observations were then added to get a score on how much gender discrimination was experienced. The observed data are given in Table 10.4.

Table 10.4 Observed responses to questions on gender discrimination of 25 random sampled students who identify as female at a small Midwestern college.

Individual	1	2	3	4	5	6	7	8	Sum
1	4	3	4	5	4	3	2	3	28
2	5	5	5	4	3	2	3	2	29
3	3	2	1	2	1	2	3	5	19
4	3	4	4	5	5	4	5	3	33
5	5	5	4	3	1	1	1	1	21
6	1	1	1	3	2	4	3	4	19
7	1	1	2	3	4	5	4	5	25
8	2	3	5	4	3	1	1	1	20
9	3	1	2	3	2	2	4	5	22
10	5	4	5	5	3	3	3	4	32
11	4	5	5	5	4	5	4	4	36
12	1	3	1	3	2	4	4	3	21
13	3	4	3	2	3	4	3	3	25
14	5	5	5	5	5	5	5	4	39
15	4	4	4	5	4	5	5	5	36
16	1	1	1	1	2	3	2	4	15
17	3	1	2	2	2	3	3	4	20
18	5	5	4	5	5	5	5	4	38
19	2	2	1	1	1	1	1	1	10
20	4	4	3	2	3	3	4	2	25
21	2	2	2	2	4	3	4	4	23
22	1	1	1	1	2	3	5	4	18
23	5	4	2	3	2	4	4	3	27
24	5	5	5	4	5	5	3	3	35
25	2	1	1	1	1	1	1	2	10

Each of the questions in the survey are qualitative on an ordinal measurement scale. When the measurement scale is ordinal, it is important to list the categories on the table in the order implied by the ordinal structure of the measurement scale. Otherwise, the calculations are the same as in the previous example. For example, in constructing a frequency table for the responses from the first question, it can be noted that 5 individuals scored the statement as 1, so the frequency for that category is 5. To get the relative frequency take the frequency and divide by the number of observations. Hence the relative frequency is $5 \div 25 = 0.20$. The corresponding percentage is computed by multiplying the relative frequency by 100%, and hence the percentage for that category is $0.20 \times \times 100\% = 20\%$. The remaining calculations follow in a similar manner and are collected in Table 10.5.

Table 10.5 Frequency distributions for the eight questions in the survey on gender discrimination from Table 10.4.

	Response					
Question	1	2	3	4	5	Total
1	5	4	5	4	7	25
	0.20	0.16	0.20	0.16	0.28	1.00
	20%	16%	20%	16%	28%	100%
2	7	3	3	6	6	25
	0.28	0.12	0.12	0.24	0.24	1.00
	28%	12%	12%	24%	24%	100%
3	7	5	2	5	6	25
	0.28	0.20	0.08	0.20	0.24	1.00
	28%	20%	8%	20%	24%	100%
4	4	5	6	3	7	25
	0.16	0.20	0.24	0.12	0.28	1.00
	16%	20%	24%	12%	28%	100%
5	4	7	5	5	4	25
	0.16	0.28	0.20	0.20	0.16	1.00
	16%	28%	20%	20%	16%	100%
6	4	3	7	5	6	25
	0.16	0.12	0.28	0.20	0.24	1.00
	16%	12%	28%	20%	24%	100%
7	4	2	7	7	5	25
	0.16	0.08	0.28	0.28	0.20	1.00
	16%	8%	28%	28%	20%	100%
8	3	3	6	9	4	25
	0.12	0.12	0.28	0.36	0.16	1.00
	12%	12%	28%	36%	16%	100%

Note that each row of Table 10.5 is a frequency distribution for the responses for each of the 8 questions. For the first question the responses look somewhat even, indicating that for that question there did not seem to be a preference for higher or lower responses. For the second and third questions, there is more division in that the higher frequencies are for the lower and higher responses. This indicates that there was little middle ground for these questions and that individuals taking part in the study either felt no discrimination for those instances or felt a great deal of discrimination. For questions 7 and 8, the opposite trend is observed; it appears that many individuals had some experience with the situation described in the questions.

This page titled 10.2: Qualitative Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.3: Quantitative Data

Frequency distributions can also be computed on quantitative data, though there is sometimes an extra step required before this can be done. To motivate our discussion, we will consider research on the number of water quality violations per year in a fictional state with 100 counties. Of interest in this research is comparing the number of violations with the median income of the counties to see if lower income counties are associated with more violations. This example is motivated by similar research from a nationwide study, but the data has been made more manageable for the purpose of this example (Bae et al. 2021). The data are given in Table 10.6.

Table 10.6 Simulated data on the number of water quality violations of public drinking sources and the median income of 100

counties of a hypothetical state over a year.

County	Violations	Income	County	Violations	Income	County	Violations	Income
1	2	31	34	0	54	67	2	63
2	2	39	35	1	64	68	1	22
3	1	30	36	4	41	69	1	17
4	0	57	37	2	34	70	1	49
5	3	24	38	1	25	71	1	23
6	8	20	39	2	28	72	0	55
7	2	56	40	2	19	73	2	18
8	1	41	41	1	55	74	1	18
9	1	81	42	0	80	75	1	54
10	2	45	43	1	10	76	1	13
11	2	23	44	1	18	77	2	47
12	1	32	45	2	20	78	6	37
13	2	25	46	0	64	79	2	44
14	9	22	47	1	12	80	2	45
15	2	26	48	2	14	81	1	27
16	2	27	49	5	24	82	0	56
17	2	28	50	1	17	83	2	48
18	1	47	51	1	35	84	2	49
19	1	19	52	1	39	85	1	32
20	1	20	53	1	18	86	2	24
21	1	42	54	1	41	87	3	25
22	2	38	55	2	40	88	1	41
23	7	19	56	2	17	89	2	34
24	1	16	57	1	22	90	2	42
25	1	33	58	2	30	91	1	49
26	1	41	59	2	26	92	1	27
27	2	36	60	1	18	93	1	18
28	2	50	61	0	76	94	2	17
29	2	29	62	2	55	95	1	34
30	0	63	63	0	56	96	2	33
31	1	24	64	1	26	97	2	18
32	2	17	65	1	58	98	2	19

33	1	37	66	1	56	99	1	28
						100	9	20

We first consider constructing a frequency distribution for the number of violations over the 100 counties in the state. Note that this is quantitative data, but it would also suffice to just treat each possible value of the variable as its own category, and then follow the same calculations shown in earlier. The resulting frequency distribution complete with relative frequencies and percentages are given in Table 10.7. This frequency distribution is interpreted just as any other frequency distribution. For example, one can observe that most of the counties have one or two violations. Of the remaining numbers of violations, having no violations happens the most. While there are several counties with more than two violations, they are relatively rare.

Table 10.7 Frequency distribution for the number of water quality violations for the 100 counties.

Number	Frequency	Relative Frequency	Percent
0	9	0.09	9%
1	44	0.44	44%
2	38	0.38	38%
3	2	0.02	2%
4	1	0.01	1%
5	1	0.01	1%
6	1	0.01	1%
7	1	0.01	1%
8	1	0.01	1%
9	2	0.02	2%
Total	100	1.00	100%

With the success of computing the frequency distribution of the data on the number of water violations, we can move to constructing a frequency distribution on the median incomes of the counties. If we consider taking the same approach, which uses each possible value of the variable as a category for the frequency table, we begin to see a potential problem. Looking through the data, we find that this variable takes on many possible values. In fact, there are 46 different values of the income variable, and each of these values would need a line in the frequency table for this data. The table also has many frequencies that are below 3 because most of the values only occur a few times. For these two reasons this table would not be very useful, as it would be difficult to read because it is so large and would not show where data is concentrated because most of the frequencies would be small.

For quantitative data with many possible values, a frequency distribution can be constructed by creating classes, which are ranges of possible values for the variable. The frequencies of the classes are then computed as the number of observed values of the variable that fall within each specified range. Relative frequencies and percentages are then computed as detailed earlier. When constructing such a table, researchers need to be careful in defining these classes. First, each data value of the variable should fall in exactly one class. That means that the classes should not overlap, and that they should cover the entire range of the variable.

Definition: Classes

A set of *classes* for a quantitative variable is a set of non-overlapping ranges that cover all the values of the variable so that each value of the variable falls in exactly one class.

The classes should not overlap because a data value should not be counted twice when computing the frequencies. Similarly, the range of the classes should cover the range of values of the variable so that each data value gets included in the frequency.

Let us consider constructing classes for the median income variable for the data in Table 10.6. Recalling the median incomes are reported in thousands of dollars, the smallest median income is 10 and the largest median income is 81. Therefore, the set of classes should cover this range. Next, the researcher must decide how to divide up this range into classes. The simplest approach is to create classes of equal length, dividing up the range of the data into equal parts. There should generally be at least four or five classes, but no more than ten. Classes that are not of equal size should generally be avoided as it makes the frequencies difficult to interpret and leaves the table open to manipulation that can be used to hide some aspects of the data. For this data it is convenient to divide it into eight classes starting with the class 10 to 19 and ending with the class 80 to 89, with equally sized classes in between. The frequency table corresponding to these classes is given in Table 10.8.

Table 10.8 Frequency distribution of the median incomes for the 100 counties.

Class	Frequency	Relative Frequency	Percent
10 to 19	21	0.21	21%
20 to 29	26	0.26	26%
30 to 39	17	0.17	17%
40 to 49	17	0.17	17%
50 to 59	12	0.12	12%
60 to 69	4	0.04	4%
70 to 79	1	0.01	1%
80 to 89	2	0.02	2%
Total	100	1.00	100%

The frequency distribution given in Table 10.8 can be interpreted just as any other frequency distribution. From the table we can observe that the median income for most of the counties is between 10 thousand and 60 thousand dollars with a slightly higher concentration on the lower end of the scale. It is most common for the counties to have a median income between 10 thousand and 20 thousand dollars. The counties in the lower portion of the table, corresponding to counties that have higher incomes, do seem unusual and extreme in that their incomes seem much larger than most of the counties.

A similar type of calculation can be done with the debt variable from the study of race and student debt. For the data shown in Table 10.1 the variable corresponding to student debt is quantitative with a range starting at 0 (no debt) to 68 thousand dollars. As with the median county income data we have constructed classes of equal length, starting with the class 0 to 9 and ending with the class 60 to 69. The corresponding frequency distribution for the data using these classes is given in Table 10.9. These data are interesting in that an unusual pattern is observed in the frequency distribution. There are two classes with relatively large frequencies, the first class being 0 to 9 with the largest frequency of 38, and the class 20 to 29 with the next largest frequency of 21. What is interesting is that in many frequency distributions the classes with large frequencies tend to be grouped together with no classes with a small frequency between them. In this case, these two classes are separated by the class 10 to 19, which has a frequency of 10, less than half that of the second largest frequency. For this data it means that there seem to be two sub-populations in the data, one with very low debt and one with debt between 10 thousand and 19 thousand dollars. Such an observation would generally warrant further investigation in a research project to see if this behavior could be explained by any of the other observed variables.

Table 10.9 Frequency distribution for the debt at graduation for the sample of 100 alumni.

Class	Frequency	Relative Frequency	Percent
0 to 9	38	0.38	38%
10 to 19	10	0.10	10%
20 to 29	21	0.21	21%
30 to 39	14	0.14	14%
40 to 49	10	0.10	10%
50 to 59	6	0.06	6%
60 to 69	1	0.01	1%
Total	100	1.00	100%

The choice of the classes can have a great effect on how the frequency table is interpreted. To demonstrate how this can occur, we will consider several hypothetical frequency distributions and then show how they can be manipulated in different ways that may make readers of the research respond in different ways. For the first case, consider a study that looks at the number of police calls per day in an urban area for a 100-day period. The original frequency distribution is given in Table 10.10. For brevity we have only included the frequencies in this table as they are sufficient for our purpose here. The number of police calls per day ranges from 0 to 8. Most days get a few calls, but there are a few days that get many more. Now consider what would happen if we collapsed the last four classes in the frequency distribution (see Table 10.11). What happens in this case is that the last class becomes one of the classes with the largest frequencies, resulting in an unnecessary emphasis on days with a relatively large number of police calls. What makes the problem worse is that the last class does not have an upper bound, so the reader cannot distinguish what the true upper bound may be. Ten calls? Twenty calls? There is no way to know from the second table that the upper bound on the range was eight calls. Someone might use a tactic like this to make the neighborhood seem more “dangerous” than it really is.

Table 10.10 Frequency distribution of the number of police calls received per day over a 100-day period in an urban area.

Calls	Frequency
0	10
1	20
2	27
3	13
4	10
5	9
6	7
7	3
8	1

Table 10.11 Frequency distribution of the number of police calls received per day over a 100-day period in an urban area.

Calls	Frequency
0	10
1	20
2	27
3	13
4	10
5 or more	20

As a second example of how frequency distributions can be manipulated, consider a study that randomly sampled individuals from a large population. The authors of the study are attempting to argue that the individuals who took part in the study came from all age groups between 20 and 60 years old. A frequency distribution for the observed sample is given in Table 10.12.

Table 10.12 Frequency distribution of the ages of participants in a social justice study.

Age	Frequency
20 to 24	71
25 to 29	8
30 to 34	5
35 to 39	86
40 to 44	82
45 to 49	65
50 to 54	76
55 to 59	6

What is immediately obvious in this frequency distribution is that although there are many age ranges with a substantial number of observations, the study found relatively few individuals willing to participate who were aged between 25 and 34, or who were older than 54. This could indicate that the conclusions from the study may not be relevant for those individuals who are in the age ranges with fewer observed individuals. Now consider combining the adjacent classes to get the frequency distribution given in Table 10.13. In this second table the gap in the age coverage has been completely obscured by combining the adjacent classes.

Table 10.13 Frequency distribution of the ages of participants in a social justice study with half as many classes.

Age	Frequency
20 to 29	79
30 to 39	91
40 to 49	147
50 to 54	82

It can be very difficult to determine if a researcher is purposely being misleading with the way they constructed a frequency distribution. When assessing frequency distributions, it is usually enough to ask yourself if the classes look strange. Usually, classes should follow some logical pattern. It seems natural to break up data into classes based on multiples of ten, for example. If you

observe classes broken up in awkward ways or in a way such that some classes are much larger than others, you should look in the research article to see if the researchers justify the way they constructed the classes. If no such justification is given, the conclusions based on these frequency tables might be suspect.

This page titled 10.3: Quantitative Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.4: Cross Classified Data

When more than one variable is observed, it is often of interest to researchers to explore whether the two variables have an association between them. That is, does one value or range of values of one of the variables tend to occur along with one value or range of values of the other variable. As we have seen, it could be that there is a direct relationship between the variables, but it could also be the case that the variables are both related some other common variables that have not been observed. In either case, knowledge of such an association can be useful to researchers either in predicting one variable from another, or in understanding how different variables are associated with or affect other variables.

If the frequency distribution of two variables are computed separately, we do not obtain any information about how the two variables behave in relation to one another. These types or relationships can be studied using frequency distributions as long as both variables are tracked simultaneously. To consider this idea in detail, we return to the example data given in Table 10.1. The frequency distribution for the race and ethnicity variable is given in Table 10.3. While the means were computed for the debt variable with respect to each category of race, the frequency distribution for the debt variable was given in Table 10.9. Note that Tables 10.3 and 10.9 do not give us the same information that we see with the means in Table 10.2, namely that African American alumni have more debt than individuals in the other categories. This is because Tables 10.3 and 10.9 are considering the two variables separately and therefore do not contain any information of the simultaneous behavior of the variables.

To get information on how the two variables behave simultaneously using frequency distributions, a researcher must construct what is called a cross-classified frequency table. This is based on the same idea as a frequency table with the difference that the table contains frequencies for all possible pairs of values of the two variables that can occur within an observation. The relative frequencies and percentages are computed as detailed earlier.

Definition: Cross Classified Frequency Distribution

A cross-classified frequency distribution of two variables is a table that contains the frequencies, relative frequencies, and percentages of the number of times each pair of values occur in an observation of the two variables in the data.

An example of a cross-classified frequency distribution is given in Table 10.14. The frequency distribution shown in this table is a cross classified table for the data on race and ethnicity and student debt at graduation for the data shown in Table 10.1. The columns of this table correspond to the different classifications of race and ethnicity, and the rows of this table correspond to the amount of debt at graduation using the classes defined earlier.

Table 10.14 Cross-classified frequency distribution of debt at graduation by race and ethnicity for the sample of 100 alumni. The first row of each entry is the observed frequency, the second row is the relative frequency, and the third row is the percentage.

	Race and Ethnicity				
Debt	AF	AS	HI	WH	Total
0 to 9	4	4	9	21	38
	0.04	0.04	0.09	0.21	0.38
	4%	4%	9%	21%	38%
10 to 19	1	0	3	6	10
	0.01	0.00	0.03	0.06	0.10
	1%	0%	3%	6%	10%
20 to 29	1	0	8	12	21
	0.01	0.00	0.08	0.12	0.21
	1%	0%	8%	12%	21%
30 to 39	1	1	1	11	14
	0.01	0.01	0.01	0.11	0.14
	1%	1%	1%	11%	14%
40 to 49	1	0	1	8	10
	0.01	0.00	0.01	0.08	0.10
	1%	0%	1%	8%	10%
50 to 59	3	0	0	3	6
	0.03	0.00	0.00	0.03	0.06
	3%	0%	0%	3%	6%
60 to 69	0	0	0	1	1
	0.00	0.00	0.00	0.01	0.01
	0%	0%	0%	1%	1%
Total	11	5	22	62	100
	0.11	0.05	0.22	0.62	1.00
	11%	5%	22%	62%	100%

While this table looks very large and complicated, it is as simple to interpret as any frequency distribution if we keep the cross-classifications in mind. For example, the uppermost left cell of this table corresponds to individuals from the study who are African American and whose debt at graduation was between 0 and 9 thousand dollars. The frequency distribution reports that there are four individuals that matched both characteristics, corresponding to 4% of the total number of observations. For the row 20 to 29 under the column for Hispanic, we can observe that there are 12 individuals that matched both characteristics, corresponding to 12% of the total number of observations. Note that this table also allows us to easily compare the frequency distributions of the debt by race and vice-versa. In this case it does not appear, for example, that other than the number of individuals within each race and ethnicity classification, there does not seem to be a major difference in the frequency distributions of debt by race, except for

the fact that African American students have a relatively large frequency for the debt class 50 to 59. This is the reason that the African American individuals have the large mean and median debt as shown in Table 10.2.

Earlier we considered a research study where twenty-five students who identified as female were randomly sampled from the graduating class at a small midwestern college. The researchers used eight questions based on the perceptions regarding gender discrimination. The observations were then added to get a score on how much gender discrimination was experienced. The observed data are given in Table 10.4. As all eight questions used by the measurement system are related to experiences with gender discrimination, we would expect that the responses should be associated with one another. That is, an individual who has experienced gender discrimination would likely give higher scores to many of the questions, whereas an individual who has not experienced gender discrimination may tend to give lower scores to each of the questions. Therefore, we would expect that if an individual score on question as a 5 on Likert scale, then they would tend to give higher scores for the other questions as well. To investigate whether this trend holds for the data observed in Table 10.4 we have computed a cross classified frequency distribution from the observed responses from the first question and the second question. This frequency distribution is given in Table 10.15.

Table 10.15 Cross-classified frequency distributions for the first two questions of the survey on gender discrimination.

Question 1	Question 2 Response					
Response	1	2	3	4	5	Total
1	4	0	1	0	0	5
	0.16	0.00	0.04	0.00	0.00	0.10
	16%	0%	4%	0%	0%	20%
2	1	2	1	0	0	4
	0.04	0.08	0.04	0.00	0.00	0.16
	4%	8%	4%	0%	0%	16%
3	2	1	0	2	0	5
	0.08	0.04	0.00	0.08	0.00	0.20
	8%	4%	0%	8%	0%	20%
4	0	0	1	2	1	4
	0.00	0.00	0.04	0.08	0.04	0.16
	0%	0%	4%	8%	4%	16%
5	0	0	0	2	5	7
	0.00	0.00	0.00	0.08	0.20	0.28
	0%	0%	0%	8%	20%	28%
Total	7	3	3	6	6	25
	0.28	0.12	0.12	0.24	0.24	1.00
	28%	12%	12%	24%	24%	100%

First note that all the frequencies are quite small in this table. Even with the small frequencies, there is an interesting trend to note in this table. Looking at the individuals who responded with a score of either 1 or 2 for the first question, one can observe that all of the corresponding responses for the second question are also either 1, 2, or 3. Similarly, individuals who responded with a score of either 4 or 5 for the first question also responded to the second question with either 3, 4, or 5. Therefore, the response to the first question tends to provide a information about what the response will be for the second question. From the data, those who gave a

low score on the first question also tended to give a low score on the second question. Those who gave a high score on the first question also tended to give a high score on the second question.

This page titled 10.4: Cross Classified Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.5: Simpson's Paradox

Capital punishment has been controversial for many decades in the United States. One of the major concerns is that capital punishment may be unevenly applied to defendants based on factors such as race, ethnicity, and socioeconomic status. One study considered data on race of the defendant and whether the death penalty was imposed for defendants from twenty Florida counties during 1976–1977 (Radelet 1981). The study looked at the issue in response to new capital punishment statutes that were enacted after a supreme court ruling. The counties were selected randomly with larger counties being more likely to be selected. Hence the data is based on a cluster sample, but without simple random sampling of the clusters. All cases with either African American or white defendants for which there was complete data were included in the study. A cross-classified frequency table of the race of the defendant and whether they received the death penalty for the observed data is given in Table 10.16.

Table 10.16 Race of the defendant and whether the death penalty was imposed for defendants from twenty Florida counties during 1976–1977.

	Death Penalty		
Defendant	Yes	No	Total
White	19	141	160
	0.058	0.433	0.491
	5.8%	43.3%	49.1%
African American	17	149	166
	0.052	0.457	0.509
	5.2%	45.7%	50.9%
Total	36	290	326
	0.110	0.890	1.0001
	11%	89%	100%

Looking at the data in table, there does not really seem to be any bias in the application of the death penalty as related to race. There are 160 white defendants in the study and 19 of them, or $19 \div 160 \times 100\% = 11.9\%$ of the white defendants received the death penalty. Similarly, there are a total of 166 African American defendants in the study and 17 of them, or $17 \div 166 \times 100\% = 10.2\%$ of the African American defendants received the death penalty. The percentages are close, and if anything, it appears that the rate at which African American defendants receive the death penalty is less than for white defendants. This is a surprising result given the secondary evidence available on this subject, and it takes a little more detailed analysis to see what is going on with this data.

The frequency distribution given in Table 10.17 has been broken down even further, and includes not only the race of the defendant, but also the race of the victim. When this second classification is added, something very interesting is observed. First consider white defendants. When a white defendant was convicted of murdering a white victim, $19 \div 151 \times 100\% = 12.6\%$ of the defendants received the death penalty. When a white defendant was convicted of murdering an African American victim, none of the defendants received the death penalty. These entries alone show an incredible bias for white defendants who murdered African American victims. Moving to the case of African American defendants, those convicted of murdering a white victim received the death penalty $11 \div 63 \times 100\% = 17\%$ of the time, which is larger than the rate for white defendants convicted of murdering white victims. Finally, for African American defendants who were convicted of murdering African American victims, the defendant received the death penalty $6 \div 103 \times 100\% = 5.8\%$ of the time. It is clear from this data that African American defendants convicted of murdering white victims receive the death penalty in far more cases than any other, and that when the victim is African American, the defendant receives the death penalty in far fewer cases.

Table 10.17 Race of the defendant and whether the death penalty was imposed for defendants from twenty Florida counties during 1976–1977 broken down by race of the victim.

		Death Penalty		
Defendant	Victim	Yes	No	Total
White	White	19	132	151
		0.058	0.405	0.463
		5.8%	40.5%	46.3%
White	African American	0	9	9
		0.000	0.003	0.003
		0%	0.3%	0.3%
African American	White	11	52	63
		0.033	0.160	0.193
		3.3%	16.0%	19.3%
African American	African American	6	97	103
		0.018	0.298	0.316
		1.8%	29.8%	31.6%
Total		36	290	326
		0.110	0.890	1.000
		11.0%	89.0%	100%

This example begs the obvious question: Why are the trends in the second table not visible in the first table? There are several things going on in Table 10.17 that contribute to why a trend is not observed in Table 10.16, but what essentially happens is that the low rate of the African American defendants being given the death penalty for African American victims is added to the high rate of African American defendants being given the death penalty for white victims. Essentially, these two opposing trends cancel each other out. In the end what is really covered up is two biases: one against African American defendants and one against African American victims.

The way the opposite trends can average themselves together to cover each other up is an example of what is known in statistics as Simpson's Paradox. Someone reading a research paper and looking at the results of cross-classified frequency distributions, or even for a single frequency distribution, should be aware of this potential problem, particularly when the trend exhibited in a frequency distribution is unusual or goes against common knowledge about an issue. It is, however, very difficult to ascertain whether a frequency distribution may have this problem unless additional data are provided. The problem here is essentially one of confounding variables, and just as we considered confounding when we considered empiricism, a critical evaluator of research must consider potential confounding variables when evaluating any frequency distributions.

This page titled 10.5: Simpson's Paradox is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.6: Critical Analysis

The critical analysis of frequency distributions for qualitative data is relatively straightforward. The frequency distribution can be used by researchers to essentially identify which values of a qualitative variable occur more often than other variables. Other trends, such as whether all the values of the variable occur at the same rate in the data can also be identified. Tables for qualitative data are usually difficult to manipulate and therefore one should just assure themselves that the authors of the research are interpreting the data in frequency distribution correctly.

The critical analysis of frequency distributions for quantitative data is more complicated. In many cases the classes must be constructed by the researcher, and there is no unique way to do this. When assessing a research paper, first see if the classes have the same size. If the classes are of different sizes, then there should be a natural reason for this implied by the nature of the data, or the researchers should explain why the classes were constructed in the way that they were. Be careful interpreting frequency distributions when the classes are of different sizes, as larger classes will often naturally have larger frequencies. Even when classes have the same size, the placement of the classes should be considered as well, as trends can be covered up, or nonexistent trends can sometimes be created just by the placement of the classes. Whenever a major hypothesis of a research paper depends on classes that have been constructed by the researcher, one should be very cautious about interpreting the results of the study.

Cross-classified frequency distributions require even more care to interpret. Along with all the potential issues of regular frequency distributions, the possibility that trends are being covered up can be a major difficulty.

In summary, the key questions in critically analyzing frequency distributions in a study are:

1. What are the classes that have high frequencies, and what are the classes that have low frequencies?
2. If the data are quantitative, how were the classes chosen? Are the classes of the same size? Have the classes been chosen in such a way that they might cover up some trend in the frequencies?
3. For cross-classified frequency distributions, what trends appear in the table? Is there some association between the variables?
4. For cross-classified data, could some trends be hidden due to the way the table was constructed?

This page titled 10.6: Critical Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.6.1: African American Student Debt

A study of college graduates for the years 2015 and 2016 focused on the amount of debt that students had undertaken to finance their education (Baum 2019). The particular focus of the study was on African American students who are more likely to borrow than students from any other racial or ethnic group. The article also focused on how this problem is often concealed when race is not considered with student debt.

Frequency distributions of the amount of debt by race and ethnicity are given in Table 10.18. Each row of this table is a frequency distribution where only the percentages are reported. For white students the highest frequency occurs at no debt, but substantial percentages remain throughout all the classes. The trend is very similar for Hispanic students. For Asian students the trend is very different. The largest frequency by far is for no debt with these students, with the remaining frequencies being much smaller. The trend for African American students is much different in that the largest frequency is for having the most amount of debt. The is concerning because it is generally African American families who have the fewest economic resources.

Table 10.18 Frequency distributions of debt by race for students earning a bachelor's degree for 2015–2016 graduates. Adapted from Baum (2019).

	Amount Borrowed					
Race or Ethnicity						
No Debt	$1 to $9,999	$10,000 to $19,999	$20,000 to $29,999	$30,000 to $39,999	$40,000 or more	
White	30%	10%	13%	18%	12%	18%
African- American	14%	12%	10%	15%	16%	33%
Hispanic	33%	14%	15%	15%	10%	13%
Asian	41%	11%	14%	16%	9%	9%
All	29%	11%	13%	17%	12%	18%

The last row in the table is the frequency distribution for the combined data of all races. Note that this frequency distribution is much closer to that of Hispanic and white students, and the overburden for African American students is completely lost. This is an example of Simpson's Paradox.

This page titled 10.6.1: African American Student Debt is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.6.2: Race and School Discipline

While the results of litigation and legislative efforts have focused on students of all races and ethnicities having equal access to education, a great deal of research has shown that opportunities within the same schools can differ greatly. Racial and ethnic disparities are an ongoing issue in education with respect to the achievement gap (Ladson-Billings 2006) and graduation rates (Wald and Losen 2007). A great deal of research has demonstrated that racial and socioeconomic disparities exist in the use of suspension and expulsion. A recent study considered racial and ethnic disparities in school discipline at the elementary and middle school level (Skiba et al. 2011).

The data for this study were based on records generated by a school-wide information system, which was being used in over 4,000 schools across the United States during the 2005–2006 academic year. The information system is an electronic platform that can be used to process educational interventions. From these 4,000 schools the researchers identified schools that used the system for the full 2005–2006 academic year, reported ethnicity information, had grade levels between kindergarten and sixth grade or sixth and ninth grade, and had agreed to share anonymous summary data. This resulted in data from 436 schools with a total enrollment of over 120 thousand elementary school students and over 60 thousand middle school students.

The frequency distribution of the number of students referred, the total enrollment, and the race or ethnicity of students for elementary school students is given in Table 10.19. From this table it can be observed that while African American students make up about 25% of the student enrollment population, they comprise 35% of the office referrals for discipline. This can be compared to white students, who make up 45% of the student enrollment population, but account for only 35% of the office referrals for discipline. These frequency distributions clearly show that African American students make up for more than what would be expected with respect to referrals for discipline.

Table 10.19 Frequency distributions of the number of students referred, the total enrollment, and race or ethnicity of students in a school-wide information system study for elementary school students.

Race or Ethnicity	Referred	Enrollment
Hispanic	4311	25021
	0.131	0.209
	13.1%	20.9%
African American	11577	30961
	0.353	0.258
	35.3%	25.8%
White	11703	54690
	0.357	0.455
	35.7%	45.5%
All Other	5212	9466
	0.159	0.079
	15.9%	7.9%
Total	32803	120148
	1.000	1.001
	100%	100%

This page titled 10.6.2: Race and School Discipline is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.7: Summary

Qualitative data are often summarized using a frequency distribution, which is a table that contains the number, proportion, and percentage of time each possible value of the qualitative variable occurs. Such a table is often the only numerical summary available for qualitative data as such data usually have a nominal or ordinal measurement scale and no mathematical calculations are generally valid with such data. The frequency distribution can be used to determine which values of the variable occur more often than the others and which values are relatively rare.

Frequency distributions can also be constructed for quantitative data, though some additional preparation is required. Unless the data takes on only a few values, classes must be constructed, and the number of times the observations fall within the classes is used to compute the frequencies, relative frequencies, and percentages for the frequency distribution. Classes should each have the same size and should be located so that they show the true structure of the data.

To investigate possible associations between variables, cross-classified frequency distributions can be constructed. These frequency distributions show how often each pair of values from two different variables occur in the data. These frequency distributions are interpreted the same way as other frequency distributions but also provide information about the joint behavior of the two variables. When considering cross-classified frequency distributions, the possibility that trends are being covered up can be a major difficulty.

This page titled 10.7: Summary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.8: About the Quotation

Stephen Jay Gould (1941–2002) was an American paleontologist, evolutionary biologist, and historian of science. He spent most of his career teaching at Harvard University and working at the American Museum of Natural History in New York. Gould was known mainly for his essays in *Natural History* and his books.

This page titled 10.8: About the Quotation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.9: Practice and Exploration

1. The data given below are the result of rolling a six-sided die fifty times.

1	1	2	6	2	2	1	2	4	1
5	5	5	6	3	1	6	5	5	5
4	4	1	6	2	3	5	3	2	3
2	4	3	3	3	5	5	6	4	3
6	3	6	5	2	5	4	6	6	5

Construct a frequency distribution for the data including the relative frequencies and the percentages. Interpret your results. Are there faces of the die that turn up more than others? Does the die seem fair to you?

2. The data given below are the result of rolling a six-sided die fifty times.

1	5	5	6	2	2	1	2	4	5
5	5	5	6	3	5	6	5	5	5
4	4	5	6	5	5	5	3	2	3
5	4	5	3	5	5	5	6	4	5
6	3	5	5	2	5	4	6	6	5

Construct a frequency distribution for the data including the relative frequencies and the percentages. Interpret your results. Are there faces of the die that turn up more than others? Does the die seem fair to you?

3. The table below shows a frequency distribution based on 60 observations of the gender identities of students who took part in a survey about campus life at a large university. The table has been left incomplete with a question mark denoting the missing entries. Using the entries of the table that have been provided, fill in the missing entries.

Identity	Relative Frequency	Frequency	Percent
Agender	2	?	3.33%
Cisgender	35	0.5833	?
Genderfluid	7	?	?
Genderqueer	?	0.0833	?
Intersex	1	?	?
Nonconforming	?	?	?
Transgender	5	?	?
Total	60	?	100

4. In a study of student debt, the total amount borrowed for undergraduate education and race and ethnicity were studied for 2015–2016 graduates. Frequency distributions of debt by race for students earning a master's degree are given in the table below. This table only includes the percentages. What general trends and conclusions do you make from these frequency distributions?

	Amount Borrowed				
Race or Ethnicity	No Debt	$1 to $24,999	$25,000 to $49,999	$50,000 to $74,999	$75,000 or more
White	43%	22%	19%	10%	7%
African American	19%	18%	26%	22%	16%
Hispanic	27%	24%	24%	16%	8%
Asian	50%	11%	14%	12%	13%

5. The table below shows a frequency distribution for the number of books checked out per week by 250 randomly sampled patrons who checked out at least one book at an inner-city library.

Number	Frequency	Relative Frequency	Percent
1	15	0.060	6.0%
2	39	0.156	15.6%
3	64	0.256	25.6%
4	50	0.200	20.0%
5	26	0.104	10.4%
Over 5	56	0.224	22.4%
Total	60	1.000	100%

What trends do you see in the table, and what types of conclusions can you make about the number of books checked out buy these individuals

?- Why do you think the last class was defined as it was? Does this hurt the interpretation of the frequency distribution?

6. In a famous study of potential gender bias (Bickel et al. 1975), a cross-classified frequency distribution of data on gender and admission to the Berkeley graduate school for the fall 1973 semester are given below:

Gender	Admit	Deny	Total
Female	1494	2827	4321
Male	3738	4704	8442
Total	5232	7531	12763

Compute the relative frequencies and percentages for this frequency distribution. What percentage of female applicants are admitted and how does this compare to the percentage of male applicants that are admitted? What conclusions can be drawn from this table?

7. Continuing with the study on potential gender bias from the previous problem, the frequency table has been split up across the six largest majors that the female applicants applied to below:

Major	Admit	Deny	Total
A	89	19	108
B	17	8	25
C	202	391	593
D	131	244	375
E	94	299	393
F	24	317	341
Total	557	1278	1835

and for the male applicants below:

Major	Admit	Deny	Total
A	512	313	825
B	353	207	560
C	120	205	325
D	138	279	417
E	53	138	191
F	22	351	373
Total	1198	1493	2691

Specific majors are not identified due to university policy (Freedman et al. 2007). Compute the relative frequencies and percentages for this frequency distributions. Looking at the trends in the frequency tables, does there seem to be gender bias in the admittance of students?

8. The table below shows the frequency distributions for the number of students referred, the total enrollment, and race or ethnicity of students in the school-wide information system study for middle school students. This is additional data for the study considered earlier.

Race or Ethnicity	Referred	Enrollment
Hispanic	4,245	10,332
	0.169	0.171
	16.9%	17.1%
African American	8,024	12,228
	0.320	0.219
	32.0%	21.9%
White	9,542	32,975
	0.381	0.545
	38.1%	54.5%
All Other	3,260	3,978
	0.130	0.066
	13.0%	6.6%
Total	25,071	60,522
	1.000	1.001
	100%	100%

What general trends and conclusions do you make from these frequency distributions?

9. In a study of unequal health service outcomes, researchers at a large hospital system gathered data on patients from the previous year that had reported trouble with alcoholism. The patients for the study were chosen based on the severity of their condition. That is, all the patients in the group should have received roughly the same diagnosis and treatment options from their physician. The cross classified frequency distribution in the table below classifies each of the patients by race and ethnicity and the type of treatment the attending physician suggested to the patient.

Race	Treatment Option			
Or Ethnicity	None	Outpatient	Inpatient	Total
African American	1001	631	315	1947
Asian	53	122	285	460
Hispanic	1507	732	744	2983
Native American	112	58	52	222
White	983	1203	2096	4237
Total	3611	2746	3492	9849

Compute the relative frequencies and percentages for this frequency distribution. What conclusions can be drawn from this table?

This page titled 10.9: Practice and Exploration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.10: Collaborative Activity

For this activity you will be supplied with three four-sided dice, two six-sided dice, and one twelve-sided die. This activity will consider how the frequency distributions of the sum of the four-sided dice, the sum of the six-sided dice, and the results of a throw of the twelve-sided die. To begin the activity, one member of the group should be designated as the dice roller and one member of the group should be designated as the data collector.

The four-sided dice will be used first. The dice roller will roll all three dice fifty times. For each roll the sum of the three dice will be observed, and the result will be recorded in a data table like the one shown in Table 10.20. As an example, consider fifty rolls of three four-sided die. Suppose that the first roll of the three four-sided dice that we observe a 2, 4, and 1. The sum of these is 7, and this is recorded in the data table as shown in Table 10.21. Suppose that the second roll of the three four-sided dice results in 4, 4, and 3. The sum is 11, and this is recorded in the data table as shown in Table 10.21. The remaining rolls are recorded in a similar manner. Example rolls are shown in Table 10.21.

Table 10.20 Data table for the collaborative activity.

Sum of Three Four-Sided Dice
Sum of Two Six-Sided Dice
One Twelve-Sided Die

Table 10.21 Data table for the collaborative activity.

Sum of Three Four-Sided Dice									
7	11	10	6	10	8	6	6	7	7
9	8	5	7	10	11	6	5	10	9
8	7	9	6	8	12	8	11	6	4
5	7	8	10	4	9	10	9	7	6
6	6	6	9	7	8	11	10	6	10
Sum of Two Six-Sided Dice									
6	3	5	5	9	3	6	7	6	12
11	8	8	7	5	5	7	9	7	11
5	7	8	6	8	8	7	6	3	7
11	9	8	8	11	8	7	10	7	4
10	5	5	6	8	6	6	11	5	11
One Twelve-Sided Die									
11	4	4	6	7	11	10	6	10	4
8	6	11	11	5	11	11	3	4	1
8	7	9	5	10	12	11	9	5	6
12	12	10	1	9	2	5	5	3	3
7	3	6	7	1	10	2	10	10	7

Next, the set of two six-sided dice are used. The dice roller will roll both dice fifty times. For each roll the sum of the two dice and each individual roll will be observed. The sum will be recorded in a data table like the one shown in Table 10.20, and the individual rolls will be recorded in a table like the one shown in Table 10.22. In this table the smaller of the rolls will be recorded first and the larger of the two rolls will be recorded second. As an example, consider fifty rolls of two six-sided dice. Suppose that the first roll of the two six-sided dice results in 2 and 4. The sum is 6, and this is recorded in the data table as shown in Table 10.21. In the second data table we will record 2 as the smaller of the two rolls and 4 as the larger of the two rolls, as shown in the example data recorded in Table 10.23. Suppose that the second roll of the two six-sided dice results in 1 and 2. The sum is 3, and this is recorded in the data table as shown in Table 10.21. Similarly, the smaller and larger numbers are recorded in Table 10.23. The remaining rolls are recorded in a similar manner. Example rolls are shown in Tables 10.21 and 10.23. Finally, the twelve-sided die is used. The dice roller will roll the die fifty times. For each roll, the outcome will be observed and the result will be recorded in a data (see Table 10.21).

Table 10.22 Data table for the collaborative activity.

Minimum of Two Six-Sided Dice									
Maximum of Two Six-Sided Dice									

Table 10.23 Example data table for the collaborative activity.

Minimum of Two Six-Sided Dice									
1	1	1	2	4	1	1	3	1	6
5	4	3	3	2	1	3	3	2	5
2	1	4	1	3	2	3	1	1	2
5	3	2	2	5	3	1	4	3	1
5	2	2	2	2	3	2	5	2	5
Maximum of Two Six-Sided Dice									
5	2	4	3	5	2	5	4	5	6
6	4	5	4	3	4	4	6	5	6
3	6	4	5	5	6	4	5	2	5
6	6	6	6	6	5	6	6	4	3
5	3	3	4	6	3	4	6	3	6

Once the data have been collected, you and your group will need to construct four frequency distributions. The first three frequency distributions correspond to the sums of the three four-sided dice, the two six-sided dice, and the twelve-sided die. Each possible sum with correspond to a class for each of these frequency tables.

Considering the example data, we begin with the sums of three four-sided dice. When rolling three four-sided dice, the possible sums of the dice can be equal to 3, 4, 5, 6, 7, 8, 9, 10, 11, or 12. Therefore we construct a frequency distribution with these classes and count the number of times each sum was observed in our data. See Table 10.24. For the first entry in the table, we note that no sums added up to 3, so the frequency is 0, along with the relative frequency and the percentage. The next possible sum is 4, and we observe two instances where a sum of 4 was observed, so the frequency is 4, and the relative frequency is $4 \div 50 = 0.08$ or 8% as indicated in Table 10.24. The remaining entries in Table 10.24 are computed in a similar manner. The frequency distributions for the sum of the two six-sided die and the twelve sided die are also constructed the same way, with the frequency distributions for the example data given in Tables 10.25 and 10.26.

Table 10.24 Example frequency distribution for the sum of three four-sided dice.

Sum	Frequency	Relative Frequency	Percentage
3	0	0.00	0%
4	2	0.04	4%
5	3	0.06	6%
6	11	0.22	22%
7	8	0.16	16%
8	7	0.14	14%
9	6	0.12	12%
10	8	0.16	16%
11	4	0.08	8%
12	1	0.02	2%
	50	1.00	100%

Table 10.25 Example frequency distribution for the sum of two six-sided dice.

Sum	Frequency	Relative Frequency	Percentage
2	0	0.00	0%
3	3	0.06	6%
4	1	0.02	2%
5	8	0.16	16%
6	8	0.16	16%
7	9	0.18	18%
8	9	0.18	18%
9	3	0.06	6%
10	2	0.04	4%
11	6	0.12	12%
12	1	0.02	2%
	50	1.00	100%

Table 10.26 Example frequency distribution for the sum of two six-sided dice.

Sum	Frequency	Relative Frequency	Percentage
1	3	0.06	6%
2	2	0.04	4%
3	4	0.08	8%
4	4	0.08	8%
5	5	0.10	10%
6	5	0.10	10%
7	5	0.10	10%
8	2	0.04	4%
9	3	0.06	6%
10	7	0.14	14%
11	7	0.14	14%
12	3	0.06	6%
	50	1.00	100%

Finally, we will construct a cross-classified frequency table for the minimum and the maximum of the two six-sided die rolls reported in Table 10.23. To construct this table, note that the possible values for both the minimum and maximum of the two dice are 1, 2, 3, 4, 5, and 6. Therefore, construct a cross classified table like the one shown in Table 10.27 that shows the results for the example data. Once the blank table is constructed, you will need to count how often each possible cross-classified minimum and maximum occurs. Remember, these observations correspond to the same roll of the two dice, so you will need to match the entry in the top part of Table 10.23 with the corresponding entry in the bottom part of Table 10.23. For the example data, there were no instances where both the minimum and the maximum of the two dice were 1, and therefore this entry is 0, with a relative frequency of 0.00 and a percentage of 0%. The next entry corresponds to the case where the minimum of the two rolls was 1 and the maximum of the two rolls was 2. In the example data, there are three instances where this occurs, and hence the corresponding frequency recorded on the table is 3, with a relative frequency of $3 \div 50 = 0.06$, or 6%. The remaining entries in the table are computed in a similar manner.

Table 10.27 Example cross classified frequency table for the minimum and maximum of the rolls of the two six-sided dice for the collaborative activity.

		Maximum						
Minimum		1	2	3	4	5	6	Total
	1	0 0.00 0%	3 0.06 6%	1 0.02 2%	2 0.04 4%	5 0.10 10%	2 0.04 4%	13 0.26 26%
	2	0 0.00 0%	0 0.00 0%	6 0.12 12%	2 0.04 4%	2 0.04 4%	4 0.08 8%	14 0.28 28%
	3	0 0.00 0%	0 0.00 0%	1 0.02 2%	5 0.10 10%	3 0.06 6%	2 0.04 4%	11 0.22 22%
	4	0 0.00 0%	0 0.00 0%	0 0.00 0%	2 0.04 4%	1 0.02 2%	1 0.02 2%	4 0.08 8%
	5	0 0.00 0%	0 0.00 0%	0 0.00 0%	0 0.00 0%	1 0.02 2%	6 0.12 12%	7 0.14 14%
	6	0 0.00 0%	0 0.00 0%	0 0.00 0%	0 0.00 0%	0 0.00 0%	1 0.02 2%	1 0.02 2%
	Total	0 0.00 0%	3 0.06 6%	8 0.16 16%	11 0.22 22%	12 0.24 24%	16 0.32 32%	50 1.00 100%

Questions

1. Compare the frequency tables for the rolls of three four-sided dice, two six-sided dice, and the twelve-sided die. What differences do you observe? Are there areas where more of the observations are concentrated, or are the frequencies evenly distributed throughout the classes?
2. Consider the cross-classified frequency table for the minimum and the maximum of the rolls of the two six-sided dice. What patterns do you observe when you look at the cross-classified frequencies?
3. Suppose that someone tells you that they had an observation of 6, and that they don't remember what dice it came from. If you had to guess based on what you observed in your frequency distributions, would you conclude that the observation was from the three four-sided dice, the two six-sided dice, or the single twelve-sided die?
4. Similarly, suppose that someone tells you that they had an observation of 3, and that they don't remember what dice it came from. What would you guess based on your observed frequency distributions?
5. Suppose that someone tells you that they had an observation of 1, and they forget what dice they rolled. Would you conclude that the observation was from the three four-sided dice, the two six-sided dice, or the single twelve-sided die?

This page titled 10.10: Collaborative Activity is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

10.11: References

Bae, SungA, So Ree Kim, Mi-Na Kim, Wan Joo Shim, and Seong-Mi Park. 2021. “Impact of Cardiovascular Disease and Risk Factors on Fatal Outcomes in Patients with Covid-19 According to Age: A Systematic Review and Meta-Analysis.” *Heart* 107, no. 5, 373–380. https://doi.org/10.1136/heartjnl-2020-317901

Baum, S. 2019. “Student Debt: The Unique Circumstances of African American Students. Race and Ethnicity in Higher Education: A Status Report. *Race and Ethnicity in Higher Education*, 201–210. https://www.equityinhighered.org/wp-...apter-8-SA.pdf

Bickel, P. J., E. A. Hammel, and J. W. O’Connell. 1975. “Sex Bias in Graduate Admissions: Data from Berkeley.” *Science* 187, 398–404. https://doi.org/10.1126/science.187.4175.398

Espinosa, Lorelle L., Jonathan M. Turk, Morgan Taylor, and Hollie M. Chessman. 2019. *Race and Ethnicity in Higher Education: A Status Report.* American Council on Education. https://www.equityinhighered.org/res...status-report/

Freedman, David, Robert Pisani, and Roger Purves. 2007. *Statistics*. Norton Publishing. https://wwnorton.com/books/Statistic...roduct-details

Ladson-Billings, G. 2006. “From the Achievement Gap to the Education Debt: Understanding Achievement in U.S. Schools.” *Educational Researcher* 35, no. 7, 3–12. https://doi.org/10.3102/0013189X035007003

Radelet, M. L. 1981. “Racial Characteristics and the Imposition of the Death Penalty.” *American Sociological Review* 46, no. 6, 918–927.

Skiba, Russell J., Robert H. Horner, Choong-Geun Chung, M. Karega Rausch, Seth L. May, and Tary Tobin. 2011. “Race is Not Neutral: A National Investigation of African American and Latino Disproportionality in School Discipline.” *School Psychology Review* 40, no. 1, 85–107. https://doi.org/10.1080/02796015.2011.12087730

Wald, J., and D. J. Losen. 2007. “Out of Sight: The Journey Through the School-to-Prison Pipeline.” In S. Books (Ed.), *Invisible Children in the Society and its Schools* (3rd ed., pp. 23–37). Lawrence Erlbaum Associates Publishers. https://www.taylorfrancis.com/chapte...d-daniel-losen

This page titled 10.11: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

CHAPTER OVERVIEW

11: Visual Methods for Summarizing Data

“Numerical quantities focus on expected values, graphical summaries on unexpected values.”

–John Tukey

11.1: What to Expect
11.2: Bar Graphs
11.3: Pie Charts
11.4: Line Graphs
11.5: Histograms
11.6: Scatterplots
11.7: Critical Analysis
11.7.1: Sampling Animals in Populations
11.8: Summary
11.9: About the Quotation
11.10: Practice and Exploration
11.11: References

This page titled 11: Visual Methods for Summarizing Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

11.1: What to Expect

The numerical methods for summarizing data described in Chapter 9 provide very efficient methods for describing very specific characteristics such as location and variation of observed data. In Chapter 10, we considered a more general type of numerical summary that provides information about where data are concentrated but can also be used to get very general ideas about location and variation. In this chapter, we will consider methods for summarizing data that are visual. The appeal of visualizing data may be obvious to many, and one is tempted to insert the obligatory quote “a picture is worth a thousand words,” which seems to have been first used in a 1918 newspaper advertisement for the San Antonio Light. Visual summaries have the advantage that they can quickly and concisely give a visual impression of the overall structure of the data.

This chapter will consider many of the popular methods for summarizing data including frequency plots, histograms, and scatterplots.

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

1. Construct and interpret bar plots for one and two qualitative variables.
2. Interpret pie charts.
3. Construct and interpret line graphs.
4. Construct and interpret histograms for quantitative data.
5. Construct histograms when provided with class intervals.
6. Construct and interpret scatterplots.

This page titled 11.1: What to Expect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

11.2: Bar Graphs

In the previous chapter we found that qualitative data could be summarized using frequency distributions that reported frequencies, the number of observations in each category, relative frequencies, the proportion of observations in each category, and percentages, the percentage of values for each category of dataset. A bar graph takes the information presented in a frequency distribution and presents it visually.

Definition: Bar Graph

A *bar graph* of a frequency distribution is a graph whose horizontal axis shows the category headings of a qualitative variable. The vertical axis shows the observed frequency, relative frequency, percentage, or numerical summary (mean, median, etc.) for each category, which correspond to the bar heights. The bars are separated by spaces.

We begin by considering the situation where we have a single categorical variable. To motivate the discussion, consider the data on race and debt from Table 10.2. Earlier we had computed the mean, median, and standard deviation of the amount of debt for each category of race. First consider constructing a bar graph for the means for each category of race. The horizontal axis of this graph will have equally spaced areas for each of the categories of race. The vertical axis corresponds to the observed mean debt for the categories. This axis is numerical and should start at 0 and should end near the largest mean debt amount. The mean debt values for the categories are 26.9 (African American), 7.6 (Asian), 14.5 (Hispanic), and 21.9 (white). Since the largest mean debt is near 30, we will start the vertical axis at 0 and end the range at 30, with marks at 0, 5, 10, 15, 20, 25, and 30. The bar graph in Figure **11.2.1** provides a quick visual summary of the mean values. Not only can one quickly observe that African Americans have the largest mean debt, and that Asians have significantly lower debt, but the relative sizes of the means can be easily observed.

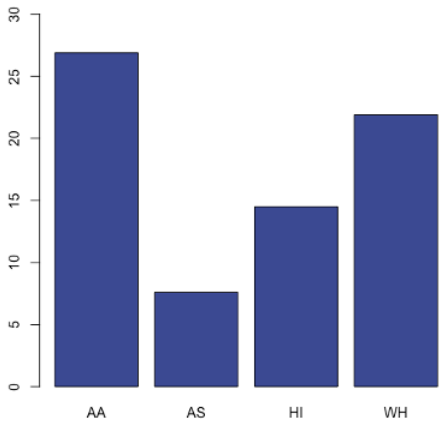


Figure **11.2.1**: Bar graph of mean debt by race information given in Table 10.2 (Public domain figure created by Alan M. Polansky).

As we discussed earlier, the mean is not a robust measure of location, meaning it can be skewed by outliers. Therefore, since debt can be extremely high and only as low as 0, we consider the median debt as well. A bar graph of the median debt for each race category is plotted in Figure **11.2.2**. In this plot the overall general trend of debt by race appears the same here with Asians having the lowest debt and African Americans having the highest.

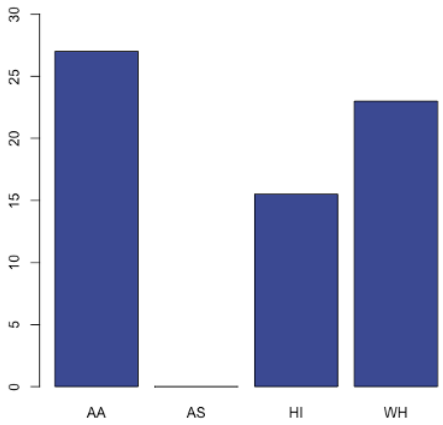


Figure **11.2.2**: Bar graph of median debt by race information given in Table 10.2 (Public domain figure created by Alan M. Polansky).

When constructing the horizontal axis of a bar graph, the order of the categories must be specified. If the categories are measured on a scale that is at least an ordinal, the categories should be arranged in the order that is specified by their structure. For example, if the categories are based on a Likert scale of 1 to 5, then the categories on the horizontal axis should respect that ordering. When the measurement scale of the categories is nominal, there is no natural ordering to the categories and one ordering is as valid as the next. When observing bar graph, one should be careful not to over interpret trends in the bars when the categories are based on a nominal measurement scale. As an example, consider the bar graph shown in Figure **11.2.3**, which is based on the same data used to construct the bar graph in Figure **11.2.2**. The bar graph in Figure **11.2.3** has the categories on the horizontal axis arranged in a different order. Note that there appears to be an increasing trend in the bars. While the bars reflect the correct levels for the medians for the different categories, the idea of a trend is illusory as the ordering of the categories for this data have no meaning.

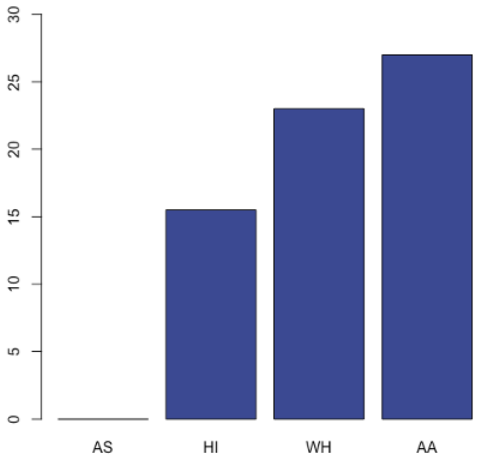


Figure **11.2.3**: Bar graph of median debt by race information given in Table 10.2 (Public domain figure created by Alan M. Polansky).

Unfortunately, visual representations of data can be manipulated in several ways to make some trends appear more apparent than the data indicates or to cover up other trends in the data. One common methods is manipulating the order of the categories when they are measured on a nominal measurement scale. For example, a low value can be masked somewhat if it is surrounded by other smaller values, or it can be emphasized by purposely putting the low category next to the highest category. The horizontal axis can also be manipulated by not starting the lower end of the range at zero. This can be used to overemphasize small differences.

Bar graphs can also be used to show more than one measure for each category of a variable. Once again, we will consider the data on race and debt from Table 10.2. We computed two measures of location for debt in each category, the mean and the median. Both measures can be represented in a side-by-side bar graph as is shown in Figure **11.2.4**. The horizontal and vertical axes are constructed in the same manner as before so that the largest of the two measures can be represented. The plot is similar in construction to the previous. Separate bar graph of the mean and the median debt. The difference with the graph in Figure **11.2.4** is that the bars representing the mean and the median debt for each category are plotted next to each other within the category. Note that because both measurements are within the same category, there is no space between mean and median debt. From the plot we can observe that except those individuals classified as Asian, there is minimal difference between the mean and the median of the observed debts. An alternative method for conveying the same information is to group the measures of location on the horizontal axis, essentially combining Figures **11.2.1** and **11.2.2** into a single plot. See Figure **11.2.5**. Note that it is possible to also include measures of variation in these plots along with the means and medians. This is strongly discouraged because the two measures provide different types of information.

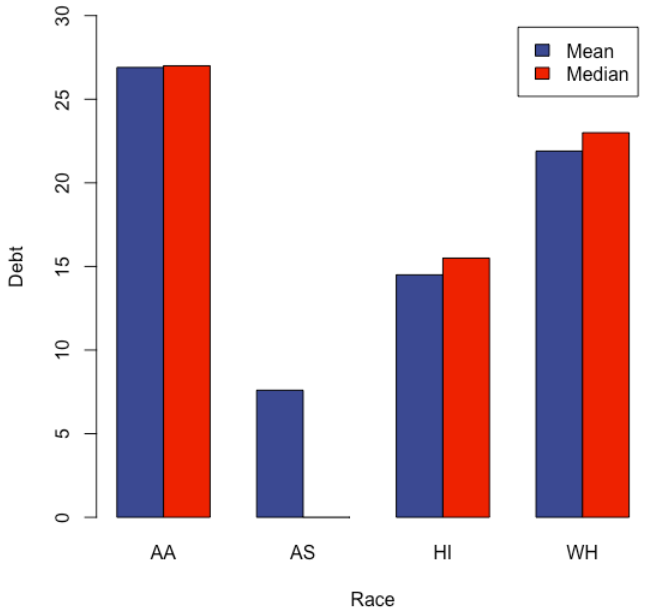


Figure **11.2.4**: Graph of a side-by-side bar graph for mean (blue) and median (orange) debt by race. Information given in Table 10.2 (Public domain image created by Alan M. Polansky).

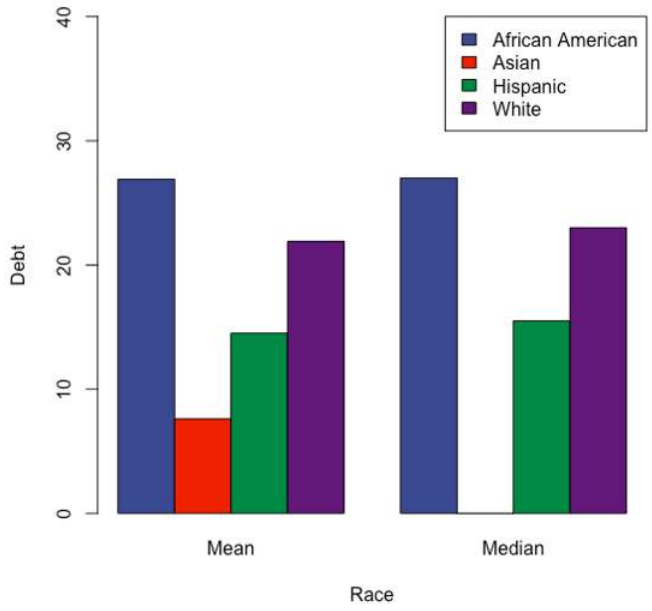


Figure 11.2.5: Graph of a side-by-side bar graph for African American (blue), Asian (red), Hispanic (green), and white (purple). Information given in Table 10.2 (Public domain image created by Alan M. Polansky).

This page titled 11.2: Bar Graphs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

11.3: Pie Charts

Pie charts are a visual method for displaying the frequency distribution of a variable measured on a nominal or ordinal scale. Consider a variable with a corresponding frequency distribution that includes the percentage of observations equal to each of the possible outcomes of the variable. One way to visually represent these percentages is to construct a circular figure that has been divided into pieces where the area of each piece is proportional to the percentage from the frequency distribution for each possible outcome of the variable. The corresponding chart resembles a pie that has been cut into pieces of various sizes.

Definition: Pie Chart

A *pie chart* of a frequency distribution is a circular graph whose "slices" corresponds to the possible outcomes for a qualitative variable. The area of the slice compared to the area of the entire pie is equal to the percentage (or relative frequency) of that response.

Earlier we looked at a survey on gender discrimination. The responses to the question are scored on a five-point Likert scale, and the frequency distribution for the first question is given in Table 11.1. A pie chart of the corresponding frequency distribution is shown in Figure **11.3.1**. In the figure we can observe a circle with slices labeled for each of the five possible observed responses. The chart has been constructed so that the area of each slice, relative to the area of the entire circle, is the same as the ratio of the corresponding frequency of the response relative to the number of responses. For example, from Table 11.1 we can observe that the number of individuals who responded with a 1 to the first question is 5, which is 20% of all the responses to the question. Therefore, when the pie chart was constructed, it was done so in such a way that the area of the slice that is labeled 1 takes up 20% of the entire area of the circle. Note that the areas labeled for responses 1 and 3 are equal. This is because both responses have the same frequency. From a pie chart one can quickly see what responses had high or low frequencies. If the slices are all about the same size, then all the frequencies are very similar. From Figure **11.3.1** we can observe that all the frequencies are similar, but response 5 has a higher frequency than the other responses.

Table 11.1 Frequency distribution for the first question in the survey on gender discrimination.

	Response					
Question	1	2	3	4	5	Total
1	5	4	5	4	7	25
	0.20	0.16	0.20	0.16	0.28	1.00
	20%	16%	20%	16%	28%	100%

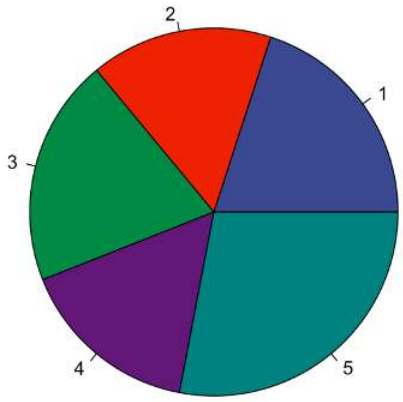


Figure 11.3.1: Pie Chart for Question 1 in Table 10.5. The numbers on the pie chart refer to the responses from the question (Public domain image created by Alan M. Polansky).

In some cases, you will see pie charts that have been enhanced for visual appeal. These enhancements are often unnecessary and, in some cases, can be used to manipulate the visual conclusions that one might make from looking at the chart. For example, some pie charts are given a three-dimensional representation as a disk. The pie chart in Figure 11.3.2: is a three-dimensional pie chart representing the same frequency data as was used to create Figure 11.3.1. Because of the illusion of perspective as well as the thickness of the disk used to represent the pie, the areas for responses 4 and 5 seem overrepresented by the amount of area that the eye perceives in the plot. Consequently, the remaining responses are underemphasized since they are at the rear of the disk.

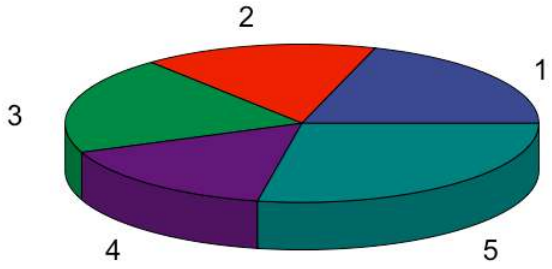


Figure 11.3.2: Three-dimensional pie Chart for Question 1 in Table 10.5. The numbers on the pie chart refer to the responses from the question (Public domain image created by Alan M. Polansky).

Often it is beneficial to list the relative frequencies of percentages for our categories on the pie chart. This is because categories can have an extremely small number of responses, and we would like to know which has the smallest. To do this, you list the category's name (in our case, 1 through 5 for the possible responses), then the percentage or relative frequency associated with that value. This is displayed in Figure **11.3.3**. Notice that the areas remain unchanged. You can also list the percentage in the pie slice itself.

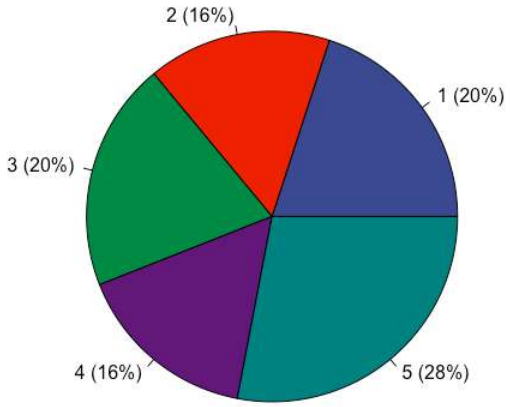


Figure **11.3.3**: Pie Chart for Question 1 in Table 10.5. The numbers on the pie chart refer to the responses from the question. The percentage of each response is also indicated (Public domain image created by Alan M. Polansky).

Pie charts are not typically the best method for displaying frequency information, though they can be useful in some contexts. Bar charts provide a much easier way to compare frequencies as the relative heights of the bars are easy to distinguish as they are plotted next to one another. It is difficult, for example, to determine visually if one pie slice has twice the area of another, whereas it is relatively easy to determine visually that one bar might be twice as high as another.

Pie charts can be used when comparing frequencies that are repeating over many situations or responses. For example, Figure **11.3.4** shows a sequence of four pie charts showing the frequency distributions of the responses for each of the first four questions. Recall that the Likert scale for this survey indicates how much the individual agrees with the question. In considering the sequence of pie charts shown in Figure **11.3.4**, we see that more people strongly disagree with questions 2 and 3 than with questions 1 and 4. Further, we can note that more people strongly agree with questions 1 and 4 than with questions 2 and 3.

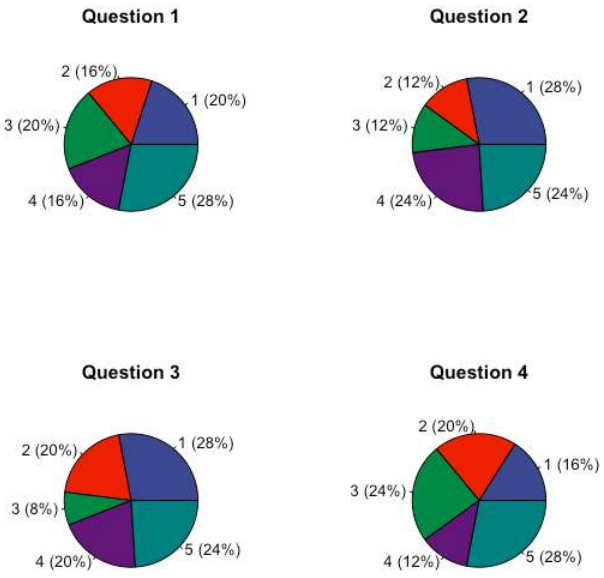


Figure **11.3.4**: Pie Chart for Questions 1–4 in Table 10.5. The numbers on the pie chart refer to the responses from the question. The percentage of each response is also indicated (Public domain image created by Alan M. Polansky).

This page titled 11.3: Pie Charts is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

11.4: Line Graphs

Line graphs can be used to represent a trend in observations with respect to a variable that has at least an ordinal measurement scale. For this chapter we will consider the special case of examples of plotting a trend of an observation that changes with respect to time. For example, you might be interested in visualizing whether there are any trends in the population of a village over a period of many years. Figure **11.4.1** shows the population of the village of Creston, IL from 1880 to 2020. The populations were taken from the United States Decennial Census, and therefore data is only available every ten years. In considering the plot in Figure 11.8, we can observe that the population was stable between 1880 and 1950, with a small dip during World War II. After that, the population increased and has somewhat stabilized once more. Note that the horizontal axis corresponds to time, in this case the years from 1880 to 2020. The vertical axis corresponds to the size of the population. The plot was constructed by plotting the population against each year and then connecting these points with lines.

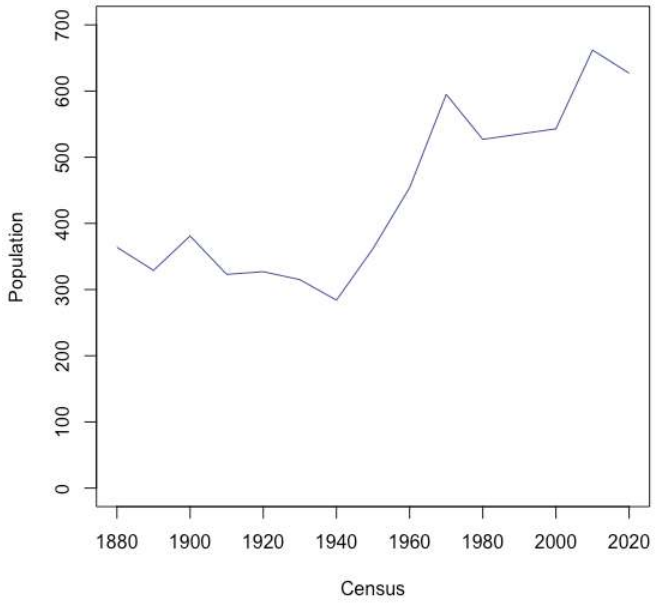


Figure **11.4.1**: The population of Creston, IL based on the U.S. Decennial Census from 1880 to 2020 (Public domain image created by Alan M. Polansky).

Definition: Line Graph

A *line graph* is a plot of points whose horizontal axis is a variable that has at least an ordinal measurement scale and whose vertical axis is an observation or measurement that has at least an ordinal measurement scale. The points are connected by line segments with respect to the order of the horizontal axis.

Line graphs can also be constructed to have more than one variable plotted on the vertical axis, which allows one to compare the trends of two or more variables that have been measured over the same time. Usually, the variables plotted against time are measured with the same units of measurement, otherwise the plot can become difficult to interpret. For example, consider the line graph shown in Figure **11.4.2**. This plot shows a measure of the number of times the phrases “same-sex marriage” and “United States Supreme Court” are searched for in the 40 months starting at February 2012 (Flores and Barclay 2016). It was during this time that two very important issues were to be decided by the United States supreme court: the Affordable Care Act, which would guarantee universal access to health care, and the issue of marriage equality, which considered whether same sex couples had the right to be married. Certain time points have also been highlighted on the plot. The first vertical line, which is a green dashed line, indicates when President Obama announced his support for same-sex marriage, and at that time there is a spike in searches for that topic. The vertical dashed purple line indicates when the Affordable Care Act was upheld by the United States Supreme Court, and

prior to that time there is a large spike in the number of searches for the phrase “supreme court” in anticipation of the ruling. The vertical teal line indicates when oral arguments for same-sex marriage were argued before the Supreme Court, and we can observe that both phrases have a spike in the number of searches at that time. Finally, the vertical red dashed line indicates when the final ruling on same sex marriage was announced, and again we can observe a spike for both phrases at that time. After that time, searches for both phrases decline and stabilize.

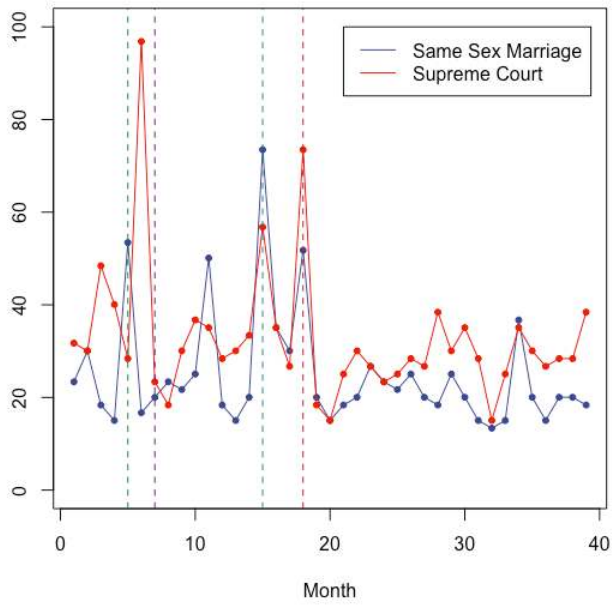


Figure **11.4.2**: Line graph of National trends in Google information searching on “same-sex marriage” (blue) and “United States Supreme Court” (red). The green dashed line indicates when President Obama announced support for same sex marriage. The purple dashed line indicates when the Affordable Care Act was upheld by the Supreme Court. The teal dashed line indicates when oral arguments were heard on same sex marriage, and the red dashed line indicates when the final court ruling was announced. The first month indicated on the plot corresponds to February 2012. The plot is adapted from information contained in (Flores and Barclay 2016). (Public domain image created by Alan M. Polansky).

Figure **11.4.3** exhibits another example of a line graph that shows trends for multiple variables. In this graph we have plotted the median income for individuals of employable age for different racial groups based on data from Wilson (2020). The horizontal axis indicates yearly data from 2000 to 2020. A first glance shows that there is a difference between the median incomes for different races, and that this difference is consistent throughout the years with Asian individuals having the highest median income followed by white individuals. African American and Hispanic individuals have the lowest median incomes, and the two are comparable. We can also observe how the median incomes change over time. For white, African American, and Hispanic individuals, the median income is stable, and perhaps slightly decreasing until around 2014 when they begin to increase. The trend for Asian individuals is somewhat different. The median income for Asian individuals increases and peaks around 2005 and then has a decrease, followed by an increase starting around 2015. We can observe from the plot that the relative increase is greater for Asian individuals near the end of the graph.

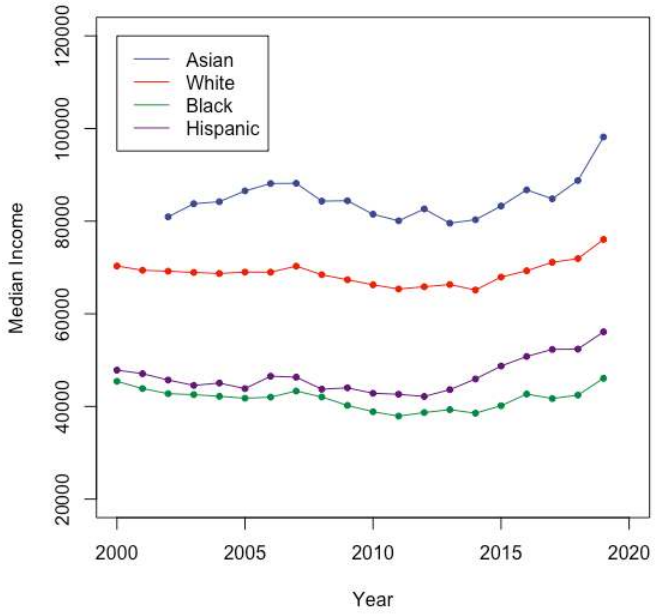


Figure **11.4.3**: Line graph of income by race from 2000–2019 based on data obtained from Wilson (2020) (Public domain image created by Alan M. Polansky)

Importantly, notice in Figure **11.4.3** is that the horizontal axis does not start at zero but at $20,000. The effect of this plot decision is to emphasize the differences between the median incomes of the racial groups. While it is not uncommon for plots of this type not to start an axis at zero, one should be aware that using different ranges for the axes can influence how differences and trends are perceived in the graph. Figure **11.4.4** shows the same data, but the vertical axis now starts at $0 instead of $20,000 and ends at $200,000 instead of $120,000. Note that the trends in the graph have been flattened out, so the differences between the lines appear less significant that in Figure **11.4.3**.

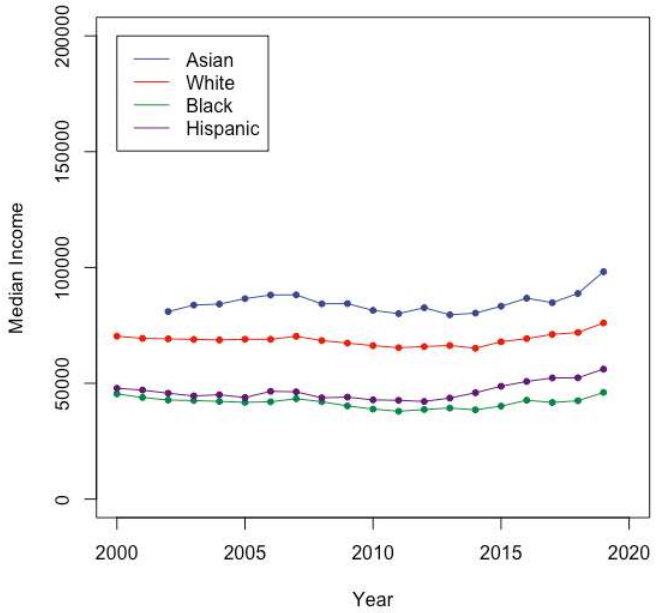


Figure **11.4.4**: Line graph of income by race from 2000–2019 based on data obtained from Wilson (2020) using an expanded vertical axis (Public domain image created by Alan M. Polansky)

This page titled 11.4: Line Graphs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

11.5: Histograms

Frequency distributions for interval or ratio data are based on defining ranges called classes and computing frequencies, relative frequencies, and percentages of the observed data points falling within these classes. The graphical version of these frequency distributions are called histograms and are constructed in a very similar way as the frequency distributions for nominal and ordinal data, except the ranges of the classes are represented directly on the horizontal axis. This means that there are no spaces between the bars in the graph.

Definition: Histogram

A *histogram* of a frequency distribution is a graph whose horizontal axis corresponds to the values of a quantitative variable and whose vertical axis corresponds to the observed frequency, relative frequency, or percentage for each class. The bars of the graph span the corresponding classes.

Let us consider constructing a histogram for a small set of data. Suppose that a small survey was conducted in which each respondent was asked how many times, in the previous two-week period, they had eaten at a fast-food restaurant. The data were observed to be 0, 2, 1, 5, 2, 2, 3, 4, 1, 2, 7, 1, 3, 4, 1, 0, 1, 4, 2, 1, 3, 3, 2, 1, 9, and 1. Based on this data, we will consider the class intervals defined by $0 \leq \text{fast food} < 1$, $1 \leq \text{fast food} < 2$, $2 \leq\leq \text{fast food} < 3$, $3 \leq \text{fast food} < 4$, $4 \leq \text{fast food} < 5$, $5 \leq \text{fast food} < 6$, $6 \leq \text{fast food} < 7$, $7 \leq \text{fast food} < 8$, $8 \leq\leq \text{fast food} < 9$, and $9 \leq \text{fast food} < 10$. The frequency distribution for this data is given in Table 11.2. Note for example, that there are two instances where an individual reported that they had not had fast food over the two-week period so that the frequency for the class 0≤ fast-food <1 is 2. Similarly, there are eight instances where an individual reported that they had fast food once over the two-week period, so that the frequency for the class 1≤ fast-food <2 is 8. The remaining frequencies are computed in a similar manner.

Table 11.2 Frequency distributions of number of times eating at a fast-food restaurant.

Number	0	1	2	3	4	5	6	7	8	9
Frequency	2	8	6	4	3	1	0	1	0	1

Using this information, we can plot the histogram in Figure 11.5.1. For the people who responded they did not eat at a fast-food restaurant last week, the height of the rectangle is 2, corresponding to the frequency for that class. There were 8 people who reported that they ate at a fast-food restaurant once, and the height of the corresponding rectangle for this class is 8. The remaining rectangles are plotted in a similar manner. Note that nobody responded with 6, so that the frequency for that class is 0, and the corresponding rectangle is not plotted leaving a gap in the graph.

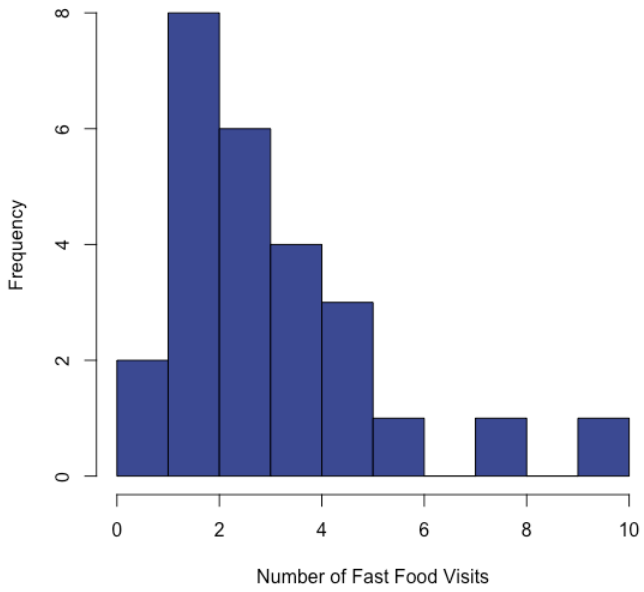


Figure **11.5.1**: Number of times people said they ate at fast-food restaurants in the last week (Public domain image created by Alan M. Polansky)

When looking for trends in histograms, it is most important to look for the general shape of that the rectangle's form. Of particular interest is areas where there are peaks in the shape of the histogram as these areas have a high concentration of data. A peak in a histogram is called a mode.

Definition: Mode

A peak on a histogram is called a *mode*.

Histograms are classified by the number of modes they have.

Definition: Modality

The *modality* of a histogram refers to how many "peaks" the histogram has. A histogram is called unimodal if there is one peak, bimodal if there are two peaks, multimodal if there are three or more peaks, and uniform if the histogram is relatively flat across the classes.

The histogram for the fast-food data shown in Figure **11.5.1** is unimodal. It is tempting to conclude that there are three peaks in this histogram. It is important that when analyzing histograms that the general trend of the data is considered, and not small details like to two small peaks on the right-hand side of Figure **11.5.1**. These small peaks are the result of a few data points, whereas the general shape of the graph is formed by most of the data. It is much safer to draw conclusions from the features that are a result of most of the data rather than relying on a few data points.

Figure **11.5.2** shows a histogram based on data that have been simulated to represent the amount of time two groups have a headache. One group was given a new headache medication while the second group was given a placebo. As we can see in Figure **11.5.2**, there are two modes. One for the length of time it takes for a headache to go away for those taking the new medication, and one for the length of time for a headache to go away for those taking a placebo. Therefore, the histogram shown in Figure **11.5.2** is bimodal.

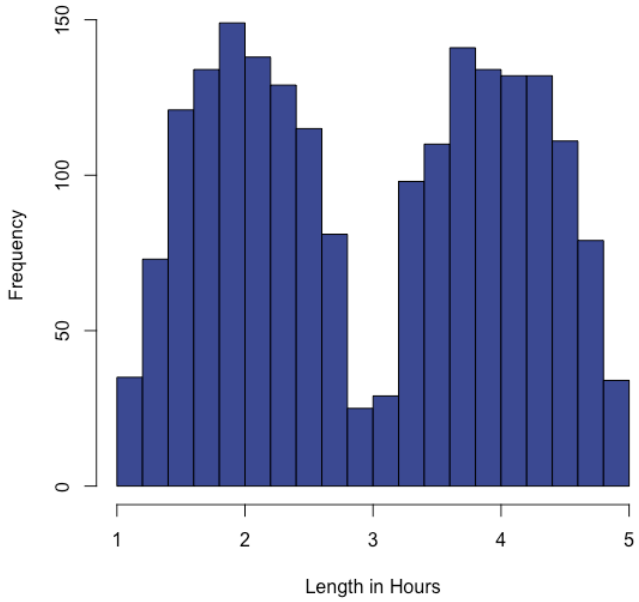


Figure 11.5.2: Histogram of simulated headache lengths for those taking a pain reliever, and those who took a placebo (Public domain image created by Alan M. Polansky)

Definition: Tail

In a unimodal histogram the *tail* of the histogram is the range on each side of the of the peak corresponding to the thinner, tapering end of the histogram where data points are less frequent.

While the definition of the tail of a histogram is somewhat vague, knowing precisely where the tail of a histogram begins is rarely required for judging the overall shape of a histogram. Rather it is the relative length and weight of the tails that is important in determining the shape characteristics of a histogram. The most important of these is the concept of symmetry, where the right and left tails of a histogram are roughly the same size. When the tails differ, the shape is described by comparing the length of the tails resulting in what is called skewness.

Definition: Skewness

The *skewness* of a unimodal histogram indicates whether there is a longer tail to the right of the peak, to the left of the peak, or if the tails are about equal on either side of the peak. If the longer tail is to the left, the histogram is left skewed. If the longer tail is to the right, then the histogram is right skewed. If the tails are about equal, the histogram is symmetric.

For the histogram given in Figure 11.5.1, which is based on the data of the number of times individuals visited a fast-food restaurant in the past two weeks, we can see that the tail on the right side of the peak is longer than on the left. Therefore, this histogram is skewed to the right. Next consider the income simulation histogram in Figure 11.5.3. From Figure 11.5.3 we again observe that the histogram is skewed to the right. This is consistent with what we know about income. While most workers will earn a salary within a certain range, some workers will earn significantly more, and a few workers will earn even more than that. Finally, in Figure 11.5.4 we have the histogram of the simulated student debts. In this case, the left tail is slightly longer than the right and we would conclude that this histogram is skewed to the left.

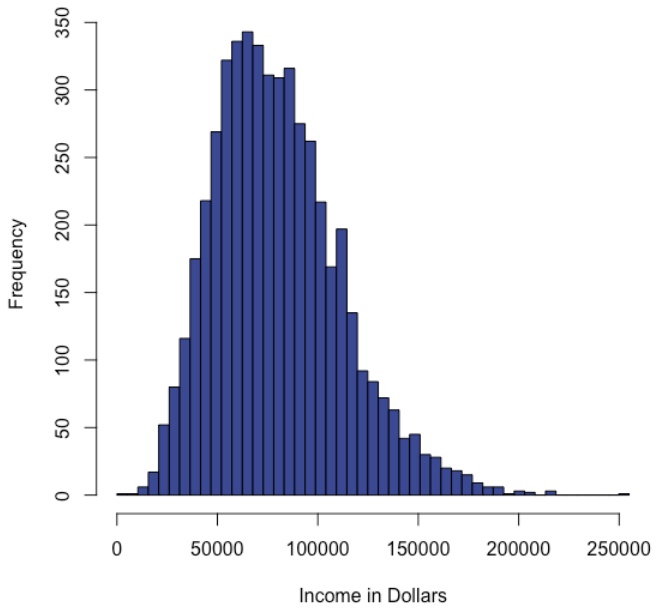


Figure **11.5.3**: Histogram of simulated incomes (Public domain image created by Alan M. Polansky)

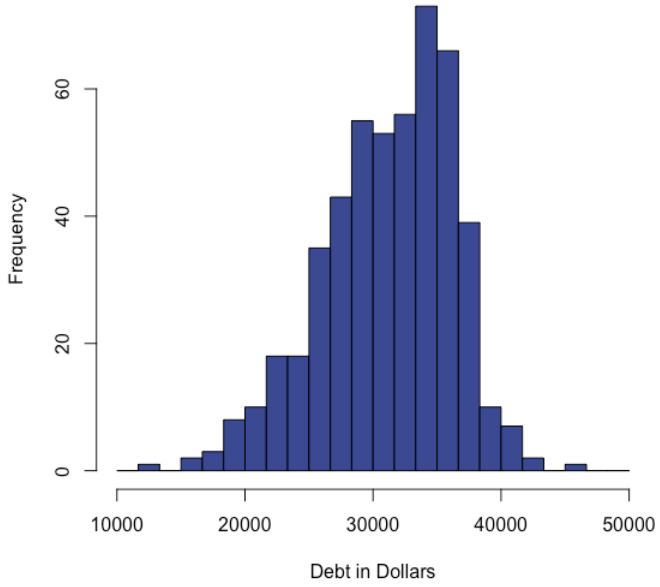


Figure **11.5.4**: Histogram of simulated student loan debt (Public domain image created by Alan M. Polansky)

Histograms are not only useful for visually depicting the frequencies of a set of data, but they can also help researchers visually determine whether there are unusual values in a set of data. Unusual values are called outliers.

Definition: Outlier

The *outliers* in histograms are values that seem significantly different compared to the rest of the values on the histogram, usually identified as values far out in either of the tails of a histogram.

Consider a set of data extracted from the National Longitudinal Study, which is sponsored by the U.S. Bureau of Labor Statistics. The data contains the yearly income approximately 7,000 individuals. A histogram of the income values is shown in Figure **11.5.5**. As is typical of income data, the histogram is skewed to the right. In further consideration of the right tail, one can also observe that some values are separated from the main bulk of the income values. These values are examples of outliers.

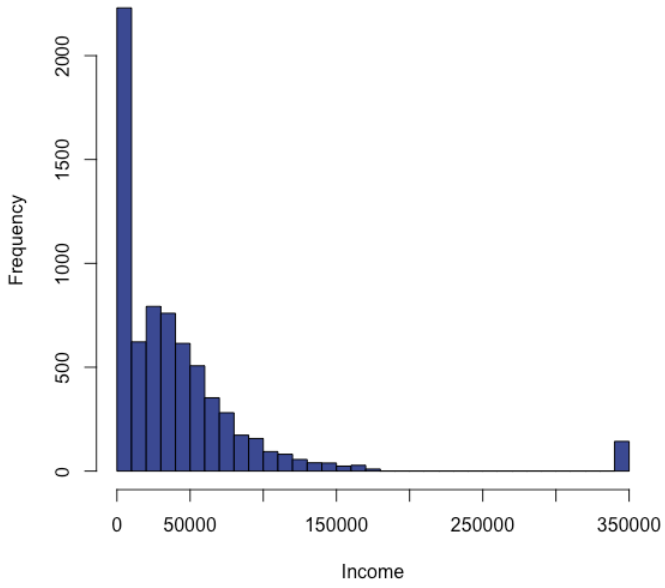


Figure **11.5.5**: The incomes of 7,006 individuals extracted from the National Longitudinal Study, which is sponsored by the U.S. Bureau of Labor Statistics (Public domain image created by Alan M. Polansky)

This page titled 11.5: Histograms is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

11.6: Scatterplots

A scatterplot is a plot of one variable against another where both variables would have at least an ordinal measurement scale. In a scatterplot only points are plotted, and the points are not connected with lines. An example of a scatterplot is given in Figure 11.17. The data used to produce this plot are from a sample of 1,078 sons' heights plotted against their fathers' heights, a dataset studied by the famous statistician Karl Pearson in 1903. Both measurements are in inches. Figure 11.17 shows a large cloud of points. Each of the points is plotted so that its position along the horizontal axis corresponds to the height of a father in the set of data. The corresponding position on the vertical axis corresponds to the height of that father's son. Note that the data are paired. That is, each height of a father is paired uniquely with a height of his son. Having two paired variables is a requirement for constructing scatterplots of data.

Definition: Scatterplot

A *scatterplot* is a plot of the individual values of two paired variables where is point is represented by a point on the plot.

Considering the plot in Figure **11.6.1**, we can see that while there is considerable variation in the height of the sons, even when the fathers' heights are quite close, there is a general pattern in the plot. For example, when we look at Figure **11.6.1**, we can observe that when the height of the father is in the range of about 60–62 inches, it would be quite unusual for the son to have a height greater that 70 inches. However, if the height of the father is around 70 inches, then about half of the sons have a height greater than 70 inches. Hence, while there is no direct relationship between the height of a father and their son, there is an association where the shorter fathers tend to have shorter sons and taller fathers tend to have taller sons.

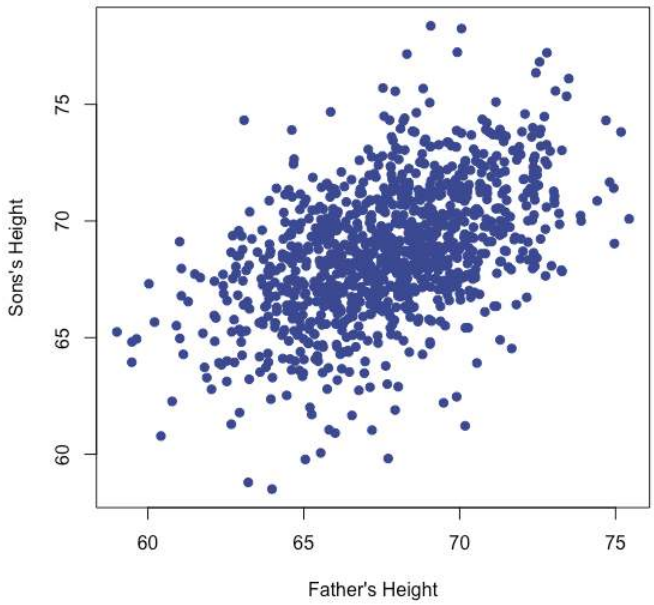


Figure **11.6.1**: The height of 1,078 sons' heights plotted against their fathers' heights, a dataset studied by the famous statistician Karl Pearson in 1903. Both measurements are in inches (Public domain image created by Alan M. Polansky)

Another interesting example is based on observations of geyser Old Faithful in Yellowstone National Park in the United States. The data discussed here is based on 299 measurements of the time between successive eruptions of the geyser and the duration of the subsequent eruption (Azzalini and Bowman 1990). Both measurements are recorded in minutes. A plot of the data is given in Figure **11.6.2**.

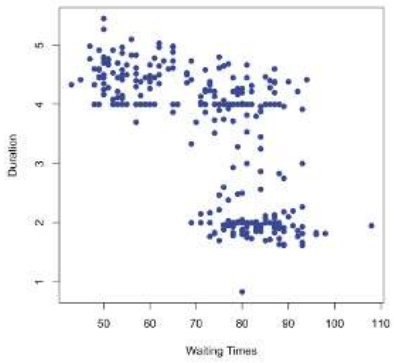


Figure **11.6.2**: Observations of the time interval between and the subsequent duration of eruptions of the geyser Old Faithful in Yellowstone National Park, Wyoming, United States from Azzalini and Bowman (1990) (Public domain image created by Alan M. Polansky)

There are many interesting details in the scatterplot shown in Figure **11.6.2**. The trend in the plot has some interesting features. For example, when the waiting time is less than about 70 minutes, the duration is almost always at least 4 minutes. When the waiting time is longer than 70 minutes, the duration can be anywhere between roughly 1 and 5 minutes. There also seems to be some anomalous behavior in the data. There are two apparent outliers. The first outlier is observed when the waiting time is around 80 minutes, and the duration is less than 1 minute. The duration is considerably less than any of the other waiting times observed in the data. The other outlier if observed when the waiting time is 100 minutes. The waiting time in this case is unusually long compared to the rest of the observed waiting times.

Another interesting aspect of this data can be observed when we consider duration times near 2 minutes and 5 minutes. The plot reveals that there are an unusually large number of observations that equal exactly 2 minutes and 5 minutes. This is the reason for the horizontal lines of points at these locations. A further analysis of the history of this data reveals that this is due to observations of the duration sometimes being rounded to the nearest minute.

This page titled 11.6: Scatterplots is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

11.7: Critical Analysis

When critically analyzing visual representations of data, it is important to observe any overall impressions or trends. Because the interpretation of visual representations of data have a large degree of subjectivity, small details in plots and graphs should not be overinterpreted. Another important issue is to determine whether the visual representation has been manipulated to cover up a trend, or perhaps make a trend appear where one does not really exist in the data.

The key questions when critically analyzing bar charts frequency distributions are:

1. How do the frequencies compare to one another?
2. Has the horizontal axis been ordered in a way to make it appear that there is a trend in the frequencies?
3. Has the horizontal axis been ordered in a way to emphasize a comparison between two categories?
4. Is the vertical axis properly scaled so that it starts at zero and has a range large enough to accommodate the data but not de-emphasize differences between categories?

The key questions when critically analyzing pie charts are:

1. Do some responses have higher frequencies than others?
2. Do some responses have lower frequencies than others?
3. Are the frequencies of all the responses about the same?
4. For a sequence of pie charts, is there a trend in the frequencies of the responses or do all the pie charts appear to be similar?
5. Have the pie charts been manipulated to emphasize or de-emphasize the frequencies for some responses, such as using a three-dimensional pie chart?

The key questions when critically analyzing line graphs are:

1. Is there a trend in time in the graph or is the line constant throughout the time period?
2. Are there times when there are peaks or valleys in the graph?
3. If more than one line is represented in the graph, how do the trends between the lines compare? Do they peak at similar times? Are there valleys at similar times?
4. Is the vertical axis properly scaled so that it starts at zero and has a range large enough to accommodate the data but not de-emphasize trends or differences?

The key questions when critically analyzing histograms are:

1. How do the frequencies compare to one another?
2. Is there a trend in the frequencies?
3. Are there gaps in the histogram where no data was observed in some classes?
4. How many modes does the histogram have?
5. If the histogram is unimodal, is it symmetric, skewed left, or skewed right?
6. Is the vertical axis properly scaled so that it starts at zero and has a range large enough to accommodate the data but not de-emphasize differences between classes?

The key questions when critically analyzing scatterplots are:

1. Does the plot show an association between the two variables?
2. Are there areas of the scatterplot where few data points were observed?
3. Are there any points far away from the clod of points that would be an outlier?
4. Are the axes properly scaled?

This page titled 11.7: Critical Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

11.7.1: Sampling Animals in Populations

In Section 4.6.5 we considered a study that directly researched animal behavior capture bias (Biro 2013). In the study four natural lakes with no fish in them and no natural or human predators were stocked with rainbow trout that had three different growth rates. At the end of the season, fish were sampled from the lakes using experimental nets designed to eliminate bias due to the size of the fish. The researcher predicted that the fast-growing fish, independent of their size, would be more likely to be captured in nets.

The capture data from the study are visually represented in Figures **11.7.1.1**–**11.7.1.3**. Each of the line graphs show the cumulative proportion of fish that were capture by each day of the study for the four lakes. The three lines represent the capture proportions for slow-, intermediate-, and fast-growing fish. All the plots indicate that that the cumulative proportion of fish captured increases. This makes logical sense because the more we go fishing in a lake, the larger proportion of the fish we will have caught. Within each plot the lines allow us to compare the trends between the three types of fish stocked in the lake. From these lines we can obtain two types of important information. In all the plots we can observe that while the trends of the three different types of fish are similar, there are differences in the proportion caught. For example, in Figure **11.7.1.2** the trend of the proportion of fish caught for the intermediate and the fast-growing fish are almost the same, the only difference being that the line for the fast-growing fish is higher than that of the intermediate-growing fish. Similar trends can be observed in Figures **11.7.1.2** and **11.7.1.3**. In a couple of cases the trend for the slow-growing fish are observed to be different than the other two types of fish. For example, in Figure **11.7.1.1** the slope of the trend line is smaller so that the line for the slow-growing fish and the fast-growing fish cross. A similar effect can be observed in Figure **11.7.1.3**.

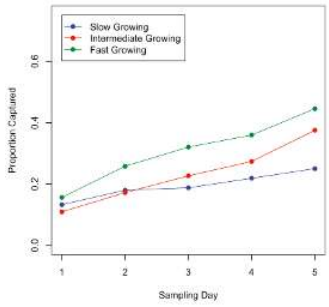


Figure **11.7.1.1**: The cumulative proportion of fish captured in the first lake for slow-, intermediate-, and fast-growing fish adapted from Biro (2013). (Public domain image created by Alan M.Polansky)

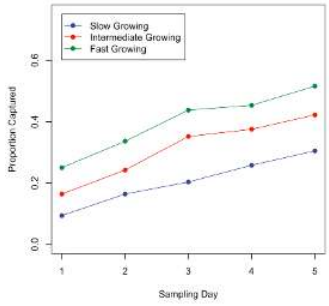


Figure **11.7.1.2**: The cumulative proportion of fish captured in the second lake for slow-, intermediate-, and fast-growing fish adapted from Biro (2013). (Public domain image created by Alan M.Polansky)

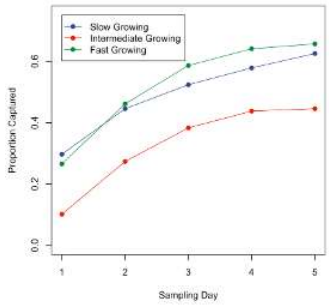


Figure **11.7.1.3**: The cumulative proportion of fish captured in the third lake for slow-, intermediate-, and fast-growing fish adapted from Biro (2013). (Public domain image created by Alan M.Polansky)

The visual representation of this data is crucial in observing these trends. One can image attempting to discern these types of conclusions from tables of the associated data. Using the line plots these trends can be observed easily and quickly.

This page titled 11.7.1: Sampling Animals in Populations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

11.8: Summary

Visual summaries of data can be an extremely helpful aid in understanding the structure of observed data. In this chapter we presented a few of the most popular types of visual summaries that you are likely to encounter in the media and research papers. Bar charts are visual representations of frequency distributions where the height of the bars corresponds to the frequencies. Bar charts can also be used to compare means, medians, ranges, standard deviations, and other numerical summaries for a variable using at least an interval measurement scale. Pie charts are an alternative method for visually representing frequencies, though they are more difficult to interpret. Line graphs are used to show how the value of a variable changes over time or space. Line plots are characterized by connecting the consecutive plotted points by lines so that any trends can be easily visually observed. Histograms are visual representations of frequency distributions where the variable of interest is measured on a interval or ordinal scale. The range of the variables is divided into classes, and the frequencies are plotted as rectangles whose width corresponds to the width of the class and whose height corresponds to the frequency. Shapes of histograms are classified by the number of peaks, or modes, that occur. Unimodal histograms are classified by symmetry. Finally, scatter plots can be used to investigate associations and relationships between two variables that are measured on interval or ratio measurement scales.

This page titled 11.8: Summary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

11.9: About the Quotation

John Tukey (1915–2000) was an American mathematician and statistician who was known for developing many new modern methods in these fields. He was particularly adept at algorithmic development and the development of new methods for visually representing data. An early pioneer in computer science, he is also credited with coining the term bit and the first published use of the word software.

This page titled 11.9: About the Quotation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

11.10: Practice and Exploration

1. In a survey of 75 graduates of a job training program, 20 of the graduates found jobs, 15 did not find jobs and were still looking for employment, 22 elected to pursue further education opportunities, and 18 did not find jobs and stopped looking for employment. Construct a bar graph of this frequency distribution.

2. Bart purchases a large bag of marbles. Each marble is either red (R), green (G), blue (B), white (W), or clear (C). One by one Bart pulls each marble from the bag until he has counted 57 marbles. The colors of the marbles are given in the table below. Construct a frequency distribution for the data and then construct a bar chart of the frequencies. What conclusions can you make from the bar chart?

R	C	B	G	C	R	B	R	R	C
G	W	W	G	G	C	R	G	R	W
R	C	B	G	R	W	G	B	W	R
R	B	C	R	B	C	C	R	C	R
C	C	W	G	G	W	R	R	R	W
B	G	R	C	B	W	C			

3. Suppose Lisa has surveyed her classmates as to where they would like to go for the next school field trip. To present the results, she has produced a pie chart showing the percentage of each response. See Figure 11.22. The possible field trip locations are the Springfield Historical Society, the Springfield Tar Pits, the Springfield Knowledgeum, the Springfield Nuclear Power Plant, and Olde Springfield Towne. Looking at Figure 11.22, which two locations had the most votes? Which location had the least votes?

4. The online automotive research group iSeeCars.com published a study that considered the colors of automobiles sold between January 2023 and April 2024 (iSeeCars 2025). Figure 11.23 shows a sequence of pie charts of the color preferences for trucks, sport utility vehicles, passenger cars, and sports cars. Based on these pie charts, do you see any differences between the color preferences based on the type of automobile?

5. The data given below correspond to the population of the town of West in Texas based on the U.S. Decennial Census from 1900 to 2020. Produce a line plot of the data and comment on any major trends that are observed.

Year	Population
1900	851
1910	1645
1920	1629
1930	1807
1940	1979
1950	2130
1960	2352
1970	2406
1980	2485
1990	2515
2000	2692
2010	2807
2020	2531

6. Suppose that the owner of a performing arts theater is interested in whether there is a trend in the sales of popcorn and drinks during an event. The owner keeps track of sales of both items during a selected concert during 10-minute intervals starting at 7:00 p.m. when the doors open and 10:00 p.m. when sales are concluded for the event. The data are given in the table below. Produce a

single line plot of the data and comment on any major trends that are observed.

Time	Popcorn	Drinks
7:00	7	3
7:10	3	3
7:20	8	9
7:30	8	16
7:40	5	15
7:50	4	14
8:00	1	12
8:10	4	3
8:20	1	11
8:30	3	4
8:40	5	5
8:50	8	11
9:00	5	8
9:10	1	5
9:20	1	11
9:30	8	10
9:40	5	10
9:50	7	11
10:00	2	5

7. The data given in the table below correspond to the mileage and asking price for 20 used cars of a similar type. The mileage is reported in thousands of miles, and the price is reported in thousands of dollars. Create a scatterplot of the data using the mileage on the horizontal axis and the price on the vertical axis. Comment on any trends that you observe. If you do not see

much of a trend, can you think of some reasons why?

Mileage	Price
9	33
8	37
2	29
20	18
30	18
32	19
4	27
40	19
36	20
9	32
31	21
9	33
19	24
13	24
31	21
12	24
20	24
22	25
21	25
22	25

8. The data in the table below show the age and race time of 25 participants in an annual 5k run. The age is reported in years, and the time is reported in minutes. Create a scatterplot of the data using the mileage on the horizontal axis and the price on the vertical axis. Comment on any trends that you observe. If you do not see much of a trend, can you think of some reasons why?

Age	Time
32	20
23	19
52	23
38	21
40	22
43	23
54	25
32	23
13	21
45	25
51	26
26	24
30	26
24	24
47	29
42	28
57	29
45	30
47	29
36	30
50	33
54	32
50	34
23	32
59	34

This page titled 11.10: Practice and Exploration is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

11.11: References

Azzalini, A., and A. W. Bowman. 1990. “A Look at Some Data on the Old Faithful Geyser.” *Applied Statistics* 39, 357–365. https://doi.org/10.2307/2347385

Biro, P. A. 2013. “Are Most Samples of Animals Systematically Biased? Consistent Individual Trait Differences Bias Samples Despite Random Sampling.” *Oecologia* 171, 339–345. https://doi.org/10.1007/s00442-012-2426-5

Flores, Andrew R., and Scott Barclay. 2016. “Backlash, Consensus, Legitimacy, or Polarization: The Effect of Same-Sex Marriage Policy on Mass Attitudes.” *Political Research Quarterly* 69, no. 1, 43–56. https://doi.org/10.1177/1065912915621175

iSeeCars. 2025. “Cars Are Half as Colorful as They Were 20 Years Ago.” https://www.iseecars.com/most-popula...r-colors-study

Wilson, Valarie. 2020. “Racial Disparities in Income and Poverty Remain Largely Unchanged Amid Strong Income Growth in 2019.” *Working Economics Blog*. https://www.epi.org/blog/racial-disp...rowth-in-2019/

This page titled 11.11: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

CHAPTER OVERVIEW

12: The Logical Structure of Formal Statistical Inference

“A useful property of a test of significance is that it exerts a sobering influence on the type of experimenter who jumps to conclusions on scanty data, and who might otherwise try to make everyone excited about some sensational treatment effect that can well be ascribed to the ordinary variation in his experiment.”

Gertrude Mary Cox

12.1: What to Expect

12.2: Hypotheses

12.3: Estimation and Evidence

12.4: Rejection Regions and Risk

This page titled 12: The Logical Structure of Formal Statistical Inference is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

12.1: What to Expect

In John Locke's *An Essay Concerning Human Understanding*, he suggests that few things in the world are known certainly to be true, but that we need to act based on the probability that something is true (Russell 2013). This line of reasoning is somewhat like the formal construction of what are known as statistical hypothesis tests and confidence intervals, though the specific interpretation of probability is somewhat different—essentially, the truth of a hypothesis about a social justice issue compared to evidence that is gathered from data observed from a sample. If there is enough evidence against the hypothesis, the hypothesis is abandoned as false; otherwise we conclude that sufficient evidence against the hypothesis was not observed. In this chapter we will first consider how statisticians construct hypotheses in a way that can be used within the formal framework required by statistical theory.

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

1. Define and discuss the concept of a hypothesis in the context of the parameters of a population.
2. Be able to define and identify the null and alternative hypotheses in the context of an empirical investigation.
3. Discuss the structure of looking for evidence in observed data for and against hypotheses.
4. Interpret rejecting and failing to reject a hypothesis.
5. Interpret significance levels and observed significance levels.
6. Interpret confidence intervals.

This page titled 12.1: What to Expect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

12.2: Hypotheses

In our earlier discussion of the scientific method, we presented the modern scientific method as an iterative five-step process based on observation, hypothesizing, testing, analysis, and drawing conclusions. As we shall see, statistical hypothesis testing encompasses the last four of these steps in a formal process. This section will focus on the second step of setting and identifying hypotheses.

A hypothesis in the abstract setting of the scientific method can be thought of as simply a statement about the universe that a researcher wishes to either prove or disprove. In the formal and practical setting of statistical hypothesis testing, such a hypothesis must be written in terms of a parameter of a population. Recall that a population is the collection of all the items or individuals that are of interest to the researcher. Each of the individuals has characteristics that can be measured, called variables, and a parameter was defined to be a summary of one or more of these characteristics computed on the measurements from the entire population. In statistics, a hypothesis is a specific statement about the value of an unknown parameter of a population.

Definition: Statistical Hypothesis

A *statistical hypothesis* is a statement about the value of a parameter of a population.

As an example, let us once again consider the professor who has 250 exams. The population in this case corresponds to the grades of the 250 students on the exam. Recalling this example, the professor needs to report how well he thinks the class did on the exam to the chair of the department. This is most succinctly reported as the mean of the two hundred and fifty exam scores. The professor has not graded the exam and hence does not know what this mean value is. This unknown mean value is the parameter of the population that is of interest to the professor. Suppose the professor thinks that the students did well. For example, suppose they think that the mean of the grade on all 250 exams will be greater than 80 on a 100-point scale. This is an example of a statistical hypothesis.

Many researchers will use the standard statistical notion when specifying hypothesis, which consists of specifying it with the letter H, followed by a colon and then statement of the hypothesis. For the example given above, the hypothesis would be represented as:

$$H: \text{The mean grade is greater than 80.}$$

Other researchers will use a more mathematical notation that is more common in statistics. In statistics the mean of a population is usually represented by the Greek letter mu, which looks like μ. For the current example, the researcher would state that μ represents the mean test grade and the hypothesis would be represented as $H: \mu > 80$. The scientific method emphasizes the statement of a hypothesis which is to be tested through observation. In the formal mathematical framework of statistics, two hypothesis are considered. These hypotheses are called the null and alternative hypotheses. These two hypotheses are set up as competitors to one another. That is, when one of the hypotheses is true, the other must be false, and in this sense statistical hypothesis testing is seen as a battle between the two that is decided by what is observed. This is a bit of a simplification of the process, and more details will be given as we proceed, but at this point it may be helpful to take a first look at setting up a statistical hypothesis test for us to understand how the null and alternative hypotheses are structured.

At the beginning of a statistical hypothesis test, the researcher sets the null hypothesis and the alternative hypothesis before any data has been observed. In the structure of the test, the null hypothesis is assumed to be true, then the data is observed. The alternative hypothesis will only be accepted as truth if substantial evidence is found in the data to indicate that the null hypothesis must be false. With this structure the null hypothesis has the comparative advantage. The researcher will assume the null is the truth, and will only abandon the null hypothesis if so much evidence is found against it that they must conclude that the alternative hypothesis is true.

Hence, at the end of a statistical hypothesis test there are two possible conclusions, and one of these conclusions is stronger and has more impact than the other. If the data are observed and substantial evidence is found against the null hypothesis, then it is rejected and the alternative hypothesis is accepted. This is a strong conclusion because the process started by assuming the null hypothesis was true, and so much evidence was found contradicting it that the researcher changed their mind and sided with the alternative hypothesis. On the other hand, if insufficient evidence is found in the observed data, then the researcher will stay with the null

hypothesis. This is the weak conclusion. Essentially, the researcher is concluding that not enough evidence was found against the null, so they will stay with the null hypothesis.

Definition: Null Hypothesis

In statistical hypothesis testing, the *null hypothesis* is initially assumed to be true. The statistical testing procedure will then determine how much evidence exists in the data contradicting the null hypothesis. The null hypothesis is usually denoted mathematically as H_0.

Definition: Alternative Hypothesis

In statistical hypothesis testing the *alternative hypothesis* is accepted if the null hypothesis is rejected. The alternative hypothesis is usually denoted mathematically as H_1.

The structure described above provides some motivation for how a researcher usually chooses the null and alternative hypothesis. If a researcher is attempting to prove that a new theory is true, they would want to put that theory in the alternative hypothesis. In that way, if the null is rejected, their theory in the alternative hypothesis would be accepted as a strong conclusion. Essentially, they would be arguing that they began the study by assuming that the current theory was true and did an experiment. In the experiment they found so much evidence contradicting the old theory that they now believe the new theory is true. If the new theory had been put into the null hypothesis, the best they could conclude is that they did not find enough evidence against the new theory to reject it, which is a much weaker conclusion.

Returning to the example of the professor sampling five exams from the class to grade, we stated earlier that the professor believes and would like to show that the mean grade on the exam for the entire class is greater than 80. The professor would want to set this conclusion to the alternative hypothesis, and hence the null hypothesis would be that the mean grade for the entire class is less than or equal to 80. That is,

$$H_0 : \text{The mean grade is less than or equal to 80,}$$

and

$$H_1 : \text{The mean grade is greater than 80,}$$

or in mathematical notation

$$H_0 : \mu \leq 80,$$

and

$$H_1 : \mu > 80.$$

The idea of comparing an old belief or old theory in the null hypothesis to a new theory in the alternative hypothesis is just one way that the null hypothesis can be considered in an intuitive way. One can also note in the null and alternative hypotheses specified for the case of the exams scores that the concept of equality is contained in the null hypothesis. That is, the null hypothesis specifies that the mean is less than or equal to 80 ($\mu \leq 80$) while the alternative hypothesis specifies that the mean is strictly greater that 80 ($\mu > 80$). This is consistent with the general rule that equality is always specified in the null hypothesis and not the alternative hypothesis. The reason for this has to do with the way risk is assessed in a statistical hypothesis test but is not important for our development here. This also applies when two or more groups are being compared. If a researcher is testing whether two gender identities have the same median salary in a geographical region, the hypothesis that the salaries are the same must be contained in the null hypothesis.

Now let us consider an example from a real study. As we have observed in several studies so far, social justice researchers have an interest in racial and ethnic disparities in the sentencing of individuals convicted of a crime. Recent research has started to consider how these disparities are shaped by the larger community. The racial threat theory states that defendants of color receive more severe punishments in places where they represent a larger or growing share of the population. This is due to their potential threat to white positions of power (Blalock 1967). A recent study explored the effects of racial, ethnic, and immigrant threat on sentence disposition and sentence length (Feldmeyer et al. 2015).

The study is based on individual-level sentencing data, county level data, and sentencing data taken from the Florida Department of Corrections Sentencing Guidelines database and its offender-based information system. The final sample included 501,027 black, Latino, and white defendants convicted of felonies and sentenced in Florida from 2000 to 2006. The composition and socioeconomic characteristics of the 67 Florida counties are taken from the 1990 and 2000 U.S. Census and the Uniform Crime Report.

The study considers two hypotheses as state in the research article. The first hypothesis is stated as (Feldmeyer et al. 2015):

Hypothesis 1: Racial/ethnic threat effects—black–white and Latino–white disparities in sentencing are larger in places with growing black and Latino populations,

and the second hypothesis is stated as,

Hypothesis 2: Immigrant threat effects—Latino–white (and possibly black–white) disparities in sentencing are larger in places with growing immigrant populations.

These are the effects that the authors are interested in demonstrating are true, and therefore these are the alternative hypotheses in the statistical tests that they use in the research article. Note that both hypotheses also imply inequality, which is consistent with the structure that the null hypothesis always contains the concept of equality or no change.

It is worth noting that an analog version of testing the hypotheses operates in the United States justice system. When a defendant is brought to trial, the jury is told that they must assume that the defendant is innocent. As a result of the evidence revealed at the trial, the jurors must decide if there is enough evidence to conclude, *beyond a reasonable doubt*, that the defendant is guilty, or if there was not sufficient evidence. At the end of the trial the defendant can be found guilty, in which case the evidence was found, or not guilty, in which case not enough evidence was found. The null hypothesis can be considered to be that the defendant is innocent, and the alternative is that the defendant is guilty. The strong conclusion, which is that there is evidence beyond a reasonable doubt, corresponds to rejecting the null hypothesis and accepting the alternative hypothesis. If not enough evidence is found, this scenario is the analog of failing to reject the null hypothesis. Note that the strong conclusion is sided with the defendant being guilty. The idea is that someone will only be punished in cases where there is so much evidence that a reasonable person would believe the defendant committed the crime.

This page titled 12.2: Hypotheses is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

12.3: Estimation and Evidence

Once the null and alternative hypotheses have been set, the scientific method proceeds to test the theory. Testing the theory in social justice studies usually consists of observing data from the population of interest based on a specified sampling plan. Once the data are observed, we proceed to the analysis step in the scientific method. That is, the observed data must be analyzed in some way that will let the researcher decide whether the data support the null or alternative hypotheses. Because of the structure of the hypotheses, what the researcher is really looking for is how much evidence is contained in the data that the null hypothesis is false. What this means is that the researcher needs to find a way of looking at the observed data that shows how much evidence exists against the null hypothesis.

Let us consider the example of the test grades once again. The null hypothesis for this situation is that the mean grade for the entire class is no greater than 80 and the alternative hypothesis is that the mean grade is greater than 80. Suppose the professor happens to sample the exams numbered 2, 6, 12, 14, and 22, whose scores are 76, 59, 74, 75, and 81, respectively. In this sample there does not seem to be much evidence against the null hypothesis that the class mean is less than or equal to 80. All but one of the exam scores is less than 80. In fact, if we compute the mean of the sampled scores comes out to be

$$(76 + 59 + 74 + 75 + 81) \div 5 = 365 \div 5 = 73.$$

If the mean score in the class was greater than 80, we would expect that our sample mean would also be greater than 80. In this case the sample mean is much less, so this is not evidence against the null hypothesis. So, what would be evidence against the null hypothesis? It makes logical sense that any time the mean of the sample is greater than 80 that this would be evidence against the null hypothesis—for example, if the professor happened to sample the exams numbered 2, 16, 17, 19, and 25, whose scores are 76, 79, 88, 82, and 97. In this sample there does seem to be some evidence against the null hypothesis. Several of the exam scores are greater than 80, and if we compute the mean of the sampled scores, it is

$$(76 + 79 + 88 + 82 + 97) \div 5 = 422 \div 5 = 84.4.$$

From this example we begin to see how evidence against the null hypothesis can be assessed. In this exposition one should notice that we are taking the mean of the sample to be an indication of the mean of the population. With the lack of any other information, we have assumed that the mean of the sample is in some way the best guess as to what the mean of the population is. That is, the mean of the sample is being used to estimate the unknown population mean.

Definition: Estimator

A statistical *estimator* is a summary computed on a sample that is taken to be the best indication of what a corresponding unknown parameter value may be.

Note that an estimator is likely not equal to the parameter value in the population. In the example considering the five sampled exams, we considered several samples in the previous section and none of the sample mean values were exactly equal to the true population mean. This is true in most applications, and the difference between the estimator and the true parameter value for the population is known as estimation error.

Definition: Estimation Error

The difference between the value of an estimator computed on a sample and the parameter value for the corresponding population is called *estimation error*.

For the five sampled exams, we first considered the case where exams numbered 2, 6, 12, 14, and 22 whose scores are 76, 59, 74, 75, and 81 were taken as the sample. The mean of the sampled scores comes out to be 73. The actual mean value for all the exams is 78, which is the value of the population parameter in this case. The estimation error is then $|73 - 78| = 5$, so that the estimate is 5 points away from the true value.

In practice the true population parameter is not known, and hence researchers are never completely sure how much error they are making when estimating a population parameter. However, for many estimators, specialized methods have been developed so that

the researcher can get an idea as to how large the error might be. This measure is called the standard error of the estimator.

Definition: Standard Error

The *standard error* of an estimator is an estimate of how large the estimation error of an estimator may be.

This estimate can be computed using the observed data and is interpreted in much the same way as a standard deviation. The standard error can be taken to be an estimate of the typical size of the error, and in an analog to the Empirical Rule discussed earlier, one would expect that the distance between the true population parameter value and the sample estimate to be less than two or three times the value of the standard error. It should be noted that there are some estimates where the standard error is interpreted differently. The problem is that the variation is not always centered on the true value of the population parameter. In most of these cases the standard error is usually not quoted by itself.

To measure the amount of evidence against the null hypothesis, a researcher will compare the estimator of a population parameter to the form of the null hypothesis. If the relationship between the estimator is consistent with the form of the null hypothesis, there is very little evidence against the truth of the null hypothesis. This corresponds to the first case with the exam data where the mean of the sample was computed to be 73, which is less than 80, and therefore is consistent with the form of the null hypothesis $H_0 : \mu \leq 80$. On the other hand, if the relationship between the estimator and the parameter is not consistent with the form of the null hypothesis, then *there is* evidence against the null hypothesis. This corresponds to the second case with the exam data where the mean of the sample was computed to be 84.4, which is greater than 80, and therefore is not consistent with the form of the null hypothesis $H_0 : \mu \leq 80$.

The question now is to determine how much evidence exists against the null hypothesis. To understand how evidence is measured, consider again the sample of exams. If the professor had happened to sample the exams numbered 17, 20, 21, 23, and 25 whose scores are 88, 91, 96, 97, and 97, the mean of the sampled scores becomes

$$(88 + 91 + 96 + 97 + 97) \div 5 = 469 \div 5 = 93.8.$$

The previous sample mean of 84.4 would provide more evidence against the null hypothesis. The reason for this is that while both sample means are consistent with the form of the alternative hypothesis and not the null hypothesis, the second sample mean is farther away from the important crossover point of 80, and therefore we should be surer that the null hypothesis should be rejected.

Think of it this way. Because our estimate is based on a sample, we know it is likely to have some error. Knowing this, it could happen that even though the sample mean is larger than 80, which is consistent with the alternative hypothesis, it could be that the error is large enough so the actual population mean is less than 80, which would be consistent with the null hypothesis. In the second case, the sample mean is farther away from the crossover point of 80, and hence in that case the error would have to be larger for this to take place. When the sample value is farther away from what the null hypothesis specifies, more evidence exists that the null hypothesis is false.

A statistical hypothesis test measures the amount of evidence by computing the distance between the estimator and the closest parameter value possible when the null hypothesis is true. This distance function is called a test statistic. In many cases the test statistic will also consider the standard error of the estimator.

Definition: Test Statistic

A *test statistic* is a function of the data that measures the evidence against the null hypothesis by computing the distance between an estimator and the closest possible parameter value specified by the null hypothesis.

For the example of the five exams sampled from the class papers, the test statistic could be written as $\text{sample mean} - 80$, which gives the raw distance between the mean of the sample and the closest value in the null hypothesis, which is 80. In most statistics textbooks the test statistic would be written as

$$\frac{\text{sample mean} - 80}{\text{standard error}}.$$

This second form of the test statistic is preferred because it tells the researcher how many standard errors the mean observed in the sample is from the closest parameter value specified by the null hypothesis. For example, if the test statistic is equal to 2, then

$$\frac{\text{sample mean} - 80}{\text{standard error}} = 2,$$

or

$$\text{sample mean} - 80 = 2 \times \text{standard error},$$

which implies that the mean of the sample is two standard errors above the closest possible parameter value specified by the null hypothesis. This value can be taken to be the amount of evidence against the null hypothesis. The question of whether this is enough evidence to conclude that the null hypothesis is false will be addressed in the next section.

This page titled 12.3: Estimation and Evidence is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

12.4: Rejection Regions and Risk

Given that a researcher has set a null and alternative hypothesis, has observed data from a population based on a sample, and has measured the amount of evidence against the null hypothesis using a test statistic, it is now time for the researcher make a conclusion. The main question that the researcher faces in this situation is whether enough evidence exists as measured by the test statistic for the researcher to reject the null hypothesis.

In the case of the sample of five exams, suppose as we did before that the mean of the sample is 84.4. This is consistent with the form of the alternative hypothesis, but as we discussed in the previous section, we also know that the mean computed on the sample is an estimate of the population mean, and there will be some error associated with the estimate. If the professor does not know the population mean (and in practical cases it will be unknown), there are two possible decisions the professor can make, and associated with these four possible outcomes on whether the correct decision has been made (see Table 12.1).

The first two possibilities can occur if the professor decides there is sufficient evidence to reject the null hypothesis. That is, the professor concludes that the mean grade on all the exams is greater than 80. In drawing this conclusion, it is possible that the professor is making the right choice, that the mean really is greater than 80. However, because the professor is deciding based on a sample, it is also possible that the professor is drawing the wrong conclusion, that is, that the mean is not greater than 80. Because of the sample that happened to be selected, the conclusion was made that it is. In terms of statistical hypothesis testing, this is equivalent to rejecting the null hypothesis when it is true. This is known as a Type I error.

Definition: Type I Error

A *Type I error* occurs when the null hypothesis is rejected when it is true.

The next two possibilities occur when the professor concludes that the mean grade on all the exams is less than or equal to 80. In drawing this conclusion, it is possible that the professor is making the right choice, that the mean really is less than or equal to 80. However, it is also possible that the professor is drawing the wrong conclusion, that is that the mean is not less than or equal to 80. In terms of statistical hypothesis testing, this is equivalent to failing to reject the null hypothesis when it is false. This is known as a Type II error.

Definition: Type II Error

A *Type II error* occurs when the null hypothesis is not rejected when it is false.

Of course, the professor, like any researcher, would like to avoid making either a Type I or Type II error. But, unless the researcher is willing to do a census of the entire population, this is not possible, and so researchers have to come to terms with the fact that these errors are always possible. However, it should be noted that once a conclusion is made, the type of potential error will be known. If a researcher rejects the null hypothesis, they know that they are either making a correct decision or they are committing a Type I Error because a Type II error can only occur when the researcher fails to reject the null hypothesis. Similarly, if the researcher fails to reject the null hypothesis, they know that they are either making the correct decision or that they are committing a Type II error because a Type I error only occurs when the null hypothesis is rejected.

Knowing that the potential for these errors cannot be avoided completely, the next best option is to attempt to ensure that they do not happen very often. As it turns out, researchers can control how often one of the errors occurs but not both, unless the researcher has a great amount of flexibility with the size of the sample. Let us assume for the moment that a sample size has been decided on, and in that case, we can control how often one of the errors occurs. Given the choice, which error do we want to control? The clear winner for this choice is the Type I error.

When structuring the null and alternative hypothesis, researchers will say the hypothesis they are trying to prove is the alternative hypothesis. Therefore, if they would like to prove their theory correct, they are interested in rejecting the null hypothesis, corresponding to the first row of Table 12.1. The only error that can occur when rejecting the null hypothesis is a Type I Error. Hence, the researcher has a keen interest in controlling how often this error occurs.

What the researcher will do is specify how much risk they are willing to take with a Type I error occurring when they analyze their data. This risk is mitigated by specifying the probability that a Type I Error will occur. This value is called the significance level of the test, and it is usually denoted using the Greek letter alpha (α).

Definition: Significance Level

The *significance level* of a statistical hypothesis test is the probability that the null hypothesis will be rejected when the null hypothesis is true.

Once the significance level is set, if the researcher rejects the null hypothesis, then they are able to say that they found sufficient evidence not to believe the null hypothesis, but then they are also able to acknowledge that they are taking an α amount of risk in drawing this conclusion.

Usually, the significance levels are set to $\alpha = 0.05$ or $\alpha = 0.01$. The level of $\alpha = 0.05$, also called the 5% significance level, is considered the standard level, but this value is largely arbitrary. This value was suggested by Sir Ronald S. Fisher in the 1930 book *Statistical Methods for Research Workers* (Cowles and Davis 1982; Fisher 1930).

The question now is why not just set the significance level to be very small, say $\alpha = 0.00000001$? That way, a researcher will be very sure that they will probably never commit a Type I error. The problem with this approach is that to ensure the type I Error rate is very small, the testing procedure will just end up rejecting the null hypothesis only when there is so much evidence against it that there is no other choice. That is, the researcher will rarely, if ever, reject the null hypothesis. When the null hypothesis is rejected, it will only be in the most obvious cases. The problem with this approach is that it goes well beyond the concept of reasonable doubt. Under this type of procedure many potential important discoveries will be ignored because of the researcher's insistence on never committing a Type I Error.

The other effect of choosing a very small significance level is that because the testing procedure will reject the null hypothesis less often, the test will usually fail to reject the null hypothesis—in this case, the rate at which the test fails to reject the null hypothesis when it should have will become larger. That is, the corresponding Type II Error rate will increase.

Hence, the choice of the significance level results from the need to balance the size of the probabilities of the Type I and Type II errors. In the way that a statistical test is set up, if a researcher specifies the Type I error rate, they must accept whatever Type II error rate is associated with it. What makes this more difficult is the fact that the Type II error will be unknown. Because of these factors statisticians have traditionally used $\alpha = 0.05$ or $\alpha = 0.01$ as a compromise.

The ability for a statistical testing procedure to detect a true alternative hypothesis is called the power of the test. Specifically, power refers to the probability that a true alternative hypothesis is detected by the test.

Definition: Power

The *power* of a statistical hypothesis test is the probability of rejecting the null hypothesis when the alternative hypothesis is true.

Note that when the null hypothesis is false, there are two possible outcomes: the correct decision (whose probability is equal to the power of the statistical test), or the Type II Error. If the alternative hypothesis is true, one of these outcomes must occur and it follows that the probability of rejecting the null hypothesis when it is false is equal to one minus the probability of a Type II Error.

Therefore a high power corresponds to a low Type II Error. As we stated above, the probability of a Type II Error is set when the significance level is set and is unknown. The only way to ensure that the power of a test will be large, and the Type II Error probability is small, is to choose a large sample from the population. Some researchers will do a power analysis. This is a statistical calculation that tells the researcher about how large the sample needs to be for the test to have certain power characteristics. These calculations can be quite complicated, but a power analysis is usually a sign of a carefully designed study.

Once the significance level is chosen, the test procedure will identify how much evidence is required to reject the null hypothesis. Specifically, the choice of the significance level will identify a range of values for the test statistic for which the researcher would

reject the null hypothesis. This range is known as the rejection region.

Definition: Rejection Region

The *rejection region* of a statistical test is a range of values such that if the test statistic is in the range, the null hypothesis is rejected for the specified significance level.

For the example of the five exams sampled from the class papers, the test statistic was be written as

$$\frac{\text{sample mean} - 80}{\text{standard error}}$$

or the number of standard errors the mean of the sample was above the mean specified in the null hypothesis, which was 80. Suppose that a researcher specified the significance level $\alpha = 0.05$ for this test. Some statistical calculations would show that for this significance level the null hypothesis should be rejected when the test statistic is greater that 2.32. This would be the rejection region for the test.

This page titled 12.4: Rejection Regions and Risk is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by .

Index

D

dire

Glossary

Alternative Hypothesis | In statistical hypothesis testing the *alternative hypothesis* is accepted if the null hypothesis is rejected. The alternative hypothesis is usually denoted mathematically as H_1.

Applicant Flow | The *applicant flow* for a position is the collection of individuals who apply for a job or promotion.

Axiom | An *axiom* is a statement that is taken to be true and serves as a starting point for further reasoning.

Bar Graph | A *bar graph* of a frequency distribution is a graph whose horizontal axis shows the category headings of a qualitative variable. The vertical axis shows the observed frequency, relative frequency, percentage, or numerical summary (mean, median, etc.) for each category, which correspond to the bar heights. The bars are separated by spaces.

Case Control Study | A *case control study* is an observational study that observes data about past differences between two groups.

Census | Observing every individual or item in a population is called a *census*.

Chance | The *chance* that an outcome from a random experiment occurs is a percentage between 0% and 100%, inclusive, that reflects the percentage of time the outcome would occur in a very large number of replications of the experiment under the same conditions for each replication.

Classes | A set of classes for a quantitative variable is a set of non-overlapping ranges that cover all the values of the variable so that each value of the variable falls in exactly one class.

Concept | A *concept* is a construct that has been clearly defined in terms of ideas that can be observed either directly or indirectly. The process of developing a construct into a concept is called *conceptualization*.

Concurrent Criterion Validity | A measurement system has *concurrent criterion validity* if it is strongly associated with a value of an established measure of a concept that is observed at the same time.

Confounding | Two or more factors are *confounded* when the effect of one of the factors cannot be distinguished from the effects of the other factors on the outcome.

Construct | *Constructs* are variables that cannot be observed either directly or indirectly.

Construct Validity | A measurement system has *construct validity* if it behaves the way researchers would expect in how it is related to measures of other concepts.

Content Validity | A measurement system has *content validity* if the dimensions and indicators cover the range of meanings within the concept.

Control Group | In an experiment that compares a standard treatment to a special treatment, the group that receives the standard treatment is called the *control group*.

Criterion Validity | A measurement system has *criterion validity* if it coincides statistically with other measures of the same concept.

Cross Classified Frequency Distribution | A cross-classified frequency distribution of two variables is a table that contains the frequencies, relative frequencies, and percentages of the number of times each pair of values occur in an observation of the two variables in the data.

Data | Measurements taken on individuals or items in a population are called *data*. A single measurement is called a *datum*.

Deduction | A conclusion is based on *deduction* if it follows logically from a set of statements that are known to be true.

Designed Experiment | A *designed experiment* is an experiment where the factors are controlled and set by the researcher prior to observing the outcome.

Dimension | A particular aspect of a concept is called a *dimension*.

Direct Observable | A variable is a *direct observable* if the observation can be measured directly and easily.

Empiricism | *Empiricism* is the philosophical theory that knowledge is gained through experience.

Estimation Error | The difference between the value of an estimator computed on a sample and the parameter value for the corresponding population is called *estimation error*.

Estimator | A statistical *estimator* is a summary computed on a sample that is taken to be the best indication of what a corresponding unknown parameter value may be.

Experiment | An *experiment* is any process that produces observations that provide evidence that can be used to determine if a hypothesis is true or not.

Extrapolation | *Extrapolation* refers to making conclusions for populations that are larger than the population that was observed.

Face Validity | A measurement system has *face validity* if the dimensions and the indicators make sense and are consistent with known research about the subject.

Factor | In an experiment a *factor* or treatment is a condition that is hypothesized to have an effect on, or is associated with, the response.

Factor Construct Validity | A measurement system has *factor construct validity* if a statistical factor analysis indicates that the number of dimensions and the grouping of the indicators within the dimensions is correct.

Frequency Distribution | The *frequency distribution* of a set of qualitative data is a table that contains the number of times each category of the data is observed. The individual counts for each category are called the *frequencies* of the categories.

Frequency Distribution (Qualitative Data) | The *frequency distribution* of a set of qualitative data is a table that contains the number of times each category of the data is observed, the proportion of data that was observed in each category, or the percentage of data that was observed in each category. The individual counts for each category are called the *frequencies* of the categories. The *relative frequencies* are computed by dividing each frequency by the total number of observations in the set of data. The *percentages* are computed by multiplying the relative frequencies by 100%.

Histogram | A *histogram* of a frequency distribution is a graph whose horizontal axis corresponds to the values of a quantitative variable and whose vertical axis corresponds to the observed frequency, relative frequency, or percentage for each class. The bars of the graph span the corresponding classes.

Indicator | A particular aspect of a dimension is called an *indicator*.

Indirect Observable | A variable is an *indirect observable* if the observation can only be measured through an additional process.

Induction | A conclusion is based on *induction* if it is generalized from a set of observations.

Internal Consistency Reliability | The reliability of a measurement system is assessed using *internal consistency reliability* if the association between the dimensions and the indicators within the dimension are all strongly associated with one another.

Line Graph | A line graph is a plot of points whose horizontal axis is a variable that has at least an ordinal measurement scale and whose vertical axis is an observation or measurement that has at least an ordinal measurement scale. The points are connected by line segments with respect to the order of the horizontal axis.

Location | A measure of *location* of a set of data is a measure that summarizes the set of data with a single value that represents the middle or center of the data.

Mean | To compute the *mean* of a set of data, add all the values in the data set and divide by the number of values in the data set.

Median | A *median* is any point which divides the lower half from the upper half of the data.

Modality | The *modality* of a histogram refers to how many "peaks" the histogram has. A histogram is called unimodal if there is one peak, bimodal if there are two peaks, multimodal if there are three or more peaks, and uniform if the histogram is relatively flat across the classes.

Mode | A peak on a histogram is called a *mode*.

Nominal Measurement Scale | A variable has a *nominal measurement scale* if the only possible valid mathematical comparison of two datum is whether they are equal.

Null Hypothesis | In statistical hypothesis testing, the *null hypothesis* is initially assumed to be true. The statistical testing procedure will then determine how much evidence exists in the data contradicting the null hypothesis. The null hypothesis is usually denoted mathematically as H_0.

Observational Study | An *observational study* is an experiment where the researcher does not control or set the factors.

Odds Ratio | If the odds of one outcome is m to $n-m$ and the odds of another outcome is k to $l-k$, then the odds ratio of the first outcome compared to the second outcome is

$$\begin{aligned}\text{odds ratio} &= \frac{m/(n-m)}{k/(l-k)} \\ &= \frac{m(l-k)}{k(n-m)}\end{aligned}$$

.

Operationalization | *Operationalization* is the process of developing the procedures that will result in the empirical observations representing the concepts.

Ordinal Measurement Scale | A variable has an *ordinal measurement scale* if the only possible valid mathematical comparisons of two datum is whether they are equal and whether one is greater than another.

Outlier | The *outliers* in histograms are values that seem significantly different compared to the rest of the values on the histogram, usually identified as values far out in either of the tails of a histogram.

Parameter | A *parameter* is a quantity or characteristic that can be computed when the whole population has been observed.

Percentile | The pth percentile of a set of data is the smallest point for which at least p% of the data is less than or equal to that point.

Pie Chart | A *pie chart* of a frequency distribution is a circular graph whose “slices” corresponds to the possible outcomes for a qualitative variable. The area of the slice compared to the area of the entire pie is equal to the percentage (or relative frequency) of that response.

Population | A *population* consists of all individuals or items that are of interest in a statistical study.

Power | The *power* of a statistical hypothesis test is the probability of rejecting the null hypothesis when the alternative hypothesis is true.

Predictive Criterion Validity | A measurement system has *predictive criterion validity* if it does a good job of predicting future values of an established measure of a concept.

Probability | The *probability* that an outcome from a random experiment occurs is a number between 0 and 1, inclusive, that reflects the proportion of time the outcome would occur in a very large number of replications of the experiment under the same conditions for each replication.

Quantitative Data | Data observed from a variable that have a natural and unique numeric value are called *quantitative data*.

Random Experiment | A *random experiment* is any experiment whose outcome cannot be predicted with absolute certainty.

Range | To compute the *range* of a set of data, subtract the smallest value in the set of data from the largest value in the set of data.

Ratio Measurement Scale | A variable has a *ratio measurement scale* if the possible valid mathematical comparisons of two datum are whether they are equal or whether one is greater than the other. Additionally, differences and ratios between two data are a valid comparison.

Rejection Region | The *rejection region* of a statistical test is a range of values such that if the test statistic is in the range, the null hypothesis is rejected for the specified significance level.

Representative Sample | A sample from a population is *representative* if the characteristics of the sample are like those of the population.

Response | In an experiment the *response* is the observed condition that is of interest to the researcher.

Sample | A *sample* is any observed part of a population.

Scatterplot | A *scatterplot* is a plot of the individual values of two paired variables where is point is represented by a point on the plot.

Science | *Science* is the systemic process of obtaining knowledge by deduction and careful observation in the form of testable hypotheses and predictions about the universe [

New Websters Dictionary and Thesaurus 1992]

Significance Level | The *significance level* of a statistical hypothesis test is the probability that the null hypothesis will be rejected when the null hypothesis is true.

Simple Random Sample | A *simple random sample* is any random method of choosing a sample from a population in such a way such that every individual or item in the population has an equal chance of being selected.

Skewness | The skewness of a unimodal histogram indicates whether there is a longer tail to the right of the peak, to the left of the peak, or if the tails are about equal on either side of the peak. If the longer tail is to the left, the histogram is left skewed. If the longer tail is to the right, then the histogram is right skewed. If the tails are about equal, the histogram is symmetric.

Standard Error | The *standard error* of an estimator is an estimate of how large the estimation error of an estimator may be.

Statistical Hypothesis | A *statistical hypothesis* is a statement about the value of a parameter of a population.

Tail | In a unimodal histogram the *tail* of the histogram is the range on each side of the of the peak corresponding to the thinner, tapering end of the histogram where data points are less frequent.

Test Statistic | A *test statistic* is a function of the data that measures the evidence against the null hypothesis by computing the distance between an estimator and the closest possible parameter value specified by the null hypothesis.

Test-Retest Reliability | The reliability of a measurement system is assessed using *test-retest reliability* if the same measurement is taken two or more times on the same group of individuals.

Treatment | In an experiment a factor or *treatment* is a condition that is hypothesized to have an effect on, or is associated with, the response.

Treatment Group | In an experiment that compares a standard treatment to a special treatment, the group that receives the special treatment is called the *treatment group*.

Type I Error | A *Type I error* occurs when the null hypothesis is rejected when it is true.

Type II Error | A *Type II error* occurs when the null hypothesis is not rejected when it is false.

Valid Measurement System | A measurement system is *valid* if the system correctly measures the concept without any systematic bias.

Variable | A single measurement that is taken on everyone observed in a population is called a *variable*.

Variation | A measure of *variation* of a set of data is a measure that summarizes the set of data with a single value that represents typically how far data points can be found from the location of the data.

Examples Used in the Text

- **Artificial Intelligence in Music Education:** Studies whether an AI chatbot is useful in enhancing the learning of students learning to play the piano. (1.4.10)

  Li, P. Ping, and B. Wang. 2023. "Artificial Intelligence in Music Education." *International Journal of Human–Computer Interaction* 40, no. 16, 4183–4192. https://doi.org/10.1080/10447318.2023.2209984

- **Attitudes About Marital Rape:** Considers the differences in attitudes about rape scenarios based on whether the attacker and the victim are in a marital relationship. (1.4.6, 5.4.2)

  Ferro, Christine, Jill Cermele, and Ann Saltzman. 2008. "Current Perceptions of Marital Rape: Some Good and Not-So-Good News." *Journal of Interpersonal Violence* 23, no. 6, 764–779. https://doi.org/10.1177/0886260507313947

- **Customer Avoidance of Advertising on Social Media:** A study that considers what factors might influence user to avoid advertising on social media.

  Li, X., Z. Liu, Y. Chen, and A. Ren. 2025. "Consumer Avoidance Toward Message Stream Advertising on Mobile Social Media: A Stimulus-Organism-Response Perspective." *Information Technology and People* 38, no. 1, 23–47. https://doi.org/10.1108/ITP-11-2020-0761

- **Economic Uncertainty and Corruption:** A study that considers how the economic conditions of a country might affect the rate of corruption. (1.4.2)

  Gyimah-Brempong, K. 2002. "Corruption, Economic Growth, and Income Inequality in Africa." *Economics of Governance* 3, 183–209. https://doi.org/10.1007/s101010200045

- **Effect of the COVID Crisis:** Studies the economic effect of the COVID pandemic for African Americans and Hispanic individuals. (1.4.5)

  Gould, E., and V. Wilson. 2020. "Black Workers Face Two of the Most Lethal Preexisting Conditions for Coronavirus—Racism and Economic Inequality." Economic Policy Institute. Retrieved from https://coilink.org/20.500.12592/5j5nvq on 04 Feb 2025. COI: 20.500.12592/5j5nvq.

- **Effects of a New Special Formula Fertilizer on Potato Growth:** Designed study that studies the yield and nutritional properties of potatoes grown using two specilized fertilizers. (3.7.4)

  Xu, F., A. Meng, Y. Liu, J. Li, and N. Wu. 2025. "Effects of New Special Formula Fertilizer on Potato Growth, Yield, and Fertilizer Utilization Efficiency." *Plants* 14, 627. https://doi.org/10.3390/plants140406

- **Gender and COVID-19 Outcomes:** This study considers whether there is a difference in outcomes from COVID-19 due to biological sex and gender identity (3.7.7)

  Danielsen, Ann Caroline, Katharine M. Lee, Marion Boulicault, Tamara Rushovich, Annika Gompers, Amelia Tarrant, Meredith Reiches, Heather Shattuck-Heidorn, Luke Weisman Miratrix, and Sarah S. Richardson. 2022. "Sex Disparities in COVID-19 Outcomes in the United States: Quantifying and Contextualizing Variation." *Social Science and Medicine* 294, 114716. https://doi.org/10.1016/j.socscimed.2022.114716

- **Improving Web Searches:** This study considers what user behaviors can be used to imply that a web search has been successful. (2.5.6)

Fox, Steve, Kuldeep Karnawat, Mark Mydland, Susan Dumais and Thomas White. 2005. “Evaluating Implicit Measures to Improve Web Search.” *ACM Transactions on Information Systems* 23, No. 2, 147–168. https://doi.org/10.1145/1059981.1059982

- **Inequities in Employment During COVID-19:** Considers how pre-existing inequalities in economic circumstances were exacerbated during the COVID-19 pandemic.

  Gemelas, Jordan, Jenna Davison, Case Keltner, and Samantha Ing. 2022. “Inequities in Employment by Race, Ethnicity, and Sector During COVID-19.” Journal of Racial and Ethnic Health Disparities 9, no. 1, 350–355. https://doi.org/10.1007/s40615-021-00963-3

- **Labor Market Discrimination:** A designed experiment that showed that resumes that use names associated with the African American community are less likely to recieve callbacks than resumes that use names that are associated with the white community. (2.5.1, 3.7.1)

  Bertrand, Marianne, and Sendhil Mullainathan. 2004. “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment
  on Labor Market Discrimination.” *American Economic Review* 94, no. 4, 991–1013. https://doi.org/10.1257/0002828042002561

- **Land Cover and Racial Segregation:** A study of how natural land cover such as trees varies based on racially segrgated neighborhoods.

  Jesdale, Bill M., Rachel Morello-Frosch, and Lara Cushing. 2013. “The Racial/Ethnic Distribution of Heat Risk–Related Land Cover in Relation to Residential Segregation.” *Environmental Health Perspectives* 121, no. 7, 811–817. https://doi.org/10.1289/ehp.1205919

- **Major League Baseball Pitch Clock:** An analysis of baseball injury data to determine whether the pitch clock has increased injuries in pitchers. (1.4.9)

  Puga, Troy, Nicholas Scigliano, McKenna Box, John Riehl. 2024. “The Major League Baseball Pitch Clock: First Year Analysis of Pitcher Injuries.” *Journal of Orthopaedics* 57, 133–136. https://doi.org/10.1016/j.jor.2024.06.022

- **Performance After Surgery in Baseball Players:** Considered whether MLB pitchers who had received ulnar collateral ligament (UCL) surgery return to pre-surgery performance after surgery. (2.5.3)

  Mastroianni, Michael A., Jennifer A. Kunes, Dany B. El-Najjar, Kyle K. Obana, Sohil S. Desai, Cole R. Morrissette , Frank J. Alexander, Alexander J. Rondon, David P. Trofa, Charles A. Popkin, William N. Levine, Christopher S. Ahmad. 2025. “Return to Performance After Ulnar Collateral Ligament Surgery in Major League Baseball Pitchers.” *The American Journal of Sports Medicine* 53, no. 1, 104–114. https://doi.org/10.1177/03635465241289142

- **Phone Use Policies in Schools in England:** A study of whether restrictive school phone policies are associated with improved student health. (4.6.3)

  Goodyear, Victoria A., Amie Randhawa, Péymane Adab, et al. 2025. "School Phone Policies and Their Association with Mental Wellbeing, Phone Use, and Social Media Use (SMART Schools): A Cross-Sectional Observational Study." *The Lancet Regional Health–Europe* 51, 101211. https://doi.org/10.1016/j.lanepe.2025.101211

- **Product Placement in Movies:** A study of how people reacted to consumer products being featured in movies. (1.4.4)

  D’Astous, Alain, and Francis Chartier. 2000. “A Study of Factors Affecting Consumer Evaluations and Memory of Product

Placements in Movies." *Journal of Current Issues and Research in Advertising* 22, no. 2, 31–40. https://doi.org/10.1080/10641734.2000.10505106

- **Race and Gender is Sentencing Disparity:** An analysis of a collection of studies that consider differences between individuals sentenced for similar crimes based on race and gender. (1.4.8)

  Lehmann Peter S. and Anna I. Gomez. 2022. "Racial, Ethnic, Gender, and Economic Sentencing Disparity." In *Handbook of Issues in Criminal Justice Reform in the United States*, 127–142. Elizabeth Jeglic and Cynthia Calkins, Editors. Springer. https://doi.org/10.1007/978-3-030-77565-0_8

- **Risk Factors for Homelessness:** Studies adolescent factors for negative psychiatric outcomes in adulthood and links these factors to homelessness.

  Herman, D. B., E. S. Susser, E. L. Struening, and B. L. Link. 1997. "Adverse Childhood Experiences: Are They Risk Factors for Adult Homelessness?" *American Journal of Public Health* 87, no. 2, 249–255. https://doi.org/10.2105/AJPH.87.2.249

- **Same Sex Marraige Laws and Hate Crimes:** Investigated whether advances in marriage equality have influenced criminal offenses against the LGBTQIA+ community. (2.5.2, 3.7.3)

  Tankard, Margaret E., and Elizabeth Levy Paluck. 2017. "The Effect of a Supreme Court Decision Regarding Gay Marriage on Social Norms and Personal Attitudes." *Psychological Science* 28, no. 9, 1334–1344. https://doi.org/10.1177/0956797617709594

- **Same-Sex Marriage and LGBT Discrimination**: Studies the effects of same-sex marriage legislation in the rate of hate crimes against individuals in the LGBTQ+ community (4.6.4)

  Nikolaou, Dimitrios. 2022. "Same-Sex Marriage Laws, LGBT Hate Crimes, and Employment Discrimination Charges." *Southern Economic Journal* 88, 869–905. https://doi.org/10.1002/soej.12548

- **Sampling Animals in Populations:** Studies the effects and inherent biases of sampling techniques when observing animal populations.

  Biro, P. A. 2013. "Are Most Samples of Animals Systematically Biased? Consistent Individual Trait Differences Bias Samples Despite Random Sampling." *Oecologia* 171, 339–345. https://doi.org/10.1007/s00442-012-2426-5

- **Satisfaction with Sport Participation:** A study of what factors may affect participation in sports. (5.4.3)

  Lepir, D., and S. Lakić. 2025. "Correlates of Sport Satisfaction: The Role of Success Level, Personality Traits, and Emotional Competence in Team and Individual Sports." *Psihologija*, 3 (online issue). https://doi.org/10.2298/PSI240723003L

- **Severity of School Discipline:** A study of the differences of the severity of school disciplinary actions taken against students with respect to race. (1.4.3)

  Skiba, Russell J., Robert H. Horner, Choong-Geun Chung, M. Karega Rausch, Seth L. May, and Tary Tobin. 2011. "Race is Not Neutral: A National Investigation of African American and Latino Disproportionality in School Discipline." *School Psychology Review* 40, no. 1, 85–107. https://doi.org/10.1080/02796015.2011.12087730

- **Sexuality, Race and Suicide:** This study, based on data from the National Violent Death Reporting System, studies the suicide rates of groups that traditionally suffer from discrimination. (1.4.7)

  Alvarez-Hernandez, Luis R., and Orion Mowbray. 2022. "Trends in Suicide Among Lesbian, Gay, and Bisexual Hispanic Individuals: National Violent Death Reporting System Data from 2012-2016." *Journal of Homosexuality* 70, no. 7, 1225–1239.

https://doi.org/10.1080/00918369.2021.2018878

- **Sleep and College Students:** A study of whether economic incentives can be used to develop better sleep habits in college students (3.7.6)

  Avery, Mallory, Osea Giuntella, and Peiran Jiao. 2025. "Why Don't We Sleep Enough? A Field Experiment Among College Students." *Review of Economics and Statistics* 107, no. 1, 65-77. https://doi.org/10.1162/rest_a_01242

- **Success Factors for Mobile Work in Health Care:** A deductive study of what factors provide satisfaction to professional users of mobile health care applications. (2.5.5)

  Chatterjee, S., S. Chakraborty, S. Sarker, S. Sarker, and F. Y. Lau. 2009. "Examining the Success Factors for Mobile Work in Healthcare: A Deductive Study." *Decision Support Systems* 46, 620–633. https://doi.org/10.1016/j.dss.2008.11.003

- **Supreme Court and Personal Atittudes:** A designed study that considers how a possible Supreme Court ruling on same-sex marriage might affect personal attitudes and same-sex marriage. (2.5.4, 3.7.5)

  Tankard, Margaret E., and Elizabeth Levy Paluck. 2016. "Norm Perception As a Vehicle for Social Change." *Social Issues and Policy Review* 10, no. 1, 181–211. https://doi.org/10.1111/sipr.12022

- **Thought and the Growth of Fungi Cultures:** A study of whether thoughts can be used to slow the growth of fungi.

  Barry, J. 1968. "General and Comparative Study of the Psychokinetic Effect on a Fungus Culture." *Journal of Parapsychology* 32, 237–243. https://www.proquest.com/openview/e9...ar&cbl=1818062

- **Tobacco Retailer Density:** Examines the association between disparities in the density of tobacco retailers and the disparities in the prevalence of tobacco use among youth and adults (4.6.7)

  Glasser, Allison M., Nathaniel Onnen, Peter F. Craigmile, Elli Schwartz, and Megan E. Roberts. 2022. "Associations Between Disparities in Tobacco Retailer Density and Disparities in Tobacco Use." *Preventive Medicine* 154, 106910. https://doi.org/10.1016/j.ypmed.2021.106910

- **Visual Exercise Environments and Treadmill Running:** A designed study that considers how the visual and audio environment affects the performance and attitudes of runners during treadmill running exercises. (3.7.2)

  Yeh, H.-P., J. A. Stone, S. M. Churchill, E. Bymer, and K. Davids. 2017. "Physical and Emotional Benefits of Different Exercise Environments Designed for Treadmill Running." *Environmental Research and Public Health* 14, no. 7, 752. https://doi.org/10.3390/ijerph14070752

- **Voter Engagement:** A study of participation rates of voters broken down by demographic groups. (1.4.1)

  Laurison, Daniel, Hana Brown, and Ankit Rastogi. 2021. "Voting Intersections: Race, Class, and Participation in Presidential Elections in the United States 2008–2016." *Sociological Perspectives* 65, no. 4, 768–789. https://doi.org/10.1177/07311214211059136

- **Young Women in the Justice System:** A study of the opportunities perceived by young women who have been incarcerated for drug offenses. (4.6.2)

  Myers, Randy. 2017. "'I Accept That I Have Nobody': Young Women, Youth Justice, and Expectations of Responsibility During Reentry." *Social Justice* 44, no. 1, 62–82.

Detailed Licensing

Overview

Title: Data and the Modern World

Webpages: 215

All licenses found:

- CC BY 4.0: 99.1% (213 pages)
- Undeclared: 0.9% (2 pages)

By Page

- Data and the Modern World - *CC BY 4.0*
  - Front Matter - *CC BY 4.0*
    - TitlePage - *CC BY 4.0*
    - InfoPage - *CC BY 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *CC BY 4.0*
    - Preface - *CC BY 4.0*
  - 1: Data and the Modern World - *CC BY 4.0*
    - 1.1: What to Expect - *CC BY 4.0*
    - 1.2: From Data to Decisions - *CC BY 4.0*
    - 1.3: The Use of Data in the Modern World - *CC BY 4.0*
      - 1.3.1: The Pharmaceutical Industry - *CC BY 4.0*
      - 1.3.2: The Retail Industry - *CC BY 4.0*
      - 1.3.3: Sports - *CC BY 4.0*
      - 1.3.4: The Online World - *CC BY 4.0*
      - 1.3.5: Education - *CC BY 4.0*
      - 1.3.6: Meteorology - *CC BY 4.0*
      - 1.3.7: Insurance - *CC BY 4.0*
      - 1.3.8: Energy - *CC BY 4.0*
    - 1.4: Some Applications - *CC BY 4.0*
      - 1.4.1: Voter Engagement - *CC BY 4.0*
      - 1.4.2: Economic Uncertainty and Corruption - *CC BY 4.0*
      - 1.4.3: Severity of School Discipline - *CC BY 4.0*
      - 1.4.4: Product Placement in Movies - *CC BY 4.0*
      - 1.4.5: The Effect of the COVID Crisis - *CC BY 4.0*
      - 1.4.6: Attitudes About Marital Rape - *CC BY 4.0*
      - 1.4.7: Sexuality, Race, and Suicide - *CC BY 4.0*
      - 1.4.8: Race and Gender in Sentencing Disparity - *CC BY 4.0*
      - 1.4.9: The Major League Baseball Pitch Clock - *CC BY 4.0*
      - 1.4.10: Artificial Intelligence in Music Education - *CC BY 4.0*
    - 1.5: About the Quotation - *CC BY 4.0*
    - 1.6: Practice and Exploration - *CC BY 4.0*
    - 1.7: References - *CC BY 4.0*
  - 2: Empiricism - *CC BY 4.0*
    - 2.1: What to Expect - *CC BY 4.0*
    - 2.2: Empiricism and Knowledge - *CC BY 4.0*
    - 2.3: The Modern Scientific Method - *CC BY 4.0*
    - 2.4: Feminist Empiricism - *CC BY 4.0*
    - 2.5: Critical Analysis - *CC BY 4.0*
      - 2.5.1: Labor Market Discrimination - *CC BY 4.0*
      - 2.5.2: Same-Sex Marriage Laws and Hate Crimes - *CC BY 4.0*
      - 2.5.3: Performance After Surgery in Baseball Players - *CC BY 4.0*
      - 2.5.4: The Supreme Court and Personal Attitudes - *CC BY 4.0*
      - 2.5.5: Success Factors for Mobile Work in Health Care - *CC BY 4.0*
      - 2.5.6: Improving Web Searches - *CC BY 4.0*
      - 2.5.7: Customer Avoidance of Advertising on Social Media - *CC BY 4.0*
    - 2.6: Summary - *CC BY 4.0*
    - 2.7: About the Quote - *CC BY 4.0*
    - 2.8: Practice and Exploration - *CC BY 4.0*
    - 2.9: Collaborative Activity - *CC BY 4.0*
    - 2.10: References - *CC BY 4.0*
  - 3: Experiments - *CC BY 4.0*
    - 3.1: What to Expect - *CC BY 4.0*
    - 3.2: Experiments - *CC BY 4.0*
    - 3.3: Designed Experiments - *CC BY 4.0*
    - 3.4: Observational Studies - *CC BY 4.0*
    - 3.5: Case Control Studies - *CC BY 4.0*
    - 3.6: Unethical Experiments - *CC BY 4.0*
    - 3.7: Critical Analysis - *CC BY 4.0*
      - 3.7.1: Labor Market Discrimination - *CC BY 4.0*

- 3.7.2: Visual Exercise Environments and Treadmill Running - *CC BY 4.0*
- 3.7.3: Same-Sex Marriage Laws and Hate Crimes - *CC BY 4.0*
- 3.7.4: Effects of a New Special Formula Fertilizer on Potato Growth - *CC BY 4.0*
- 3.7.5: The Supreme Court and Personal Attitudes - *CC BY 4.0*
- 3.7.6: Sleep and College Students - *CC BY 4.0*
- 3.7.7: Gender and COVID-19 Outcomes - *CC BY 4.0*
- 3.7.8: Thought and the Growth of Fungi Cultures - *CC BY 4.0*
- 3.7.9: Risk Factors for Homelessness - *CC BY 4.0*

- 3.8: Summary - *CC BY 4.0*
- 3.9: About the Quote - *CC BY 4.0*
- 3.10: Practice and Exploration - *CC BY 4.0*
- 3.11: Collaborative Activity - *CC BY 4.0*
- 3.12: References - *CC BY 4.0*

- 4: The Nature of Populations - *CC BY 4.0*
  - 4.1: What to Expect - *CC BY 4.0*
  - 4.2: Defining a Population - *CC BY 4.0*
  - 4.3: Practicality - *CC BY 4.0*
  - 4.4: Established Data - *CC BY 4.0*
  - 4.5: Defining Populations in Litigation - *CC BY 4.0*
  - 4.6: Critical Analysis - *CC BY 4.0*
    - 4.6.1: The Presidential Election of 1936 - *CC BY 4.0*
    - 4.6.2: Young Women in the Justice System - *CC BY 4.0*
    - 4.6.3: Phone Use Policies in Schools in England - *CC BY 4.0*
    - 4.6.4: Same-Sex Marriage and LGBT Discrimination - *CC BY 4.0*
    - 4.6.5: Sampling Animals in Populations - *CC BY 4.0*
    - 4.6.6: Land Cover and Racial Segregation - *CC BY 4.0*
    - 4.6.7: Tobacco Retailer Density - *CC BY 4.0*
  - 4.7: Summary - *CC BY 4.0*
  - 4.8: About the Quote - *CC BY 4.0*
  - 4.9: Practice and Exploration - *CC BY 4.0*
  - 4.10: Collaborative Activity - *CC BY 4.0*
  - 4.11: References - *CC BY 4.0*
- 5: The Nature of Data - *CC BY 4.0*
  - 5.1: What to Expect - *CC BY 4.0*
  - 5.2: Quantitative and Qualitative Data - *CC BY 4.0*
  - 5.3: The Structure of Data - *CC BY 4.0*
  - 5.4: Critical Analysis - *CC BY 4.0*
    - 5.4.1: Inequities in Employment During COVID-19 - *CC BY 4.0*
    - 5.4.2: Attitudes About Marital Rape - *CC BY 4.0*
    - 5.4.3: Satisfaction with Sport Participation - *CC BY 4.0*
    - 5.4.4: The Supreme Court and Personal Attitudes - *CC BY 4.0*
    - 5.4.5: Product Placement in Movies - *CC BY 4.0*
    - 5.4.6: The Racial Wealth Gap - *CC BY 4.0*
  - 5.5: Summary - *CC BY 4.0*
  - 5.6: About the Quotation - *CC BY 4.0*
  - 5.7: Practice and Exploration - *CC BY 4.0*
  - 5.8: Collaborative Activity - *CC BY 4.0*
  - 5.9: References - *CC BY 4.0*
- 6: Measuring Abstract Information - *CC BY 4.0*
  - 6.1: What to Expect - *CC BY 4.0*
  - 6.2: Abstract Concepts - *CC BY 4.0*
  - 6.3: Conceptualization and Operationalization - *CC BY 4.0*
  - 6.4: Validity and Reliability - *CC BY 4.0*
  - 6.5: Established Measurement Systems - *CC BY 4.0*
  - 6.6: Critical Analysis - *CC BY 4.0*
    - 6.6.1: Attitudes About Marital Rape - *CC BY 4.0*
    - 6.6.2: Gender and Race-Related Stress - *CC BY 4.0*
    - 6.6.3: Measuring Satisfaction in Sports - *CC BY 4.0*
  - 6.7: Summary - *CC BY 4.0*
  - 6.8: About the Quotation - *CC BY 4.0*
  - 6.9: Practice and Exploration - *CC BY 4.0*
  - 6.10: Collaborative Activity - *CC BY 4.0*
  - 6.11: References - *CC BY 4.0*
- 7: Probability and Randomness - *CC BY 4.0*
  - 7.1: What to Expect - *CC BY 4.0*
  - 7.2: Randomness - *CC BY 4.0*
  - 7.3: Defining Randomness - *CC BY 4.0*
  - 7.4: Probability - *CC BY 4.0*
  - 7.5: Odds - *CC BY 4.0*
  - 7.6: Critical Analysis - *CC BY 4.0*
    - 7.6.1: The Birthday Problem - *CC BY 4.0*
    - 7.6.2: Risk of Death from a Shark Attack - *CC BY 4.0*
    - 7.6.3: Bias in the Application of the Death Penalty - *CC BY 4.0*
    - 7.6.4: The Vietnam War Draft Lottery - *CC BY 4.0*
    - 7.6.5: Ovid and Gambling - *CC BY 4.0*
    - 7.6.6: Sports Activities and Lower Back Pain in Adolescence - *CC BY 4.0*

- 7.7: Summary - *CC BY 4.0*
- 7.8: About the Quotation - *CC BY 4.0*
- 7.9: Practice and Exploration - *CC BY 4.0*
- 7.10: Collaborative Activity - *CC BY 4.0*
- 7.11: References - *CC BY 4.0*

8: Observing Data from Populations - *CC BY 4.0*

- 8.1: What to Expect - *CC BY 4.0*
- 8.2: Motivation - *CC BY 4.0*
- 8.3: The Concept of Sampling - *CC BY 4.0*
- 8.4: Randomness and Observing Data - *CC BY 4.0*
- 8.5: More Complicated Sampling Methods - *CC BY 4.0*
- 8.6: Nonresponse - *CC BY 4.0*
- 8.7: Critical Analysis - *CC BY 4.0*
  - 8.7.1: The Achievement Gap - *CC BY 4.0*
  - 8.7.2: Tobacco Retailer Density - *CC BY 4.0*
  - 8.7.3: Gender and Race-Related Stress - *CC BY 4.0*
  - 8.7.4: Sampling Animals in Populations - *CC BY 4.0*
  - 8.7.5: Sports Activities and Lower Back Pain in Adolescence - *CC BY 4.0*
  - 8.7.6: Sleep and College Students - *CC BY 4.0*
  - 8.7.7: Visual Exercise Environments and Treadmill Running - *CC BY 4.0*
- 8.8: Summary - *CC BY 4.0*
- 8.9: About the Quotation - *CC BY 4.0*
- 8.10: Practice and Exploration - *CC BY 4.0*
- 8.11: Collaborative Activity - *CC BY 4.0*
- 8.12: References - *CC BY 4.0*

9: Numerical Methods for Summarizing Data - *CC BY 4.0*

- 9.1: What to Expect - *CC BY 4.0*
- 9.2: Summarizing Data - *CC BY 4.0*
- 9.3: Location - *CC BY 4.0*
- 9.4: Variation - *CC BY 4.0*
- 9.5: Percentiles - *CC BY 4.0*
- 9.6: Critical Analysis - *CC BY 4.0*
  - 9.6.1: Race and COVID-19 Mortality - *CC BY 4.0*
  - 9.6.2: Effects of a New Special Formula Fertilizer on Potato Growth - *CC BY 4.0*
  - 9.6.3: Visual Exercise Environments and Treadmill Running - *CC BY 4.0*
  - 9.6.4: Satisfaction with Sport Participation - *CC BY 4.0*
- 9.7: Summary - *CC BY 4.0*
- 9.8: About the Quotation - *CC BY 4.0*
- 9.9: Practice and Exploration - *CC BY 4.0*
- 9.10: Collaborative Activity - *CC BY 4.0*
- 9.11: References - *CC BY 4.0*

10: Frequency Distributions - *CC BY 4.0*

- 10.1: What to Expect - *CC BY 4.0*
- 10.2: Qualitative Data - *CC BY 4.0*
- 10.3: Quantitative Data - *CC BY 4.0*
- 10.4: Cross Classified Data - *CC BY 4.0*
- 10.5: Simpson's Paradox - *CC BY 4.0*
- 10.6: Critical Analysis - *CC BY 4.0*
  - 10.6.1: African American Student Debt - *CC BY 4.0*
  - 10.6.2: Race and School Discipline - *CC BY 4.0*
- 10.7: Summary - *CC BY 4.0*
- 10.8: About the Quotation - *CC BY 4.0*
- 10.9: Practice and Exploration - *CC BY 4.0*
- 10.10: Collaborative Activity - *CC BY 4.0*
- 10.11: References - *CC BY 4.0*

11: Visual Methods for Summarizing Data - *CC BY 4.0*

- 11.1: What to Expect - *CC BY 4.0*
- 11.2: Bar Graphs - *CC BY 4.0*
- 11.3: Pie Charts - *CC BY 4.0*
- 11.4: Line Graphs - *CC BY 4.0*
- 11.5: Histograms - *CC BY 4.0*
- 11.6: Scatterplots - *CC BY 4.0*
- 11.7: Critical Analysis - *CC BY 4.0*
  - 11.7.1: Sampling Animals in Populations - *CC BY 4.0*
- 11.8: Summary - *CC BY 4.0*
- 11.9: About the Quotation - *CC BY 4.0*
- 11.10: Practice and Exploration - *CC BY 4.0*
- 11.11: References - *CC BY 4.0*

12: The Logical Structure of Formal Statistical Inference - *CC BY 4.0*

- 12.1: What to Expect - *CC BY 4.0*
- 12.2: Hypotheses - *CC BY 4.0*
- 12.3: Estimation and Evidence - *CC BY 4.0*
- 12.4: Rejection Regions and Risk - *CC BY 4.0*

Back Matter - *CC BY 4.0*

- Index - *CC BY 4.0*
- Glossary - *CC BY 4.0*
- Examples Used in the Text - *CC BY 4.0*
- Detailed Licensing - *CC BY 4.0*
- Detailed Licensing - *Undeclared*

Detailed Licensing

Overview

Title: Data and the Modern World

Webpages: 215

All licenses found:

- CC BY 4.0: 99.1% (213 pages)
- Undeclared: 0.9% (2 pages)

By Page

- Data and the Modern World - *CC BY 4.0*
  - Front Matter - *CC BY 4.0*
    - TitlePage - *CC BY 4.0*
    - InfoPage - *CC BY 4.0*
    - Table of Contents - *Undeclared*
    - Licensing - *CC BY 4.0*
    - Preface - *CC BY 4.0*
  - 1: Data and the Modern World - *CC BY 4.0*
    - 1.1: What to Expect - *CC BY 4.0*
    - 1.2: From Data to Decisions - *CC BY 4.0*
    - 1.3: The Use of Data in the Modern World - *CC BY 4.0*
      - 1.3.1: The Pharmaceutical Industry - *CC BY 4.0*
      - 1.3.2: The Retail Industry - *CC BY 4.0*
      - 1.3.3: Sports - *CC BY 4.0*
      - 1.3.4: The Online World - *CC BY 4.0*
      - 1.3.5: Education - *CC BY 4.0*
      - 1.3.6: Meteorology - *CC BY 4.0*
      - 1.3.7: Insurance - *CC BY 4.0*
      - 1.3.8: Energy - *CC BY 4.0*
    - 1.4: Some Applications - *CC BY 4.0*
      - 1.4.1: Voter Engagement - *CC BY 4.0*
      - 1.4.2: Economic Uncertainty and Corruption - *CC BY 4.0*
      - 1.4.3: Severity of School Discipline - *CC BY 4.0*
      - 1.4.4: Product Placement in Movies - *CC BY 4.0*
      - 1.4.5: The Effect of the COVID Crisis - *CC BY 4.0*
      - 1.4.6: Attitudes About Marital Rape - *CC BY 4.0*
      - 1.4.7: Sexuality, Race, and Suicide - *CC BY 4.0*
      - 1.4.8: Race and Gender in Sentencing Disparity - *CC BY 4.0*
      - 1.4.9: The Major League Baseball Pitch Clock - *CC BY 4.0*
      - 1.4.10: Artificial Intelligence in Music Education - *CC BY 4.0*
    - 1.5: About the Quotation - *CC BY 4.0*
    - 1.6: Practice and Exploration - *CC BY 4.0*
    - 1.7: References - *CC BY 4.0*
  - 2: Empiricism - *CC BY 4.0*
    - 2.1: What to Expect - *CC BY 4.0*
    - 2.2: Empiricism and Knowledge - *CC BY 4.0*
    - 2.3: The Modern Scientific Method - *CC BY 4.0*
    - 2.4: Feminist Empiricism - *CC BY 4.0*
    - 2.5: Critical Analysis - *CC BY 4.0*
      - 2.5.1: Labor Market Discrimination - *CC BY 4.0*
      - 2.5.2: Same-Sex Marriage Laws and Hate Crimes - *CC BY 4.0*
      - 2.5.3: Performance After Surgery in Baseball Players - *CC BY 4.0*
      - 2.5.4: The Supreme Court and Personal Attitudes - *CC BY 4.0*
      - 2.5.5: Success Factors for Mobile Work in Health Care - *CC BY 4.0*
      - 2.5.6: Improving Web Searches - *CC BY 4.0*
      - 2.5.7: Customer Avoidance of Advertising on Social Media - *CC BY 4.0*
    - 2.6: Summary - *CC BY 4.0*
    - 2.7: About the Quote - *CC BY 4.0*
    - 2.8: Practice and Exploration - *CC BY 4.0*
    - 2.9: Collaborative Activity - *CC BY 4.0*
    - 2.10: References - *CC BY 4.0*
  - 3: Experiments - *CC BY 4.0*
    - 3.1: What to Expect - *CC BY 4.0*
    - 3.2: Experiments - *CC BY 4.0*
    - 3.3: Designed Experiments - *CC BY 4.0*
    - 3.4: Observational Studies - *CC BY 4.0*
    - 3.5: Case Control Studies - *CC BY 4.0*
    - 3.6: Unethical Experiments - *CC BY 4.0*
    - 3.7: Critical Analysis - *CC BY 4.0*
      - 3.7.1: Labor Market Discrimination - *CC BY 4.0*

- 3.7.2: Visual Exercise Environments and Treadmill Running - *CC BY 4.0*
- 3.7.3: Same-Sex Marriage Laws and Hate Crimes - *CC BY 4.0*
- 3.7.4: Effects of a New Special Formula Fertilizer on Potato Growth - *CC BY 4.0*
- 3.7.5: The Supreme Court and Personal Attitudes - *CC BY 4.0*
- 3.7.6: Sleep and College Students - *CC BY 4.0*
- 3.7.7: Gender and COVID-19 Outcomes - *CC BY 4.0*
- 3.7.8: Thought and the Growth of Fungi Cultures - *CC BY 4.0*
- 3.7.9: Risk Factors for Homelessness - *CC BY 4.0*

- 3.8: Summary - *CC BY 4.0*
- 3.9: About the Quote - *CC BY 4.0*
- 3.10: Practice and Exploration - *CC BY 4.0*
- 3.11: Collaborative Activity - *CC BY 4.0*
- 3.12: References - *CC BY 4.0*

- 4: The Nature of Populations - *CC BY 4.0*
  - 4.1: What to Expect - *CC BY 4.0*
  - 4.2: Defining a Population - *CC BY 4.0*
  - 4.3: Practicality - *CC BY 4.0*
  - 4.4: Established Data - *CC BY 4.0*
  - 4.5: Defining Populations in Litigation - *CC BY 4.0*
  - 4.6: Critical Analysis - *CC BY 4.0*
    - 4.6.1: The Presidential Election of 1936 - *CC BY 4.0*
    - 4.6.2: Young Women in the Justice System - *CC BY 4.0*
    - 4.6.3: Phone Use Policies in Schools in England - *CC BY 4.0*
    - 4.6.4: Same-Sex Marriage and LGBT Discrimination - *CC BY 4.0*
    - 4.6.5: Sampling Animals in Populations - *CC BY 4.0*
    - 4.6.6: Land Cover and Racial Segregation - *CC BY 4.0*
    - 4.6.7: Tobacco Retailer Density - *CC BY 4.0*
  - 4.7: Summary - *CC BY 4.0*
  - 4.8: About the Quote - *CC BY 4.0*
  - 4.9: Practice and Exploration - *CC BY 4.0*
  - 4.10: Collaborative Activity - *CC BY 4.0*
  - 4.11: References - *CC BY 4.0*
- 5: The Nature of Data - *CC BY 4.0*
  - 5.1: What to Expect - *CC BY 4.0*
  - 5.2: Quantitative and Qualitative Data - *CC BY 4.0*
  - 5.3: The Structure of Data - *CC BY 4.0*
  - 5.4: Critical Analysis - *CC BY 4.0*
    - 5.4.1: Inequities in Employment During COVID-19 - *CC BY 4.0*
    - 5.4.2: Attitudes About Marital Rape - *CC BY 4.0*
    - 5.4.3: Satisfaction with Sport Participation - *CC BY 4.0*
    - 5.4.4: The Supreme Court and Personal Attitudes - *CC BY 4.0*
    - 5.4.5: Product Placement in Movies - *CC BY 4.0*
    - 5.4.6: The Racial Wealth Gap - *CC BY 4.0*
  - 5.5: Summary - *CC BY 4.0*
  - 5.6: About the Quotation - *CC BY 4.0*
  - 5.7: Practice and Exploration - *CC BY 4.0*
  - 5.8: Collaborative Activity - *CC BY 4.0*
  - 5.9: References - *CC BY 4.0*
- 6: Measuring Abstract Information - *CC BY 4.0*
  - 6.1: What to Expect - *CC BY 4.0*
  - 6.2: Abstract Concepts - *CC BY 4.0*
  - 6.3: Conceptualization and Operationalization - *CC BY 4.0*
  - 6.4: Validity and Reliability - *CC BY 4.0*
  - 6.5: Established Measurement Systems - *CC BY 4.0*
  - 6.6: Critical Analysis - *CC BY 4.0*
    - 6.6.1: Attitudes About Marital Rape - *CC BY 4.0*
    - 6.6.2: Gender and Race-Related Stress - *CC BY 4.0*
    - 6.6.3: Measuring Satisfaction in Sports - *CC BY 4.0*
  - 6.7: Summary - *CC BY 4.0*
  - 6.8: About the Quotation - *CC BY 4.0*
  - 6.9: Practice and Exploration - *CC BY 4.0*
  - 6.10: Collaborative Activity - *CC BY 4.0*
  - 6.11: References - *CC BY 4.0*
- 7: Probability and Randomness - *CC BY 4.0*
  - 7.1: What to Expect - *CC BY 4.0*
  - 7.2: Randomness - *CC BY 4.0*
  - 7.3: Defining Randomness - *CC BY 4.0*
  - 7.4: Probability - *CC BY 4.0*
  - 7.5: Odds - *CC BY 4.0*
  - 7.6: Critical Analysis - *CC BY 4.0*
    - 7.6.1: The Birthday Problem - *CC BY 4.0*
    - 7.6.2: Risk of Death from a Shark Attack - *CC BY 4.0*
    - 7.6.3: Bias in the Application of the Death Penalty - *CC BY 4.0*
    - 7.6.4: The Vietnam War Draft Lottery - *CC BY 4.0*
    - 7.6.5: Ovid and Gambling - *CC BY 4.0*
    - 7.6.6: Sports Activities and Lower Back Pain in Adolescence - *CC BY 4.0*

- 7.7: Summary - *CC BY 4.0*
- 7.8: About the Quotation - *CC BY 4.0*
- 7.9: Practice and Exploration - *CC BY 4.0*
- 7.10: Collaborative Activity - *CC BY 4.0*
- 7.11: References - *CC BY 4.0*

- 8: Observing Data from Populations - *CC BY 4.0*
  - 8.1: What to Expect - *CC BY 4.0*
  - 8.2: Motivation - *CC BY 4.0*
  - 8.3: The Concept of Sampling - *CC BY 4.0*
  - 8.4: Randomness and Observing Data - *CC BY 4.0*
  - 8.5: More Complicated Sampling Methods - *CC BY 4.0*
  - 8.6: Nonresponse - *CC BY 4.0*
  - 8.7: Critical Analysis - *CC BY 4.0*
    - 8.7.1: The Achievement Gap - *CC BY 4.0*
    - 8.7.2: Tobacco Retailer Density - *CC BY 4.0*
    - 8.7.3: Gender and Race-Related Stress - *CC BY 4.0*
    - 8.7.4: Sampling Animals in Populations - *CC BY 4.0*
    - 8.7.5: Sports Activities and Lower Back Pain in Adolescence - *CC BY 4.0*
    - 8.7.6: Sleep and College Students - *CC BY 4.0*
    - 8.7.7: Visual Exercise Environments and Treadmill Running - *CC BY 4.0*
  - 8.8: Summary - *CC BY 4.0*
  - 8.9: About the Quotation - *CC BY 4.0*
  - 8.10: Practice and Exploration - *CC BY 4.0*
  - 8.11: Collaborative Activity - *CC BY 4.0*
  - 8.12: References - *CC BY 4.0*
- 9: Numerical Methods for Summarizing Data - *CC BY 4.0*
  - 9.1: What to Expect - *CC BY 4.0*
  - 9.2: Summarizing Data - *CC BY 4.0*
  - 9.3: Location - *CC BY 4.0*
  - 9.4: Variation - *CC BY 4.0*
  - 9.5: Percentiles - *CC BY 4.0*
  - 9.6: Critical Analysis - *CC BY 4.0*
    - 9.6.1: Race and COVID-19 Mortality - *CC BY 4.0*
    - 9.6.2: Effects of a New Special Formula Fertilizer on Potato Growth - *CC BY 4.0*
    - 9.6.3: Visual Exercise Environments and Treadmill Running - *CC BY 4.0*
    - 9.6.4: Satisfaction with Sport Participation - *CC BY 4.0*
  - 9.7: Summary - *CC BY 4.0*
  - 9.8: About the Quotation - *CC BY 4.0*
  - 9.9: Practice and Exploration - *CC BY 4.0*
  - 9.10: Collaborative Activity - *CC BY 4.0*
  - 9.11: References - *CC BY 4.0*
- 10: Frequency Distributions - *CC BY 4.0*
  - 10.1: What to Expect - *CC BY 4.0*
  - 10.2: Qualitative Data - *CC BY 4.0*
  - 10.3: Quantitative Data - *CC BY 4.0*
  - 10.4: Cross Classified Data - *CC BY 4.0*
  - 10.5: Simpson's Paradox - *CC BY 4.0*
  - 10.6: Critical Analysis - *CC BY 4.0*
    - 10.6.1: African American Student Debt - *CC BY 4.0*
    - 10.6.2: Race and School Discipline - *CC BY 4.0*
  - 10.7: Summary - *CC BY 4.0*
  - 10.8: About the Quotation - *CC BY 4.0*
  - 10.9: Practice and Exploration - *CC BY 4.0*
  - 10.10: Collaborative Activity - *CC BY 4.0*
  - 10.11: References - *CC BY 4.0*
- 11: Visual Methods for Summarizing Data - *CC BY 4.0*
  - 11.1: What to Expect - *CC BY 4.0*
  - 11.2: Bar Graphs - *CC BY 4.0*
  - 11.3: Pie Charts - *CC BY 4.0*
  - 11.4: Line Graphs - *CC BY 4.0*
  - 11.5: Histograms - *CC BY 4.0*
  - 11.6: Scatterplots - *CC BY 4.0*
  - 11.7: Critical Analysis - *CC BY 4.0*
    - 11.7.1: Sampling Animals in Populations - *CC BY 4.0*
  - 11.8: Summary - *CC BY 4.0*
  - 11.9: About the Quotation - *CC BY 4.0*
  - 11.10: Practice and Exploration - *CC BY 4.0*
  - 11.11: References - *CC BY 4.0*
- 12: The Logical Structure of Formal Statistical Inference - *CC BY 4.0*
  - 12.1: What to Expect - *CC BY 4.0*
  - 12.2: Hypotheses - *CC BY 4.0*
  - 12.3: Estimation and Evidence - *CC BY 4.0*
  - 12.4: Rejection Regions and Risk - *CC BY 4.0*
- Back Matter - *CC BY 4.0*
  - Index - *CC BY 4.0*
  - Glossary - *CC BY 4.0*
  - Examples Used in the Text - *CC BY 4.0*
  - Detailed Licensing - *CC BY 4.0*
  - Detailed Licensing - *Undeclared*